

Sachiyo Arai · Kazuhiro Kojima
Koji Mineshima · Daisuke Bekki
Ken Satoh · Yuiko Ohta (Eds.)

LNAI 10838

New Frontiers in Artificial Intelligence

JSAI-isAI Workshops, JURISIN,
SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI
Tsukuba, Tokyo, November 13–15, 2017
Revised Selected Papers

 Springer

Lecture Notes in Artificial Intelligence

10838

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/1244>

Sachiyo Arai · Kazuhiro Kojima
Koji Mineshima · Daisuke Bekki
Ken Satoh · Yuiko Ohta (Eds.)

New Frontiers in Artificial Intelligence

JSAI-isAI Workshops, JURISIN,
SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI
Tsukuba, Tokyo, November 13–15, 2017
Revised Selected Papers

Editors

Sachiyo Arai
Chiba University
Chiba
Japan

Kazuhiro Kojima
National Institute of Advanced Industrial
Science and Technology
Ibaraki
Japan

Koji Mineshima
Ochanomizu University
Tokyo
Japan

Daisuke Bekki
Ochanomizu University
Tokyo
Japan

Ken Satoh
National Institute of Informatics
Tokyo
Japan

Yuiko Ohta
Fujitsu Laboratories Limited
Kanagawa
Japan

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-93793-9 ISBN 978-3-319-93794-6 (eBook)
<https://doi.org/10.1007/978-3-319-93794-6>

Library of Congress Control Number: 2018947317

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

JSAI (The Japanese Society for Artificial Intelligence) is a premier academic society that focuses on artificial intelligence in Japan and was established in 1986. JSAI-isAI (JSAI International Symposium on Artificial Intelligence) 2017 was the ninth international symposium on AI supported by the JSAI. JSAI-isAI 2017 was successfully held during November 13–15 at University of Tsukuba in Tokyo, Japan. In all, 203 people from 15 countries participated.

JSAI-isAI 2017 included seven workshops, where 16 invited talks and 91 papers were presented. This volume, *New Frontiers in Artificial Intelligence: JSAI-isAI 2017 Workshops*, is the proceedings of JSAI-isAI 2017. From the seven workshops (JURISIN 2017, SKL 2017, AI-Biz 2017, LENLS 14, AAA 2017, SCIDOCA 17, and kNeXI 2017) 30 papers were carefully selected and revised according to the comments of the workshop Program Committees. The acceptance rate was about 32%. This resulted in the excellent selection of papers that are representative of some of the topics of AI research both in Japan and in other parts of the world.

JURISIN 2017 was the 11th international workshop on Juris-informatics. Juris-informatics is a new research area that studies legal issues from the perspective of informatics. The purpose of this workshop was to discuss both fundamental and practical issues among people from various backgrounds such as law, social science, information and intelligent technology, logic and philosophy, including the conventional “AI and law” area.

SKL 2017 (the 4th International Workshop on Skill Science) aimed to internationalize research on skill sciences by organizing the meeting. Human skills involve well-attuned perception and fine motor control, often accompanied by thoughtful planning. The involvement of body, environment, and tools mediating them makes the study of skills unique among research on human intelligence.

AI-Biz 2017 (Artificial Intelligence of and for Business) was the second workshop held to foster the concepts and techniques of business intelligence (BI) in artificial intelligence. BI should include such cutting-edge techniques as data science, agent-based modeling, complex adaptive systems, and IoT. The main purpose of this workshop is to provide a forum for participants to discuss important research questions and practical challenges in business intelligence, business informatics, data analysis and agent-based modeling, to exchange the latest results, and to join efforts in solving common challenges.

LENLS 14 was the 14th event in the series, and it focused on the formal and theoretical aspects of natural language. LENLS (Logic and Engineering of Natural Language Semantics) is an annual international workshop recognized internationally in the formal syntax-semantics-pragmatics community. It has been bringing together for discussion and interdisciplinary communication researchers working on formal theories of natural language syntax, semantics and pragmatics, (formal) philosophy, artificial intelligence, and computational linguistics.

AAA 2017 (the Third International Workshop on Argument for Agreement and Assurance) focused on argumentation, which has now become an interdisciplinary research subject receiving much attention from diverse communities including formal logic, informal logic, and artificial intelligence. It aims at analyzing, evaluating, and systematizing various aspects of human argument appearing in the media — television, newspapers, WWW, etc. — and also artificial arguments constructed from structured knowledge with logical language and inference rules. Their research achievements are widely applicable to various domains such as safety, political, medical, and legal domains.

SCIDOCA 17 (the Second International Workshop on Scientific Document Analysis) focused on the recent proliferation of scientific papers and technical documents that has become an obstacle to efficient acquisition of new information in various fields. It is almost impossible for individual researchers to check and read all related documents. Even retrieving relevant documents is becoming more difficult. This workshop gathered together all researchers and experts who work on scientific document analysis from various perspectives, and invited technical paper presentations and system demonstrations that cover any aspects of scientific document analysis.

HAT-MASH 2016 (Healthy Aging Tech Mashup Service, Data, and People) was the second international workshop in the series that bridges healthy aging and elderly care technology, information technology, and service engineering. The main objective of this workshop was to provide a forum for participants to discuss important research questions and practical challenges in healthy aging and elderly care support and to promote transdisciplinary approaches.

kNeXI 2017 (Knowledge Explication in Industry) was aimed at promoting research about improving the skill and knowledge of employees in industry. kNeXI focuses on industry in which the knowledge of employees is important to their daily work. Such industry includes elderly care, education, health-promotion services, etc.

It is our great pleasure to be able to share some highlights of these fascinating workshops in this volume. We hope this book will introduce readers to the state-of-the-art research outcomes of JSAI-isAI 2017, and motivate them to participate in future JSAI-isAI events.

April 2018

Sachiyo Arai
 Kazuhiro Kojima
 Koji Mineshima
 Ken Satoh
 Daisuke Bekki
 Yuiko Ohta

Organization

JURISIN 2017

Workshop Chair

Ken Satoh National Institute of Informatics, Japan

Steering Committee

Takehiko Kasahara Toin Yokohama University, Japan
Makoto Nakamura Nagoya University, Japan
Katsumi Nitta Tokyo Institute of Technology, Japan
Seiichiro Sakurai Meiji Gakuin University, Japan
Ken Satoh National Institute of Informatics, Japan
Satoshi Tojo Advanced Institute of Science and Technology, Japan
Katsuhiko Toyama Nagoya University, Japan

Advisory Committee

Trevor Bench-Capon The University of Liverpool, UK
Tomas Gordon Fraunhofer FOKUS, Germany
Henry Prakken University of Utrecht and Groningen, The Netherlands
John Zeleznikow Victoria University, Australia
Robert Kowalski Imperial College London, UK
Kevin Ashley University of Pittsburgh, USA

Program Committee

Thomas Ågotnes University of Bergen, Norway
Ryuta Arisaka National Institute of Informatics, Japan
Kristijonas Cyras Imperial College, UK
Marina De Vos University of Bath, UK
Phan Minh Dung AIT, Thailand
Randy Goebel University of Alberta, Canada
Guido Governatori NICTA, Australia
Tatsuhiko Inatani Kyoto University, Japan
Tokuyasu Kakuta Chuo University, Japan
Yoshinobu Kano Shizuoka University, Japan
Mi-Young Kim University of Alberta, Canada
Nguyen Le Minh Japan Advanced Institute of Science and Technology,
Japan
Beishui Liao Zhejiang University, China
Hatsuru Morita Tohoku University, Japan
Makoto Nakamura Nagoya University, Japan

Katumi Nitta	Tokyo Institute of Technology, Japan
Paulo Novais	University of Minho, Portugal
Julian Padgett	University of Bath, UK
Ginevra Peruginelli	ITTIG-CNR, Italy
Seiichiro Sakurai	Meiji Gakuin University, Japan
Katsuhiko Sano	Hokkaido University, Japan
Giovanni Sartor	EUI/CIRSFID, Italy
Ken Satoh	National Institute of Informatics and Sokendai, Japan
Akira Shimazu	Japan Advanced Institute of Science and Technology, Japan
Fumio Shimpo	Keio University, Japan
Satoshi Tojo	Japan Advanced Institute of Science and Technology, Japan
Katsuhiko Toyama	Nagoya University, Japan
Rob van den Hoven van Genderen	VU University Amsterdam, The Netherlands
Leon van der Torre	University of Luxembourg, Luxembourg
Bart Verheij	University of Groningen, The Netherlands
Katsumasa Yoshikawa	IBM Research Tokyo, Japan
Masaharu Yoshioka	Hokkaido University, Japan
Harumichi Yuasa	Institute of Information Security, Japan
Yueh-Hsuan Weng	Tohoku University, Japan
Adam Wyner	University of Aberdeen, UK

SKL 2017

Workshop Chair

Tsutomu Fujinami	Japan Advanced Institute of Science and Technology
------------------	--

Program Committee

Masaki Suwa	Keio University, Japan
Ken Hashizume	Osaka University, Japan
Mihoko Otake	RIKEN, Japan
Yoshifusa Matsuura	Yokohama National University, Japan
Yuta Ogai	Tokyo Polytechnic University, Japan
Kentaro Kodama	Kanagawa University, Japan

AI-Biz 2017

Workshop Chair

Takao Terano	Tokyo Institute of Technology, Japan
--------------	--------------------------------------

Workshop Co-chairs

Hiroshi Takahashi	Keio University, Japan
Setsuya Kurahashi	University of Tsukuba, Japan

Program Committee

Takao Terano	Tokyo Institute of Technology, Japan
Hiroshi Takahashi	Keio University, Japan
Setsuya Kurahashi	University of Tsukuba, Japan
Hiroshi Deguchi	Tokyo Institute of Technology, Japan
Reiko Hishiyama	Waseda University, Japan
Manabu Ichikawa	National Institute of Public Health, Japan
Yoko Ishino	Yamaguchi University, Japan
Hajime Kita	Kyoto University, Japan
Hajime Mizuyama	Aoyama Gakuin University, Japan
Masakazu Takahashi	Yamaguchi University, Japan
Shingo Takahashi	Waseda University, Japan
Takashi Yamada	Yamaguchi University, Japan

LENLS 14**Workshop Chair**

Katsuhiko Sano	Hokkaido University, Japan
----------------	----------------------------

Workshop Co-chairs

Daisuke Bekki	Ochanomizu University/JST CREST, Japan
Koji Mineshima	Ochanomizu University/JST CREST, Japan
Elin McCready	Aoyama Gakuin University, Japan

Program Committee

Katsuhiko Sano	Hokkaido University, Japan
Daisuke Bekki	Ochanomizu University/JST CREST, Japan
Koji Mineshima	Ochanomizu University/JST CREST, Japan
Elin McCready	Aoyama Gakuin University, Japan
Alastair Butler	Faculty of Humanities, Hirosaki University, Japan
Richard Dietz	iCLA, Yamanashi Gakuin University, Japan
Naoya Fujikawa	Tokyo Metropolitan University, Japan
Yoshiki Mori	University of Tokyo, Japan
Yasuo Nakayama	Osaka University, Japan
David Y. Oshima	Nagoya University, Japan
Osamu Sawada	Mie University, Japan
Wataru Uegaki	Leiden University, The Netherlands
Katsuhiko Yabushita	Naruto University of Education, Japan
Tomoyuki Yamada	Hokkaido University, Japan
Shunsuke Yatabe	Kyoto University, Japan
Kei Yoshimoto	Tohoku University, Japan

AAA 2017

Workshop Chair

Kazuko Takahashi Kwansei Gakuin University, Japan

Workshop Co-chairs

Yoshiki Kinoshita Kanagawa University, Japan
Tim Kelly University of York, UK
Hiroyuki Kido Sun Yat-sen University China

Program Committee

Martin Caminada Cardiff University, UK
Ewen Denney SGT/NASA Ames Research Center, USA
Juergen Dix Clausthal University of Technology, Germany
Phan Minh Dung Asian Institute of Technology, Thailand
Richard Hawkins University of York, UK
C. Michael Holloway NASA Langley Research Center, USA
Antonis Kakas University of Cyprus, Cyprus
Tim Kelly University of York, UK
Hiroyuki Kido Sun Yat-sen University, China
Yoshiki Kinoshita Kanagawa University, Japan
Yutaka Matsuno Nihon University, Japan
Nir Oren The University of Aberdeen, UK
John Rushby SRI International, USA
Chiaki Sakama Wakayama University, Japan
Ken Satoh National Institute of Informatics and Sokendai, Japan
Guillermo Ricardo Simari Universidad del Sur in Bahia Blanca, Argentina
Kenji Taguchi National Institute of Advanced Industrial Science
and Technology, Japan
Kazuko Takahashi Kwansei Gakuin University, Japan
Toshinori Takai Nara Institute of Science and Technology, Japan
Makoto Takeyama Kanagawa University, Japan
Paolo Torrioni University of Bologna, Italy
Charles Weinstock Software Engineering Institute, USA
Stefan Woltran Vienna University of Technology, Austria
Shuichiro Yamamoto Nagoya University, Japan

SCIDOCA 2017

Workshop Chairs

Yuji Matsumoto Nara Institute of Science and Technology, Japan
Hiroshi Noji Nara Institute of Science and Technology, Japan

Program Committee

Takeshi Abekawa	National Institute of Informatics, Japan
Akiko Aizawa	National Institute of Informatics, Japan
Naoya Inoue	Tohoku University, Japan
Kentaro Inui	Tohoku University, Japan
Yoshinobu Kano	Shizuoka University, Japan
Yusuke Miyao	National Institute of Informatics, Japan
Junichiro Mori	University of Tokyo, Japan
Hidetsugu Nanba	Hiroshima City University, Japan
Shoshin Nomura	National Institute of Informatics, Japan
Ken Satoh	National Institute of Informatics, Japan
Hiroyuki Shindo	Nara Institute of Science and Technology, Japan
Yoshimasa Tsuruoka	University of Tokyo, Japan
Minh Le Nguyen	Japan Advanced Institute of Science and Technology, Japan
Pontus Stenetorp	University College London, UK

kNeXI 2017**Workshop Chair**

Satoshi Nishimura	National Institute of Advanced Industrial Science and Technology, Japan
-------------------	--

Workshop Co-chairs

Takuichi Nishimura	National Institute of Advanced Industrial Science and Technology, Japan
Ken Fukuda	National Institute of Advanced Industrial Science and Technology, Japan

Program Committee

Satoshi Nishimura	National Institute of Advanced Industrial Science and Technology, Japan
Takuichi Nishimura	National Institute of Advanced Industrial Science and Technology, Japan
Ken Fukuda	National Institute of Advanced Industrial Science and Technology, Japan
Kentaro Watanabe	National Institute of Advanced Industrial Science and Technology, Japan
Yasuyuki Yoshida	National Institute of Advanced Industrial Science and Technology, Japan
Nami Iino	National Institute of Advanced Industrial Science and Technology, Japan

Sponsored By

The Japan Society for Artificial Intelligence (JSAI)



Contents

JURISIN2017

Analysis of COLIEE Information Retrieval Task Data	5
<i>Masaharu Yoshioka</i>	
From Case Law to Ratio Decidendi	20
<i>Josef Valvoda and Oliver Ray</i>	
Textual Entailment in Legal Bar Exam Question Answering Using Deep Siamese Networks	35
<i>Mi-Young Kim, Yao Lu, and Randy Goebel</i>	

SKL2017

A Study on Intellectual Tasks Influenced by the Embodied Knowledge	51
<i>Itsuki Takiguchi and Akinori Abe</i>	

AI-Biz2017

Agent-Based Simulation for Evaluating Signage System in Large Public Facility Focusing on Information Message and Location Arrangement	67
<i>Eriko Shimada, Shohei Yamane, Kotaro Ohori, Hiroaki Yamada, and Shingo Takahashi</i>	
Developing an Input-Output Table Generation Algorithm Using a Japanese Trade Database: Dealing with Ambiguous Export and Import Information	83
<i>Takaya Ohsato, Kaya Akagi, and Hiroshi Deguchi</i>	
Stock Price Prediction with Fluctuation Patterns Using Indexing Dynamic Time Warping and k^* -Nearest Neighbors	97
<i>Kei Nakagawa, Mitsuyoshi Imamura, and Kenichi Yoshida</i>	
Characterization of Consumers' Behavior in Medical Insurance Market with Agent Parameters' Estimation Process Using Bayesian Network	112
<i>Ren Suzuki, Yoko Ishino, and Shingo Takahashi</i>	
Do News Articles Have an Impact on Trading? - Korean Market Studies with High Frequency Data	129
<i>Sungjae Yoon, Aiko Suge, and Hiroshi Takahashi</i>	

Detecting Short-Term Mean Reverting Phenomenon in the Stock Market and OLMAR Method	140
<i>Kazunori Umino, Takamasa Kikuchi, Masaaki Kunigami, Takashi Yamada, and Takao Terano</i>	
A Study on Technology Structure Clustering Through the Analyses of Patent Classification Codes with Link Mining.	157
<i>Masashi Shibata and Masakazu Takahashi</i>	
LENLS 14	
Relating Intensional Semantic Theories: Established Methods and Surprising Results	171
<i>Kristina Liefke</i>	
Expressive Small Clauses in Japanese	188
<i>Yu Izumi and Shintaro Hayashi</i>	
Pictorial and Alphabet Writings in Asymmetric Signaling Games	200
<i>Liping Tang</i>	
Collecting Weighted Coercions from Crowd-Sourced Lexical Data for Compositional Semantic Analysis.	214
<i>Mathieu Lafourcade, Bruno Mery, Mehdi Mirzapour, Richard Moot, and Christian Retoré</i>	
How Dogwhistles Work.	231
<i>R. Henderson and Elin McCready</i>	
Transformational Semantics on a Tree Bank.	241
<i>Oleg Kiselyov</i>	
Derived Nominals and Concealed Propositions	253
<i>Iliara Frana and Keir Moulton</i>	
Denials and Negative Emotions: A Unified Analysis of the Cantonese Expressive <i>Gwai2</i>	266
<i>Grégoire Winterstein, Regine Lai, and Zoe Pei-sui Luk</i>	
Evidentials in Causal Premise Semantics: Theoretical and Experimental Investigation	282
<i>Yurie Hara, Naho Orita, and Hiromu Sakai</i>	
Annotating Syntax and Lexical Semantics With(out) Indexing	299
<i>Alastair Butler and Stephen Wright Horn</i>	
Discontinuity in Potential Sentences in Japanese	314
<i>Hiroaki Nakamura</i>	

AAA 2017

Invited Talk: Structured Engineering Argumentation 335
Robin E. Bloomfield

Invited Talk: Computational Persuasion with Applications
in Behaviour Change 336
Anthony Hunter

SCIDOCA 17

A Hierarchical Neural Extractive Summarizer for Academic Papers 339
Kazutaka Kinugawa and Yoshimasa Tsuruoka

Leveraging Document-Specific Information for Classifying Relations
in Scientific Articles 355
Qin Dai, Naoya Inoue, Paul Reisert, and Kentaro Inui

kNeXI2017

Investigating Classroom Activities in English Conversation Lessons Based
on Activity Coding and Data Visualization 375
*Zilu Liang, Satoshi Nishimura, Takuichi Nishimura,
and Mario Alberto Chapa-Martell*

On-site Knowledge Representation Tool for Employee-Driven
Service Innovation 390
Kentaro Watanabe

Consideration of Application Cases of Structured Manual
and Its Utilization 401
Satoshi Nishimura, Ken Fukuda, and Takuichi Nishimura

Author Index 415

JURISIN2017

Juris-Informatics (JURISIN) 2017

Ken Satoh

National Institute of Informatics, Japan

The Eleventh International Workshop on Juris-Informatics (JURISIN 2017) was held with a support of the Japanese Society for Artificial Intelligence (JSAI) in association with JSAI International Symposia on AI (JSAI-isAI 2017). JURISIN was organized to discuss legal issues from the perspective of informatics. Compared with the conventional AI and law, the scope of JURISIN covers a wide range of topics, which includes model of legal reasoning, argumentation/negotiation/argumentation agent, legal term ontology, formal legal knowledge-base/intelligent management of legal knowledge-base, translation of legal documents, information retrieval of legal texts, computer-aided law education, use of Informatics and AI in law, legal issues on applications of robotics and AI to society social implications of use of informatics and AI in law, natural language processing for legal knowledge, verification and validation of legal knowledge systems and any theories and technologies which is not directly related with juris-informatics but has a potential to contribute to this domain.

Thus, the members of Program Committee (PC) are leading researchers in various fields: Thomas Ágotnes (University of Bergen, Norway), Ryuta Arisaka (National Institute of Informatics, Japan), Kristijonas Cyras (Imperial College, UK), Marina De Vos (University of Bath, UK), Phan Minh Dung (AIT, Thailand), Randy Goebel (University of Alberta, Canada), Guido Governatori (NICTA, Australia), Tatsuhiko Inatani (Kyoto University, Japan), Tokuyasu Kakuta (Chuo University, Japan), Yoshinobu Kano (Shizuoka University, Japan), Mi-Young Kim (University of Alberta, Canada), Nguyen Le Minh (Japan Advanced Institute of Science and Technology, Japan), Beishui Liao (Zhejiang University, China), Hatsuru Morita (Tohoku University, Japan), Makoto Nakamura (Nagoya University, Japan), Katumi Nitta (Tokyo Institute of Technology, Japan), Paulo Novais (University of Minho, Portugal), Julian Padget (University of Bath, UK), Ginevra Peruginelli (ITTIG-CNR, Italy), Seiichiro Sakurai (Meiji Gakuin University, Japan), Katsuhiko Sano (Hokkaido University, Japan), Giovanni Sartor (EUI/CIRSFID, Italy), Ken Satoh (National Institute of Informatics and Sokendai, Japan), Akira Shimazu (Japan Advanced Institute of Science and Technology, Japan), Fumio Shimo (Keio University, Japan), Satoshi Tojo (Japan Advanced Institute of Science and Technology, Japan), Katsuhiko Toyama (Nagoya University, Japan), Rob van den Hoven van Genderen (VU University Amsterdam, The Netherlands), Leon van der Torre (University of Luxembourg, Luxembourg), Bart Verheij (University of Groningen, The Netherlands), Katsumasa Yoshikawa (IBM Research Tokyo, Japan), Masaharu Yoshioka (Hokkaido University, Japan), Harumichi Yuasa (Institute of Information Security, Japan), Yueh-Hsuan Weng (Tohoku University, Japan), and Adam Wyner (University of Aberdeen, UK). The collaborative work of computer scientists, lawyers and

philosophers is expected to contribute to the advancement of juris-informatics and it is also expected to open novel research areas.

Fifteen papers were submitted to JURISIN 2017. Each paper was reviewed by at least three members of PC. Thirteen papers were accepted in total. The collection of papers covers various topics such as legal reasoning, argumentation theory, impact of informatics and AI application into the society, application of natural language processing and so on. As invited speakers, Professor Katsumi Nitta from Tokyo Institute of Technology, Japan gave a talk on “Development of Argumentation Agent”, and Professor Kevin Ashley from The University of Pittsburg, USA, gave a talk on “Mining Information from Statutory Texts: A Case Study”. Moreover, we have a joint invited talk with AAA2017 Workshop whose speaker is Professor Anthony Hunter from University College London, UK talking on “Computational Persuasion with Applications in Behaviour Change”.

After the workshop, seven papers were submitted for the post proceedings. They were reviewed by PC members again and four papers were finally selected, but one paper was withdrawn. Followings are their synopses. Masaharu Yoshioka gave a detailed analysis of retrieval tasks studied in the international competition on legal information extraction/entailment (COLIEE 2017). Josef Valvoda and Oliver Ray proposed a method of extracting main judgement called “Ratio Decidendi” from UK legal cases. Mi-Young Kim, Yao Lu and Randy Goebel introduced their implementation of a Siamese deep Convolutional Neural Network for textual entailment in the question answering process. They evaluated their system using the data from the competition on legal information extraction/entailment (COLIEE).

Finally, we wish to express our gratitude to all those who submitted papers, PC members, discussant and attentive audience.



Analysis of COLIEE Information Retrieval Task Data

Masaharu Yoshioka^(✉)

Graduate School of Information Science and Technology, Hokkaido University,
N-14 W-9, Kita-ku, Sapporo 060-0814, Japan
yoshioka@ist.hokudai.ac.jp

Abstract. The Competition on Legal Information Extraction/Entailment (COLIEE) involves the legal question answering task. The information retrieval task for finding relevant articles to questions is one of the subtasks in COLIEE. In this paper, we compare the characteristics of the test data provided in two different language (English and Japanese) and analyze topic difficulty based on the submission data by using the retrieval results of Indri, a state-of-the-art information retrieval system. We also discuss issues relating to the design of new COLIEE information retrieval tasks.

1 Introduction

The Competition on Legal Information Extraction/Entailment (COLIEE) is a series of competitions that explore issues related to legal information extraction and entailment [1, 2]. The purpose of competition tasks is to design a legal information entailment system to answer questions in the Japanese bar exam (Civil Code). Competition tasks are divided into two phases: information retrieval (IR) for retrieving relevant articles for entailment and entailment using relevant articles (answering yes/no to exam questions).

In this paper, we focus on issues related to the IR task. The aim of this task is to find relevant Japanese Civil Code articles to check whether questions are true statements or not. Various techniques, such as least squares method, ranking SVM, neural net, WordNet, distributed representation, legal terminology, and n-gram words, have been used for constructing IR systems [3–8]. With each competition, the performance of systems according to F-measure increases slightly.

However, several issues have been raised in the evaluation of these systems. One relates to the variation of questions [3, 8]. For example, there are questions where the wording is similar to the articles themselves. Those questions are thus easy to retrieve by using any type of IR system. In contrast, there are types of questions that require matching of concrete example words against abstract legal terminology (e.g., “17 years old” and “minors”). For those questions, it is necessary to introduce other knowledge resources to estimate the relationships. Other issues concern the language [8]. The original Japanese civil laws came into

operation in 1890 and are revised every year. In contrast, the English version of the laws is translated from a specific version of Japanese civil laws. Even though the English version is also updated every year, it has a more uniform style description than the Japanese version. However, such issues concerning language have not been thoroughly investigated.

In this paper, to investigate the characteristics of the COLIEE IR task and the state-of-the-art of proposed IR systems, we analyze all submitted runs for the COLIEE IR task by using information of the baseline IR system Indri. Indri is a state-of-the-art IR system based on language modeling. Since Indri employs phrase matching that is widely used in the submitted system as word n-grams (sequence of words), the performance of the system is better or almost equivalent to the baseline system using TF-IDF for English data reported in the overview papers (baseline for COLIEE 2016 [1] and UA-TFIDF for COLIEE 2017 [2]).

2 COLIEE IR Task

2.1 Task Description

The COLIEE IR task retrieves a static set of relevant civil code articles for answering the question given as a query. Relevant articles are ones that are necessary to check the appropriateness of the given question statement. There are questions that require only one article to check the appropriateness (one relevant article) and ones that require multiple articles (multiple relevant articles).

The participant systems are requested to return relevant article sets for the question test data. The performance of the system is evaluated by F-measure (mean average of precision and recall). In recent years, four competitions have been conducted. Table 1 gives a summary of the information concerning the number of questions (number of questions with multiple relevant articles) and the number of relevant articles in total. (In each competition, a set of questions from the bar exam of the same year was used: H22 (2010), H24 (2012), H26 (2014), and H28 (2016) for COLIEE 2014, 2015, 2016, and 2017, respectively.¹) Most of the questions have only one relevant article and less than one-third of questions have multiple (two or more) relevant articles.

Table 1. Test data

Year	Number of questions (with multiple relevant articles)	Number of relevant articles
H22	47 (3)	51
H24	79 (12)	102
H26	95 (29)	131
H28	78 (20)	110

¹ One COLIEE competition uses one year bar exam data (e.g., H22 for COLIEE 2014) for IR task and another one year for entailment (e.g. H23(2011) for COLIEE 2014).

2.2 Submitted Runs

Run is a retrieval results (list of candidate relevant articles for each question) submitted by a participant. Each participating team can submit multiple runs as candidates for evaluation.²

Table 2 presents information about the runs. Except for COLIEE 2017 [2], there is no clear description about the language used for generating runs. However, after checking the papers, we found that most of the submissions used English data except for Kanolab (H24), HUKB (H26 and H28), and KIS (H28).

Table 2. Number of submissions (teams runs)

Year	English		Japanese	
	Runs	Teams	Runs	Teams
H22	1	1	0	0
H24	6	4	6	1
H26	10	6	4	1
H28	11	5	5	2

In this analysis, since the characteristics of the runs from the same team are similar, we use the best performance runs for each team for further analysis.

2.3 Indri

Indri is a state-of-the-art IR system based on language modeling [9]³. Indri is used in varieties of IR tasks including web retrieval [10], blog retrieval [11]. Retrieval performance of Indri is good even for the out-of-box Indri implementation [11]. Indri uses phrase-based queries⁴ that are useful for identifying legal terminology represented as compound terms. In this experiment, we do not conduct any parameter tuning for this task (out-of-box setting same as [11]). Figure 1 shows the procedures to translate a question query into one for Indri in English and Japanese.

For both languages, the original question queries are translated by using the operator “#combine” for representing all combinations of the keywords in the query are used for calculating the similarity. In addition, for English, words in the stop-words list⁵ are excluded from the index and the Krovetz stemmer is used for normalizing terms in the documents and queries. For Japanese, MeCab morphological analyzer with original dictionary (no special tuning for handling

² Since participants didn’t have information about the relevant articles, they can submit multiple runs with different settings as candidates for evaluation.

³ <https://www.lemurproject.org/indri/>.

⁴ Phrase-based queries can take into account the word ordering in the query.

⁵ <http://www.lemurproject.org/stopwords/stoplist.dft>.

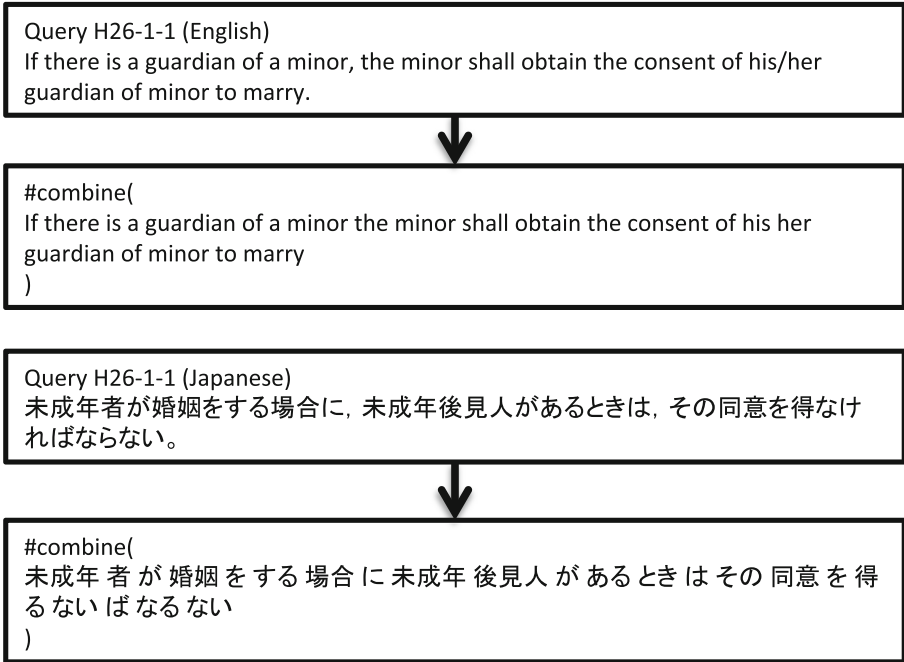


Fig. 1. Indri query construction

legal terms) is used to normalize and segment a query text into a sequence of normalized Japanese words and no stop-words are used. To construct a retrieval result, a top-ranked retrieved document for each query is returned as a relevant article (this approach is the most commonly used among all submitted data).

Table 3 shows the system performance of Indri for the test data. From this table, it is clear that the difficulty of Japanese and English test data differs year by year.

Table 3. Retrieval performance of Indri (English and Japanese)

Year	English			Japanese		
	F	Precision	Recall	F	Precision	Recall
H22	0.551	0.574	0.529	0.735	0.766	0.706
H24	0.486	0.557	0.431	0.409	0.468	0.363
H26	0.496	0.589	0.427	0.460	0.547	0.397
H28	0.585	0.705	0.500	0.436	0.526	0.372

3 Analysis of the COLIEE Competition Data

In this analysis, we use ranked results of Indri retrieval results as a feature to show the similarity between a given question and a related article or articles. This feature allows the classification of the question type based on its difficulty.

3.1 Comparison Between English and Japanese Test data

First, we compare the rank of relevant articles of the English and Japanese test data (Table 4). From this table, we can see that the correlations between the ranks of English and Japanese test data are similar, although there are a few cases where the results are completely different. H24-27-O⁶ is a case that Japanese data is easier than English one (rank 18 for Japanese and rank 495 for English). On the contrary, H26-27-3 is a case that English one is easier than Japanese one (article 650: rank 9 for English and rank 678 for Japanese; article 701: rank 300 for English and 98 for Japanese; article 702 rank 561 for English and 114 for Japanese). Underlines highlight the corresponding parts among question and article pairs.

- H24-7-O: In cases where (A) and (B) is a married couple, if A engages in juristic act with (C) in the single name of oneself regarding everyday household matters, the extinctive prescription of C’s credit against B shall be nullified upon the judicial claim of C’s credit against B.

AとBが夫婦の場合、Aが自己の単独名義でCと日常の家事に関して契約を締結して債務を負ったとき、CのAに対する債権の裁判上の請求により、CのBに対する債権の消滅時効も中断する。

Answer: 434

(Request for Performance to One Joint and Several Obligor)

Article 434 A request for performance made to one joint and several obligor shall also be effective with respect to other joint and several obligor(s).

(連帯債務者の一人に対する履行の請求)

第四百三十四条 連帯債務者の一人に対する履行の請求は、他の連帯債務者に対しても、その効力を生ずる。

- H26-27-3: In cases where A found a collared dog whose owner is unknown, and took care of it for the unknown owner; A took the dog to his/her house and took care of it, then the dog pushed down a vase on a shoe

⁶ Capital letter such as “A”, “B”, and “C” are used for anonymize original name in Japanese judicial precedent and also be used in the exam.

cupboard and broke it. In this case, if A was free from any negligence, A may claim compensation for the loss from the owner of the dog.⁷

Aが首輪の付いている飼い主不明の犬を発見し、その不明の飼い主のために犬の世話をした場合に関する次のアからオまでの各記述のうち、正しいものを組み合わせたものは、後記1から5までのうちどれか。Aが自分の家に犬を連れて帰り、世話をしていたところ、犬が下駄箱の上に置かれていた花瓶を倒し、壊してしまった。この場合、Aに過失がなかったとすると、Aは犬の飼い主に対して損害賠償を請求することができる。

Answer: 650,701,702

(Mandatory's Claims for Reimbursement of Expense)

Article 650 (1) If the mandatory has incurred costs found to be necessary for the administration of the mandated business, the mandatory may claim reimbursement of those costs from the mandator and any interest on the same from the day the costs were incurred.

(2) If the mandatory has incurred any obligation found to be necessary for the administration of the mandated business, the mandatory may demand that the mandator perform the obligation on the mandatory's behalf. In such cases, if the obligation has not yet fallen due, the mandatory may require the mandator to tender reasonable security.

(3) If the mandatory suffers any loss due to the administration of the mandated business without negligence in the mandatory, he/she may claim compensation for the loss from the mandator.

(受任者による費用等の償還請求等)

第六百五十条 受任者は、委任事務を処理するのに必要と認められる費用を支出したときは、委任者に対し、その費用及び支出の日以後におけるその利息の償還を請求することができる。

2 受任者は、委任事務を処理するのに必要と認められる債務を負担したときは、委任者に対し、自己に代わってその弁済をすることを請求することができる。この場合において、その債務が弁済期にないときは、委任者に対し、相当の担保を供させることができる。

3 受任者は、委任事務を処理するため自己に過失なく損害を受けたときは、委任者に対し、その賠償を請求することができる。

For the first case, “債務者” are translated into “obligor” in the article and “債務”(obligation) in the question as “credit”. However, Japanese system can use “債務”(obligation) as a part of “連帯債務者”(joint obligor) that are split as “連帯”(joint), “債務”(obligation), “者”(person; suffix) using Japanese Morphological analyzers.

⁷ A part of Japanese questions include description about the explanation of their question types and this part is not translated into English one. For example, Japanese question has “次のアからオまでの各記述のうち、正しいものを組み合わせたものは、後記1から5までのうちどれか。” (There are five descriptions about the legal decision ア (a) to オ (o). Please select a combination of correct description(s) from 1 to 5 below.), but no corresponding description in English one.

For the latter case, the style of writing for article 650 (3) and questions are very similar (share long phrases) in English and not for Japanese. Another problem is related to description about the explanation of their question types. When we use the question that removes this explanation: “Aが首輪の付いている飼い主不明の犬を発見し、その不明の飼い主のために犬の世話をした場合。Aが自分の家に犬を連れて帰り、世話をしていたところ、犬が下駄箱の上に置かれていた花瓶を倒し、壊してしまった。この場合、Aに過失がなかったとすると、Aは犬の飼い主に対して損害賠償を請求することができる。”, the rank of the article 650 rise to 100.

Table 4. Comparison of rank of relevant articles in Japanese and English

J	E										
	1	2	3	4	5	-10	-30	-50	-100	101-	Sum
1	145	10	6	1	1	3	0	0	0	0	166
2	18	13	6	2	0	2	3	0	0	0	44
3	7	6	4	2	0	3	2	0	0	0	24
4	2	3	3	2	4	1	2	0	1	0	18
5	1	0	0	3	3	2	0	0	0	0	9
-10	4	3	3	3	2	5	4	0	0	1	25
-30	5	0	1	0	6	5	16	2	2	3	40
-50	0	1	0	0	0	1	11	3	3	1	20
-100	0	0	0	0	0	1	0	4	7	6	18
101-	0	0	0	0	0	1	0	2	4	23	39
Sum	182	36	23	13	16	24	38	11	17	34	394

Table 4 shows the comparative analysis results between the system based on English test data and Japanese one. Columns of the table represents rank for the English test data of the relevant articles and lows represents ones for Japanese test data. Most of the top one ranked questions for the English are also ranked as top for the Japanese, but there are several exceptional cases.

Difficult Japanese ones.

Based on the results of Tables 3 and 4, it is therefore necessary to analyze the runs for English and Japanese data separately.

3.2 Analysis of English Test Data

The numbers of teams that used English data for each run (H22(2010), H24(2012), H26(2014), and H28(2016)) were 1, 4, 6, and 5, respectively. For Japanese data, the numbers were 0, 1, 1, 2, respectively. Since the numbers of Japanese runs are small, we will first focus on the English runs.

Table 5 shows the cross table of Indri rank (English) for all relevant articles and the number of team best performance runs that find those articles. First and second row of Table 5 represent test data used in the competition and number of team best performance runs that find those articles (maximum number of teams are different based on the number of participant teams for each test data 1, 4, 6, 5 for H22, H24, H26, H28) respectively. When there is no team that can be retrieve relevant articles, it was categorized as 0 (no team). From this table, we confirmed questions with Indri rank 1 are likely to be retrieved by many teams, but questions with Indri rank lower than 100 are not retrieved by any teams.

Table 5. Cross table of Indri rank (English) and number of team best performance runs that find relevant articles

Indri rank	H22		H24						H26						H28					
	0	1	0	1	2	3	4	0	1	2	3	4	5	6	0	1	2	3	4	5
1	1	26	1	0	6	19	18	0	2	0	5	4	24	21	2	2	3	7	6	35
2	1	6	1	1	5	3	2	4	1	1	0	2	0	1	2	2	1	1	1	1
3	2	2	3	2	1	0	0	0	2	1	0	0	2	0	3	1	2	1	1	0
4	0	0	3	1	1	0	0	1	0	2	1	2	0	0	0	1	0	1	0	0
5	0	0	1	1	0	0	0	2	6	1	0	0	0	0	2	1	1	1	0	0
-10	1	2	3	0	0	0	0	7	2	1	0	0	1	0	5	0	1	0	0	1
-30	2	1	8	2	0	0	0	12	2	1	0	0	0	0	7	2	1	0	0	0
-50	1	0	4	1	0	0	0	1	0	0	0	0	0	0	4	0	0	0	0	0
-100	3	0	3	1	0	0	0	4	1	0	0	0	0	0	5	0	0	0	0	0
101-	3	0	11	0	0	0	0	14	0	0	0	0	0	0	6	0	0	0	0	0
Sum	14	37	38	9	13	22	20	45	16	7	6	8	27	22	36	9	9	11	8	37

Since most of the runs return only one article per question, the second- and third-ranked relevant articles are not likely to be retrieved. Therefore, most of the second- and third-ranked relevant articles are not retrieved by all runs regardless of Indri rank. To reduce the effect of questions with multiple answers, Table 6 shows the cross table of Indri rank (English) for the top-ranked relevant article for each question and the number of team best performance runs that find those articles. From this table, we see that there are several questions whose retrieved relevant articles differ according to the submitted runs. For example, in the case of Indri top-ranked questions in H28 data, the number of questions that are retrieved by all teams (five) increases from 35 to 38, because the following questions have two different relevant articles that are retrieved by the different systems.

- H28-11-5 (The statutory lien over movables shall prevail over the pledge of movables.): Only Indri found 339 and 5 other runs found 334.

- H28-16-1 (Before the principal is fixed, a revolving mortgagee may assign a Revolving Mortgage or effect a partial assignment of the Revolving Mortgage, without the approval of the revolving mortgagor.): Indri and 3 other runs found 398-12 and 2 other runs found 398-13.
- H28-24-4 (In cases where there is any latent defect in the subject matter of a sale, if the buyer demands compensation for damages based on a warranty against defects within one year from the time when he/she came to know the existence of the defect, extinctive prescription of the right to demand compensation for damages shall not be completed even if 10 years have passed at that time since he/she received the delivery of the subject matter of the sale.): Indri and 2 other runs found 566 and 3 other runs found 570.

Those questions have multiple answers for similar contents. In contrast, the number of questions for Indri rank higher than 100 decreases, meaning that there are a certain number of questions whose multiple relevant articles are ranked lower than 100. We will discuss the characteristics of questions with multiple relevant articles in Sect. 3.4.

Table 6. Cross table of Indri rank (English) of the top-ranked relevant article and the number of team best performance runs that find at least one relevant article

Indri rank	H22		H24					H26								H28					
	10	1	0	1	2	3	4	0	1	2	3	4	5	6	0	1	2	3	4	5	
1	1	26	1	0	6	19	18	0	0	1	5	2	24	24	1	2	2	6	6	38	
2	1	6	1	1	5	3	2	2	0	1	0	2	0	1	0	2	0	1	1	1	
3	2	2	1	2	1	0	0	0	1	1	0	0	1	0	2	0	3	0	1	0	
4	0	0	2	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	
5	0	0	0	1	0	0	0	1	4	1	0	0	0	0	0	0	0	1	0	0	
-10	1	2	0	0	0	0	0	5	2	1	0	0	1	0	1	0	1	0	0	0	
-30	2	1	3	1	0	0	0	3	1	1	0	0	0	0	4	0	1	0	0	0	
-50	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
-100	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
101-	2	0	7	0	0	0	0	5	0	0	0	0	0	0	1	0	0	0	0	0	
Sum	10	37	18	6	13	22	20	16	9	7	6	5	27	25	11	4	7	9	8	39	

Table 7 shows the cross table of Indri rank and the results of the best-run of all submission data using English collection for each year (H22: UA; H24: UA; H26: iLIS7; and H28: iLIS7 (Year: teamID)). Number of teams for second row of Table 7 is 0 (the best-run fails to retrieve the article) and 1 (the best-run succeeds to retrieve the article). From this table, it is clear that the best-run systems are effective at retrieving articles with higher Indri rank, but it is difficult for those systems to retrieve articles with lower Indri rank. The characteristics of those articles are discussed in Sect. 4.

Table 7. Cross table of Indri rank (English) and best-run of all submission data using English collection

Indri rank	H22		H24		H26		H28	
	0	1	0	1	0	1	0	1
1	1	26	5	39	2	54	5	50
2	1	6	5	7	6	3	6	2
3	2	2	3	3	2	3	4	4
4	0	0	4	1	2	4	1	1
5	0	0	2	0	5	4	4	1
-10	1	2	3	0	9	2	5	2
-30	2	1	10	0	13	2	9	1
-50	1	0	5	0	1	0	4	0
-100	3	0	4	0	5	0	5	0
101-	3	0	11	0	14	0	6	0
Sum	14	37	52	50	59	72	49	61

3.3 Analysis of Japanese Test Data

Since the number of teams using Japanese test data is just one—except for H28—and the second-best team in H28 is relatively worse than the best-run of all submission data and Indri, we use cross tables for the best run (Tables 8 and 9) only for its analysis. The general tendency of the results is almost identical to the English data.

Table 8. Cross table of Indri rank (Japanese) and best-run of all submission data using Japanese collection

Indri rank	H24		H26		H28	
	0	1	0	1	0	1
1	18	19	2	50	6	35
2	8	4	10	8	5	6
3	8	2	4	1	6	2
4	4	0	5	1	4	2
5	3	0	2	1	3	0
-10	4	0	7	1	10	2
-30	9	1	11	2	10	5
-50	5	0	9	0	3	0
-100	7	0	5	0	6	0
101-	10	0	12	0	5	0
Sum	76	26	67	64	58	52

Table 9. Cross table of Indri rank (Japanese) and best-run of all submission data that finds at least one relevant article

Indri rank	H24		H26		H28	
	0	1	0	1	0	1
1	16	21	2	50	4	37
2	6	2	5	8	4	5
3	7	2	4	0	4	2
4	3	0	1	1	0	3
5	2	0	2	1	2	0
-10	3	0	4	2	6	1
-30	5	1	5	1	4	3
-50	3	0	3	0	1	0
-100	3	0	3	0	2	0
101-	5	0	3	0	0	0
Sum	53	26	32	63	27	51

3.4 Analysis of Questions with Multiple answers

In the test data, some questions require multiple articles to check their appropriateness (Table 1). There are two types of such multiple relevant article sets: the first case discusses similar issues, for which all articles have similar Indri ranks; in the second case, a set contains one main article that discusses the main legal decision with supplemental articles to justify the decision. In this case, the supplemental articles have lower Indri rank.

Since more than half of the questions in the training data require only one article, most teams return one article only per question. In addition to the general procedures, a number of teams have special procedures to select additional articles. Therefore, there are only a few cases in which two relevant articles were successfully retrieved in the submitted runs. Moreover, no submitted runs successfully retrieved three or more articles for one question. The number of questions that successfully retrieved two articles/number of questions with multiple relevant articles is H22: 0/3, H24: 4/12, H26: 1/29, and H28: 4/20 for the English data and H24: 0/12, H26: 1/29, and H28: 1/20 for the Japanese data. For the latter, HUKB (the best-run for H26 and H28) used a special procedure to handle *mutatis mutandis* and found (H26-36-1 and H28-26-1). For the former, various kinds of techniques are introduced (e.g., handling *mutatis mutandis*, a machine learning method for analyzing the question to decide the number of relevant articles). However, the second-ranked number of retrieved articles has a higher Indri rank (rank:count, 2:2, 3:2, -10:1, -30:2, -50:1, -100(51):1) and most of them have reference between first- and second-ranked articles (6/9).

4 Discussion

From the analysis of the submitted runs, we confirmed that Indri rank can be used effectively to estimate the difficulty of questions and characteristics of the submitted runs. Especially for the questions whose relevant articles are highly ranked (but not ranked 1st). For example, several submitted runs improve the quality of retrieved results by introducing domain-oriented techniques (such as sentence structure analysis and legal terminology). For example, in the case of H28-1-5: “A minor may not become an agent”, Indri returns article 158, which includes the keywords “minor” and “agent” as first ranked, but two English runs (iLis7 [6] and JNLP [7]) found article 102 (relevant article), which discusses the requirement to be an agent as first ranked. Since this article does not have the term “minor” in the article, the Indri rank is 21 for English.

(Agent’s Capacity to Act)

Article 102 An agent need not to be a person with the capacity to act.

However, it is difficult to retrieve relevant articles with lower Indri rank. Even for best-run systems, it is difficult to find relevant articles for those whose Indri rank is lower than 5. Thus, even though many teams have introduced a number of semantic handling methods (e.g., WordNet [12] and word2vec [13]), they are not currently adequate to handle such semantic relationships. For the next COLIEE IR task, the results from this analysis should be considered for constructing test data and evaluation. For example, it may be better to introduce question type categories, according to which we can classify topics into the following categories.

Simple. Simple questions that do not require advanced techniques such as sentence structure analysis or semantic knowledge. Relevant articles for these questions may have a high Indri rank.

Example of this question is H24-16-4. Question in English is “If an inheritance of a revolving mortgagee commences before the principal is fixed, the principal secured by a Revolving Mortgage shall be fixed at the time of the commencement of the inheritance, unless there is the agreement between the heirs and the revolving mortgagor to success the Revolving Mortgage.”, and relevant article is Article 398-8.

(Inheritances of Revolving Mortgagees or Obligors)

Article 398-8 (1) If an inheritance of a revolving mortgagee commences before the principal is fixed, the Revolving Mortgage shall secure the claims that exist at the time of the commencement of the inheritance and shall otherwise secure claims the heir prescribed by agreement between the heirs and the revolving mortgagor acquires after the commencement of the inheritance.

(2) If an inheritance of an obligor commences before the principal is fixed, the Revolving Mortgage shall secure the obligations that exist at the time of the commencement of the inheritance and shall otherwise secure the claims that the heir prescribed by agreement between the revolving mortgagee and the revolving mortgagor assumes after the commencement of the inheritance.

– snip –

Since sentence (2) in the article is very similar to the question, it is easy to retrieve this document as relevant one.

Complex. Questions that require several techniques to find out one relevant answer. For example, there are questions that require sentence analysis to estimate the importance of the keyword to re-rank the ordinal IR system's result. An example of this category is H28-1-5 discussed above. Another type of questions are ones that require semantic matching. Relevant articles for these questions may not have a high Indri rank and they contain large numbers of words that are not used in articles. In addition, there are several questions that requires all of those techniques.

However, there are several questions whose Indri ranks are high even though they contain large numbers of words that are not used in articles. Example of the question is H26-10-3. Question in English is "A library of a university A has a book "P", and a professor B borrows the same and uses his/her office in the same campus. In the case where B left "P" in a train on the way home, and F found and possesses the same, B may claim for the restoration of "P" by bringing an action for recovery of possession against F." and relevant article is Article 200.

(Actions for Recovery of Possession)

Article 200 (1) When a possessor is forcibly dispossessed, he/she may claim for the restoration of the Thing and compensation for damages by bringing an action for recovery of possession.

(2) An Action for recovery of possession cannot be filed against a specific successor of the usurper of possession; provided, however, that this shall not apply if that successor had knowledge of the fact of usurpation.

Since they share good phrases "Action for recovery of possession" and Indri can retrieve this article as rank 1. However, in order to conduct entailment, it is necessary to deal with words that are not used in articles for analyzing the condition of this article is satisfied or not.

Multiple answers. Questions that have multiple relevant articles should be categorized into subcategories such as similar, reference, and supplemental.

The performance of systems should be evaluated by using the performance for questions of each category. In addition, it may be better to take into account the difference for each category.

Using of this category, we can summarize the status of legal IR systems developed in COLIEE competition as follows. At this moment, the best-run system for English and Japanese are good to retrieve "Simple" questions, but they are not good enough to retrieve "Complex" questions. Further failure analysis may be necessary to improve the retrieval performance of the systems. For the multiple answers, there are quite few attempts to return multiple answers. It is necessary to propose a method that is not used in the competition for handling all types of this question.

This analysis suggests that for the next series of the competition, it may be necessary to include more questions for "Complex" and "Multiple answers" and

compare retrieval performance of each category may help the participants to understand the characteristics of their systems.

It may be also meaningful to use this category for analyzing the entailment results. For example, for the questions that are difficult to retrieve relevant articles, most of the entailment results are derived from non-relevant articles. However, since answers for the entailment is true or false, there are many cases that systems find correct answers by chance for those questions. It is necessary to take into account such effect for the analysis.

5 Conclusion

In this paper, we proposed a method to analyze COLIEE IR test data by using the state-of-the-art IR system Indri. From this analysis, we confirmed that the best-run systems are good at handling simple questions and ones that require sentence structure analysis; however, they are not effective at retrieving documents that require semantic knowledge and in cases of multiple relevant articles. We also proposed a new criterion for designing the next COLIEE task to evaluate system performance for questions for which it is difficult to retrieve articles with current best-run systems.

Acknowledgment. We would like to thank organizers of COLIEE to provide submitted runs data for this analysis and would also like to thank all participants of COLIEE to provide valuable information. This work was partially supported by JSPS KAKENHI Grant Number 16H01756.

References

1. Kim, M.Y., Goebel, R., Kano, Y., Satoh, K.: COLIEE-2016: evaluation of the competition on legal information extraction and entailment. In: The Proceedings of the 10th International Workshop on Juris-Informatics (JURISIN2016), Paper 11 (2016)
2. Kano, Y., Kim, M.Y., Goebel, R., Satoh, K.: Overview of COLIEE 2017. In: Satoh, K., Kim, M.Y., Kano, Y., Goebel, R., Oliveira, T. (eds.) 4th Competition on Legal Information Extraction and Entailment, COLIEE 2017. EPiC Series in Computing, vol. 47, pp. 1–8. EasyChair (2017)
3. Kim, M.-Y., Xu, Y., Goebel, R.: Legal question answering using ranking SVM and syntactic/semantic similarity. In: Murata, T., Mineshima, K., Bekki, D. (eds.) JSAI-isAI 2014. LNCS (LNAI), vol. 9067, pp. 244–258. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48119-6_18
4. Kim, K., Heo, S., Jung, S., Hong, K., Rhim, Y.Y.: An ensemble based legal information retrieval and entailment system. In: The Proceedings of the 10th International Workshop on Juris-Informatics (JURISIN2016), Paper 11 (2016)
5. Onodera, D., Yoshioka, M.: Civil code article information retrieval system based on legal terminology and civil code article structure. In: The Proceedings of the 10th International Workshop on Juris-Informatics (JURISIN2016), Paper 19 (2016)

6. Heo, S., Hong, K., Rhim, Y.Y.: Legal content fusion for legal information retrieval. In: Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, pp. 277–281. ACM, New York (2017)
7. Carvalho, D.S., Tran, V., Tran, K.V., Minh, N.L.: Improving legal information retrieval by distributional composition with term order probabilities. In Satoh, K., Kim, M.Y., Kano, Y., Goebel, R., Oliveira, T. (eds.) 4th Competition on Legal Information Extraction and Entailment, COLIEE 2017. EPIc Series in Computing, vol. 47, pp. 43–56. EasyChair (2017)
8. Yoshioka, M., Onodera, D.: A civil code article information retrieval system based on phrase alignment with article structure analysis and ensemble approach. In: Satoh, K., Kim, M.Y., Kano, Y., Goebel, R., Oliveira, T. (eds.) 4th Competition on Legal Information Extraction and Entailment, COLIEE 2017. EPIc Series in Computing, vol. 47, pp. 9–22. EasyChair (2017)
9. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, pp. 2–6 (2005)
10. Metzler, D., Strohman, T., Croft, W.: Indri at trec 2006: lessons learned from three terabyte tracks. In: Proceedings of the Text REtrieval Conference (2006)
11. Ernsting, B., Weerkamp, W., de Rijke, M., et al.: Language modeling approaches to blog post and feed finding. In: Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007) (2007)
12. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**, 39–41 (1995)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)



From Case Law to Ratio Decidendi

Josef Valvoda^(✉) and Oliver Ray

Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK
{jv16618, csxor}@bristol.ac.uk

Abstract. This paper is concerned with the task of automatically identifying legally binding principles, known as *ratio decidendi* or just *ratio*, from transcripts of court judgements, also called *case law* or just *cases*. After briefly reviewing the relevant definitions and previous work in the area, we present a novel system for automatically extracting ratio from cases using a combination of natural language processing and machine learning. Our approach is based on the hypothesis that the ratio of a given case can be reliably obtained by identifying statements of legal principles in paragraphs that are cited by subsequent cases. Our method differs from related recent work by extracting principles from the text of the *cited* paragraphs (in the given case) as opposed to the text of the *citing* paragraphs (in a subsequent case). We conduct our own independent small-scale annotation study which reveals that this seemingly subtle shift of focus substantially increases reliability of finding the ratio. Then, by building on previous work in the automatic detection of legal principles and cross citations, we present a fully automated system that successfully identifies the ratio (in our study) with an accuracy of 72%.

Keywords: Ratio decidendi · Case law · Natural language processing
Machine learning · Principle detection · Cross reference resolution

1 Introduction

In common law *ratio decidendi*, or just *ratio* [2], are the key principles used to decide the outcome of a legal case and the doctrine of *stare decisis*, or simply *precedent* [5], requires that ratio is applied on subsequent cases with similar facts to decide their outcome. This is different from the law used in civil law countries, where the legal principles are spelled out directly in statutes. Thus the identification of ratio is crucial to the work of lawyers and judges in common law countries such as the UK.

To find the ratio of a case of interest, lawyers must typically perform a detailed analysis of the text of the given case along with the text of key citing and cited cases and other cases related by topic [9]. Although legal search engines, such as Westlaw¹, are typically used to find the related cases relatively efficiently, there is currently a lack of technological support for the task of actually identifying the ratio contained within the transcripts of those cases. While

¹ Westlaw UK, Online legal research from Sweet & Maxwell, <http://westlaw.co.uk>.

there are human generated summaries of a case available, these are not focused on the ratio and provide only a simplified overview. Since a single case can consist of more than 150 paragraphs spread over 30 pages of text, the task of finding a ratio is an extremely time-consuming and error-prone process for humans to perform.

To automatically identify ratio, two issues must be addressed. The first issue is to distinguish statements of legal principles from discussion of specific facts to which those principles are applied in a particular case. The second issue is to determine which of those principles are pivotal to the outcome of the case (and therefore constitute the legally binding ratio), as opposed to those principles which are merely incidental and which are formally known as *obiter dicta*, or just *obiter* [2]. Automating these processes would be of great value to lawyers.

This paper presents a novel system for automatically extracting ratio from cases using a combination of natural language processing and machine learning to solve the two issues noted above. To achieve this we essentially build upon and integrate the recent work of Shulayeva et al. on the detection of principles and facts in case law [8] and the earlier work of Adedjouma et al. on cross reference resolution in statutory law [1].

The contribution of our work stems from several differences from Shulayeva et al. [8]. Whereas Shulayeva et al. specifically emphasise that they are *not* attempting to identify ratio, we specifically emphasise that we *are*. Whereas Shulayeva et al. are primarily concerned with the distinction between cited *facts* and *principles*, we are primarily concerned with the distinction between *orbiter* and *ratio*. Whereas the method of Shulayeva et al. is mainly concerned with analysing *citing* paragraphs, our method is mainly concerned with analysing *cited* paragraphs.

To illustrate these differences and demonstrate the effectiveness of our approach, we conducted our own independent small-scale annotation study which supports our hypothesis that the ratio of a given case can be reliably obtained by identifying the statements of legal principles in paragraphs that are cited by subsequent cases. In fact, our investigation shows that the principles in cited paragraphs correspond to ratio (as manually determined by a human expert using Wambaugh’s Inversion Test [2]) with a precision of 76%, whereas the principles extracted by Shulayeva et al. from the citing paragraphs correspond to ratio with an accuracy of only 68%.

In order to automate our new approach we do three things. First we build upon Shulayeva et al.’s methodology for principle identification in order to achieve a 96% accuracy on this individual task. Second, inspired by Adedjouma’s research on cross reference identification and resolution, we develop our own legal text schema to achieve an accuracy of 94% on the individual task of cited paragraph identification. Third, by combining the two classifiers above, we demonstrate a fully automatic system that successfully identifies ratio with 72% accuracy.

We conclude the paper by discussing how the bar we have set can be potentially raised in future work.

2 Background

In this section we explain our terminology and the distinction between ratio and obiter as discussed in the legal informatics literature [2]. We further describe Shulayeva et al.’s work concerned with automated fact and principle identification [8], Saravanan et al.’s work on case summarisation [7] and Adedjouma et al.’s work on cross reference resolution [1].

2.1 Citing vs. Cited: Cases, Paragraphs and Principles

When referring to paragraphs in legal cases connected by a citation we will use the terminology illustrated in Fig. 1, which is a natural extension of the nomenclature used in Westlaw. Suppose a paragraph (A) in a citing case (X) contains a citation to some paragraph (B) in cited case (Y). Then the former is called the *citing paragraph* (A) while the latter is called the *cited paragraph* (B).

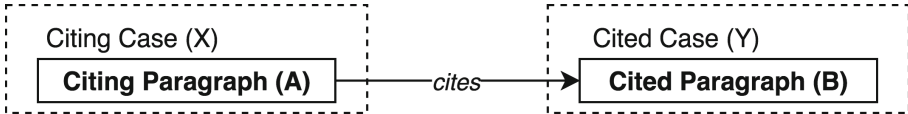


Fig. 1. The distinction between citing cases and paragraphs.

2.2 Defining Ratio via Wambaugh’s Inversion Test

Defining ratio decidendi and establishing a test for finding it in case law is essential for ratio identification. There are many different legal interpretations on what ratio actually is [5,6]. Branting’s research elegantly summarises the differing points of view on the matter [2]. He identifies two general areas of focus. One is on the material facts, the other on the deciding principles. These further translate into five different viewpoints [2]:

1. *The ratio decidendi of a precedent consists of propositions of law explicit or implicit in the opinion that are necessary to the decision.*
2. *A unique proposition of law necessary to a decision can seldom be determined. Instead a gradation of propositions ranging in abstraction from the specific facts of the case to abstract rules can satisfy this condition.*
3. *The ratio decidendi of a precedent must be grounded in the specific facts of the case.*
4. *The ratio decidendi of a precedent includes not only the precedent’s material facts and decisions but also the theory under which the material facts lead to the decision.*
5. *Subsequent decisions can limit, extend, overturn earlier precedents.*

In this paper we choose the first definition from above, since it can be tested for using Wambaugh’s Inversion Test [2]. Under this test, a principle in the case is inverted in its meaning; and if such inversion would affect the outcome of the case, then the principle is deemed a ratio. On the other hand, if the inversion would not affect the outcome, the principle is deemed an obiter. This test is widely used by legal practitioners and lends our research a solid grounding. Note that the outcome of the case is not only the immediate application of the law used in the case (i.e. which party won), but also the points of law established by reaching the conclusion (i.e. the legal precedent set by the case).

We now illustrate the Inversion Test by applying it to the landmark case of *Stack v Dowden* (commonly referred to as just *Stack*) concerned with division of family property after the breakup of a (unmarried) cohabiting couple. The question was whether the parties held the property as beneficial joint tenants, entitling them each to half of the property, or whether their conduct suggests they held the property otherwise, entitling the court to divide the property unequally. To decide on this the court needed to establish what should be considered to determine the parties intention. Thus, applying the Inversion Test on a sentence: *“The search is to ascertain the parties’ shared intentions, actual, inferred or imputed, with respect to the property in the light of their whole course of conduct in relation to it.”* identifies this sentence as a ratio. This is because Lady Hale employs the principle in this sentence, considering the parties conduct as a whole, to ultimately divide the shares in the property 65 to 35%, departing from the default legal position. Reversing the logic of this sentence would therefore disrupt the outcome of the case.

2.3 Automatic Case Summarisation

As part of their work on text summarisation Saravanan et al. consider the identification of ratio as one of seven different categories which they use to classify sentences in judgements: *“1. Identifying the case, 2. Establishing the facts of the case, 3. Arguing the case, 4. Arguments, 5. Ratio of the decision, 6. History of the case, 7. Final decision”* [7]. Upon first look, it might be tempting to compare our respective work on ratio identification.

However, Saravanan et al. are working with a very different case law to us. Where we focus on long English precedent setting cases from higher courts, they focus on shorter applicational cases from the Indian law. Perhaps because the cases they work with are much shorter, the example they provide only contains six paragraphs and a single judgement while our cases contain upwards of hundred and fifty paragraphs and five judges², they classify individual sentences whereas we classify paragraphs.

Because we are looking at such radically different case law to Saravanan et al., the search for the ratio is also a completely different task. This is illustrated

² Unfortunately, we could not access the data corpus used by Saravanan et al. since the links in their paper are out of service and we have not received a reply to our email asking for additional information. Thus, we could only inspect the examples in the paper itself.

by Saravanan et al. reliance on cue phrases to identify ratio. Phrases such as “*We are of the view*”, would not work in English law for ratio identification as judges do not use this or any other phrase to point out where they are dictating the ratio and where an obiter.

Therefore, Saravanan et al.’s work on ratio identification method does not capture the complexity we are dealing with in English Supreme Court cases.

2.4 Detection of Legal Principles

Shulayeva et al. address the task of automatically identifying (re)statements of facts and principles that are being cited from an earlier case. They refer to these as *cited* facts and principles (though it is important to note that they actually appear in the *citing* paragraph (A) in Fig. 1).

Their research demonstrates an agreement³ between two human annotators on annotating cited facts and principles of $\kappa = 0.60$. They have automated their task with 85% accuracy using supervised machine learning framework based on linguistic features. The features they use are: part of speech tags, unigrams, dependency pairs, length of sentence, position in the text and an indicator if/whether the sentence contains a full case citation. Their method is described below [8]:

1. Feature counts were normalised by term frequency and inverse document frequency.
2. Attribute selection (InfoGainAttributeEval in combination with Ranker (threshold = 0) search method) was performed over the entire dataset.
3. The Naive Bayes Multinomial classifier was used for the classification task.
4. Results are reported for tenfold cross-validation. The 2659 sentences in the dataset were randomly partitioned into 10 subsamples. In each fold one of the subsamples was used for testing after training on the remaining 9 subsamples. Results are reported over the 10 testing subsamples, which constitute the entire dataset.

Shulayeva et al. have applied their framework on their so-called Gold Standard corpus comprising of 2659 sentences selected from 50 common law reports that had been taken from the British and Irish Legal Institute (BAILII) website in RTF format. The corpus contained human annotated sentences labeled 60% as neutral, 30% as principles and 10% as facts. Their complete results are in Table 1 of the Appendix.

2.5 Cited Paragraph Detection

Adedjouna et al. focused on cross reference identification (CRI) and resolution (CRR) in legislature [1]. They demonstrate that by developing a legal text

³ κ is the predominant agreement measure that corrects raw agreement $P(A)$ for agreement by chance $P(E)$ [3, 8]:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}.$$

schema, it is possible to define a structure capable of CRI, but also CRR, without the need for understanding the semantics of a sentence [1]. Reporting that using natural language processing techniques, a schema can be automatically applied on Luxembourg’s legislation to identify references with 99.7% precision and resolve them with 99.9% precision. They evaluate their method on 1223 references selected from Luxembourg’s income tax law.

3 From Case Law to Ratio Decidendi

According to the doctrine of stare decisis, cases are decided based on the precedent set out in the case law preceding them. Judges cite previous cases to apply the ratio from preceding cases on the facts before them [4]. Therefore, when judges cite specific paragraphs, it is reasonable to expect they are citing the paragraphs mostly for the ratio contained in them. To find the ratio we would thus like to identify cited paragraphs and identify the principles in them.

Shulayeva et al.’s method cannot be used to identify ratio directly because, as they themselves point out, it tends to confuse the ratio of the earlier case with that of the subsequent case: *“The main cause of error for the automatic annotation of principles was that the Gold Standard only annotated principles from cited cases, but often these were linguistically indistinguishable (in our machine learning approach) from discussions of principles by the current judge”* [8]. Perhaps for this reason, Shulayeva et al. explicitly say that *“the term ratio will be avoided”* [8].

Instead, we hypothesise it is better to try and identify ratio by identifying cited paragraphs (B) and extracting principles directly from them.

4 Testing Our Approach

To test if cited paragraphs correlate with ratio more than citing paragraphs, we conducted an empirical study on what we have called the Stack corpus. We created the Stack corpus by collecting and analysing cases from the Westlaw legal search engine, which contains a list of key citing cases for every major case in its database. As the case of interest, Stack v Dowden [2007] UKHL 17 was selected. This had 51 citing cases, of which Westlaw contained 41 in its database. Our Stack corpus is based on the 798 specific citations to Stack contained in these 41 citing cases.

We selected Stack for two reasons. First, it is both cited and cites frequently as a major Supreme Court judgement and therefore gives us plenty of data to work with. Second, because it contains a famous dissenting judgement of Lord Neuberger, containing principles reaching conclusion disagreeing with the rest of the judgement, a good example of an obiter. It is therefore a very challenging and rich case for the purposes of this task.

Our study essentially compares the occurrence of ratio in cited paragraphs containing a principle, with the occurrence of ratio in citing paragraphs containing a principle. We therefore manually identified the citing and cited paragraphs, paragraphs containing ratio and paragraphs with principles.

To determine whether a paragraph contains ratio or obiter, we manually applied the Inversion Test. Under the Inversion Test, the meaning of the principle is inverted and if this disrupts the logic for arriving at the outcome of the case, the principle is ratio, otherwise it is an obiter. Paragraphs without a principle were labeled as an obiter. Employing the Inversion Test we have found out that out of 158 paragraphs in Stack, only 34 contain ratio.

Manually analysing all the 41 citing cases in Stack corpus we found all the paragraphs in Stack which have been cited. Out of 798 citations in the cases, we identified 72 distinct cited paragraphs in Stack. Out of these 62 contain principles. Going through Stack paragraph by paragraph we identified 46 citing paragraphs, out of these 34 contain principles.

The confusion matrix in Table 3 of the Appendix shows the results for cited paragraphs. The data suggest correlation between cited paragraphs and the ratio. Applying this method achieves 76% accuracy in identifying paragraphs containing ratio and $\kappa = 0.45$. We have further tested the opposite assumption looking at citing paragraphs. As per Table 4 of the Appendix there is a drop in accuracy to 68% and most importantly the κ coefficient falls down to mere 0.06. These results support our hypothesis that cited paragraphs are a better indicator of ratio than citing paragraphs. Cited paragraphs are almost twice as precise and have more than three times the recall in identifying ratio than citing paragraphs.

Analysing the mislabeled paragraphs, we discovered two common reasons why judges cite paragraphs without the ratio.

1. The judge might refer to obiter of the cited case. There are numerous reasons why a judge might do this. For example he might highlight a shortcoming of current law:

“27 There is obvious force in the claimant’s contention that, as she and the defendant took out a mortgage in joint names for 43,000, for which they were jointly and separately liable, in respect of a property which they jointly owned, this should be treated in effect as representing equal contributions of 21,500 by each party to the acquisition of the property. It is right to mention that I pointed out in paras 118–119 in the Stack case that, although simple and clear, such a treatment of a mortgage liability might be questionable in terms of principle and authority.” (Lord Neuberger citing himself in *Laskar v Laskar* to suggest he would prefer a different (his own) interpretation of the law.)

Interestingly, a judge might also be using parts of the logic of the dissenting judge, in our case Lord Neuberger, to explain the reasoning of the majority of judges in the case:

*“125 While an intention may be inferred as well as express, it may not, at least in my opinion, be imputed. That appears to me to be consistent both with normal principles and with the majority view of this House in *Pettitt v Pettitt* [1970] AC 777, as accepted by all but Lord Reid in *Gissing v Gissing* [1971] AC 886, 897 h, 898 b - d, 900 e - g, 901 b - d, 904 e - f, and reiterated by the Court of Appeal in*

Grant v Edwards [1986] Ch 638, 651 f - 653 a. The distinction between inference and imputation may appear a fine one (and in Gissing v Gissing [1971] AC 886, 902 g - h, Lord Pearson, who, on a fair reading I think rejected imputation, seems to have equated it with inference), but it is important." (Lord Neuberger's obiter in *Stack v Dowden*, often cited as an explanation of the ratio.)

2. The judge might refer to the facts of the cited case. This happens when the judge is linking two cases on their facts to establish the applicability of the precedent:

"77 A great deal of work was done on the Purves Road property, some of it redecoration and repairs, some of it alterations and improvements. There is no doubt that the parties worked on this together, although there was a dispute as to exactly how much work each did and the judge found that Mr Stack probably did 'more than Ms Dowden gave him credit for' and eventually concluded that 'he had been responsible for making most of these improvements'. But he could not put a figure on their value to the sale price." (Facts cited in Paragraph 77 of *Stack v Dowden*.)

5 Automating Our Approach

To automate the methodology from Sect. 3 we had to resolve two problems. First, how to automatically identify principles and second, how to automatically identify cited paragraphs. Each is explored in a subsection below. Finally we combine the two solutions to automatically identify the ratio.

5.1 Principle Identification

Shulayeva et al. report two problems with their method of cited facts and principles annotation. First, cited principles and facts are often *"linguistically indistinguishable (in [their] machine learning approach) from discussions of principles by the current judge"* and second, sentences can contain both facts and principles at the same time, making it difficult to classify them as only one or the other. Under the definition of a ratio we employ we need only to identify the principles, as opposed to both facts and principles. Further, because we identify cited paragraphs via a separate method (see Sect. 5.2), we are interested in all principles, as opposed to only the restatements of facts and principles that are being cited from an earlier case.

We have therefore relabelled 96 sentences originally labelled neutral (because they are discussions of principles of a current judge) or fact (because they contain both fact and principle) as principles, 4% of the Gold Standard corpus, and created the New corpus. We implemented Shulayeva et al.'s framework in Weka according to their instructions and trained it on the New corpus. Below are some examples of the types of sentences we have re-annotated.

First is a sentence introducing a new principle, instead of restating a cited one, labelled as neither in Gold Standard corpus. We have re-labelled this sentence as a principle in the New corpus, because we are interested in identifying all the principles.

“He is not obliged to comply with any request that may be made to him by the borrower let alone by a surety if he judges it to be in his own interests not to do so.”

Second is a sentence classified in Gold Standard corpus as a fact. This sentence however contains both a fact and a principle, the latter is highlighted by us in bold. For our purposes such sentence still carries a principle and we therefore relabel it from a fact to a principle.

*“Thus for example if the manager explains either when making the request for payment to a third party or on it being questioned by the customer that the third party is a supplier and that the object is to obtain necessary materials for the work more quickly or that the third party is an associated company carrying on the same business **such an explanation might well bring the request for the payment to the third party within the usual authority of a person in his position and therefore within his apparent authority.**”*

The complete results are reported in Table 2 of the Appendix. They show a high accuracy for the task of principle identification, with over 10% increase in accuracy and a jump from $\kappa = 0.72$ to $\kappa = 0.90$ when applying Shulayeva’s model trained on our New corpus on the task of identifying principles compared to using the same model, trained on their Gold Standard corpus to identify restatements of facts and principles.

5.2 Cited Paragraphs Identification

There are noticeable similarities between cross references in legal statutes and paragraph citations in case law. For example in Fig. 2 Adedjouma’s cross reference identification (CRI) would identify “Art. 156.” using regular expression. In case law a paragraph would be cited as “para. 156”. For cross reference resolution (CRR), statutes seem to be comparable to case law, with implicit references such as “the next article” or explicit “the law of April 13th, 1995” which would be “above” or “in *Stack v Dowden*” in case law.

Art. 2. [...] Individuals are considered non-resident taxpayers if they do not reside in Luxembourg but have a local income as per the definition of Art. 156.

Fig. 2. CRI in Dutch legislation.

To identify cited paragraphs we have analysed 798 citations in Stack corpus. The 41 cases in the corpus are concerned with different issues, some citing to distinguish their problem, some to criticise it, and some to apply the law. They also come from a variety of courts and judges. Together, this gives us a representative sample of the variety of approaches judges and transcript writers could use to cite another case.

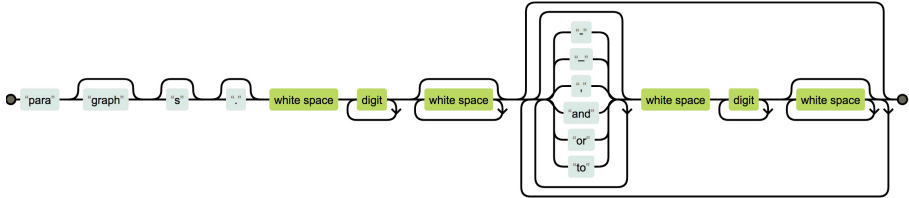


Fig. 3. Regular expression for the first type of paragraph citation.

Just like Adedjouma et al. we start by identifying all the patterns possible for recognizing the citation itself (CRI). There seem to be two general approaches. Under the first approach, the paragraph is either cited as *paragraph* or its abbreviation such as *para* or *paras*, see Fig. 3. The second approach, on the other hand, begins the citation with *at* followed by the paragraph numbers in square brackets, see Fig. 4.

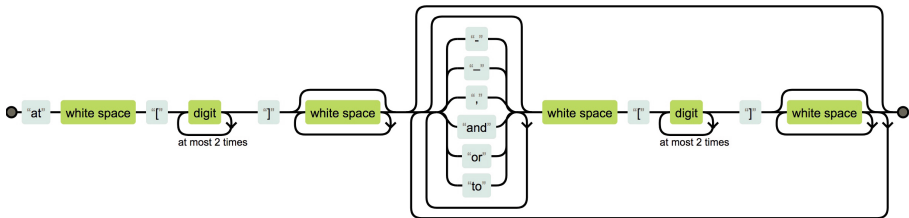


Fig. 4. Regular expression for the second type of paragraph citation.

The paragraph numbers themselves are expressed either as a single *number*, which correspond to Adedjouma’s simple cross reference expressions (CRE), or as a *list of numbers* or a *range of numbers*, that correspond to Adedjouma’s multivalued CRE’s [1]. First we extract these citations using regular expression, then we interpret them to get all the implicit numbers in a range.

With the numbers of paragraphs cited identified the task of CRR remains. Because Adedjouma et al. are focused on statutes, they can simply extract the “*Article 156*”, and assume it’s the Article 156 of the same document the citation has been extracted from. Our task isn’t as simple. Unlike Adedjouma et al., we

need to distinguish if the paragraph citations is pointing to the case we are interested in or not, since many cases are cited by a single case.

On top of that, we need to do this for case law which is by its very nature less structured than statutes. Consider the example sentences below. While we can relatively easily identify paragraph citation pointing to *Stack v Dowden* in example 1 and 2 below, where the name of the case or its abbreviation is included in the sentence, the same can't be said about examples 3 and 4.

1. *“Indeed, this would be rare in a domestic context, but might perhaps arise where domestic partners were also business partners: see Stack v Dowden, para 32.”*
2. *“First, as in the Stack case (see paras 90–92), the two parties in this case kept their financial affairs separate.”*
3. *“Fourthly, however, if the task is embarked upon, it is to ascertain the parties’ common intentions as to what their shares in the property would be, in the light of their whole course of conduct in relation to it: Lady Hale, at para 60.”*
4. *“In paragraph 42, Lady Hale rejected this approach in Jones v Kernott.”*

Carefully analysing the sentences citing *Stack v Dowden*, we have come up with a schema capable of identifying citations of our case of interest. Our schema takes an advantage of the knowledge of the names of the judges and the name of the case of interest, as both are written at the top of each case, and can be easily identified in the text. We also count the number of paragraphs in the case of interest to be able to reject any citation of a paragraph larger than this as impossible. The case abbreviation can be easily identified, since it is the first name in the case (e.g. for *Stack v Dowden*, it would be *Stack*). Employing these features, we constructed and implemented the Schema in Fig. 5 and applied it on the *Stack* corpus.

Under this schema we can resolve the examples above, including the problematic examples 3 and 4. Example 4 can be rejected as it contains a citation of another case (eg. *Jones v Kernott*, which we identify using regular expression). On the other hand example 3 would be classified as citing *Stack v Dowden*, because it contains the name of Baroness Hale, a judge in *Stack v Dowden*.

Comparing 3 and 4, one might notice there is no smoking gun evidence suggesting *Lady Hale* of example 3 is referencing *Lady Hale’s* judgement in *Stack v Dowden* instead of any other case she has judged, such as *Jones v Kernott* in example 4. However in our approach we work with the knowledge the case we analyse is citing the case of interest at some point. We know this since the cases we are analysing are selected from the list of cases citing *Stack v Dowden*, Westlaw provides. Therefore assuming that the case cited with a judge of the case of interest is indeed the case of interest, is a reasonable assumption to make. And as we report below, while a weakness of our approach, it still allows us to identify cited paragraphs with very high accuracy. Moreover, this is the best we can do without engaging with full analysis of the paragraph, or indeed the full case, which would be necessary to fully resolve this problem.

From 798 citations in the *Stack* corpus, the schema identifies 176 citations of *Stack*, 175 out of these are true positives. The single false positive is of a

sentence where the same Judge, who gave the judgement in the case of interest, is reported citing a paragraph but of a different case that is not mentioned in the same sentence. A schema analysing semantics of a sentence would be required to resolve this issue. There are also 8 citations that the program fails to recognise as citing Stack. These false negatives have not been extracted because the citing entity is contained in a different sentence or paragraph from the citing expression. Despite these shortcomings of our schema we achieve 98.7% precision, comparable to Adedjouma’s 99.9%.

However, since we are focused on identifying paragraphs citing Stack, it’s better to evaluate on how accurate the classifier is at identifying cited paragraphs. Out of 72 paragraphs we have manually identified as cited from 158 paragraphs in Stack, the classifier identifies 64 true positives 85 true negatives, giving it a decent accuracy of 94%. The full results are in Table 5 of the Appendix.

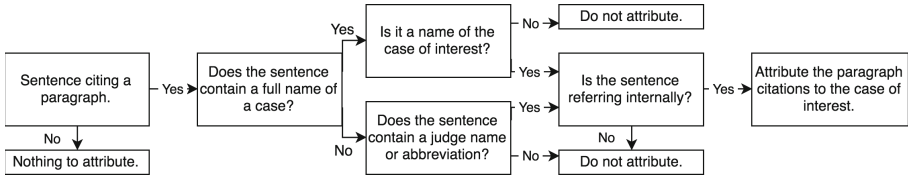


Fig. 5. Schema for attributing a citation to the case of interest from a present case (i.e. the case we are in).

5.3 Automatically Identifying Cited Principles

Finally, simply combining the principle and cited paragraph classifier described above, we evaluate how accurate our method is in identifying the ratio. As per Table 6 the new classifier identifies ratio with 72% accuracy on our Stack corpus.

Combining the principle and cited paragraph classifier therefore not only pin-points the position of the ratio in the paragraph, but also filters the cited paragraphs that do not contain principles, removing the instances where only facts are cited, further improving the performance.

Our automated approach therefore nearly matches it’s theoretical ceiling performance of 76% in the task of ratio identification (as established by our manual annotation study in Sect. 4), proving our hypothesis that focus on principles in cited paragraphs is a possible way of tackling the difficult task of automatic ratio identification.

6 Conclusion and Future Work

In this paper we have presented a novel approach to ratio identification. We demonstrate that identifying ratio can be automated by looking at cited paragraphs with 72% accuracy. We have also improved upon Shulayeva et al.’s work,

adapting their model for the sole identification of principles. Applying similar ideas to Adedjouma et al. we demonstrate that cross reference identification and resolution and cited paragraph identification can be achieved with almost equally high precision. Given the time-consuming nature of manually identifying ratio, our approach is a step forward in helping lawyers and judges spend their time on applying the law rather than looking for it. However, our work is only a preliminary study. A larger dataset, with more cases and human annotators, would be necessary for a full evaluation of the accuracy of our approach as well as the scope of it’s applicability. After all, only heavily cited cases can take benefit of our method. A very recent case, or a case without citations, will naturally be impossible to analyse. This is a limitation of our approach.

In our research, we have focused on all cited paragraphs without discrimination. However, not all citations are of equal importance. They are cited with different frequency, they are cited by cases from courts of different importance (Supreme court might cite differently than the Court of Appeal), the citing cases themselves might be reported immediately after the case comes out, as well as several years or even decades later and the citing case might approve as well as disapprove of a paragraph. All of the above could be used as features discriminating between cited paragraphs. Further, it is not only paragraphs that are cited, a sentence can be directly quoted and paragraphs might be cited individually or in a range. Discriminating between the precision with which text is referred to could again help distinguish between ratio and obiter. In the future, we would like to explore how the features above could further reduce the misclassification of obiter as ratio in our method.

Appendix

Table 1. Per category and aggregated statistics for the original Shulayeva et al.’s principle and fact classifier trained on Gold Standard corpus.

Classified as →	Principle	Fact	Neither
Principle	646	5	160
Fact	4	198	41
Neither	135	38	1432
Type	Precision	Recall	F-measure
Principle	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884
Accuracy	0.85	κ	0.72

Table 2. Per category and aggregated statistics for Shulayeva et al.’s classifier trained on New corpus for extraction of principles only.

Classified as →	Principle	Neither	
Principle	837	70	
Neither	48	1769	
Type	Precision	Recall	F-measure
Principle	0.946	0.923	0.934
Neither	0.962	0.974	0.968
Accuracy	0.96	κ	0.90

Table 3. Distribution of ratio and obiter between cited and not cited paragraphs containing principle.

Contained in →	Cited	Not-cited	
Ratio	31	3	
Obiter	41	83	
Type	Precision	Recall	F-measure
Cited	0.468	0.853	0.604
Not-cited	0.948	0.734	0.827
Accuracy	0.76	κ	0.45

Table 4. Distribution of ratio and obiter between citing and not citing paragraphs containing principle.

Classified as →	Citing	Not-citing	
Ratio	9	25	
Obiter	25	99	
Type	Precision	Recall	F-measure
Citing	0.265	0.265	0.265
Not-citing	0.798	0.798	0.798
Accuracy	0.68	κ	0.06

Table 5. Per category and aggregated statistics for cited paragraph classifier.

Classified as →	Cited	Not-cited	
Cited	64	8	
Not-cited	1	85	
Type	Precision	Recall	F-measure
Principle	0.985	0.889	0.935
Neither	0.914	0.988	0.950
Accuracy	0.94	κ	0.88

Table 6. Per category and aggregated statistics for Ratio Decidendi classifier.

Classified as →	Ratio	Obiter	
Ratio	22	12	
Obiter	33	91	
Type	Precision	Recall	F-measure
Principle	0.400	0.647	0.494
Neither	0.884	0.734	0.802
Accuracy	0.72	κ	0.31

References

1. Adedjouma, M., Sabetzadeh, M., Briand, L.C.: Automated detection and resolution of legal cross references: approach and a study of Luxembourg’s legislation. In: 2014 IEEE 22nd International Requirements Engineering Conference (RE), pp. 63–72, August 2014
2. Branting, K.: Four challenges for a computational model of legal precedent. *THINK (J. Inst. Lang. Technol. Artif. Intell.)* **3**, 62–69 (1994)
3. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.* **22**(2), 249–254 (1996)
4. Elliott, C., Quinn, F.: *English Legal System*, 1st edn. Pearson Education, New York (2012)
5. Greenawalt, K.: Interpretation and judgment. *Yale J. Law* **9**(2), 415 (2013)
6. Raz, M.: Inside precedents: the ratio decidendi and the obiter dicta. *Common L. Rev.* **3**, 21 (2002)
7. Saravanan, M., Ravindran, B.: Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artif. Intell. Law* **18**(1), 45–76 (2010). <https://doi.org/10.1007/s10506-010-9087-7>
8. Shulayeva, O., Siddharthan, A., Wyner, A.: Recognizing cited facts and principles in legal judgements. *Artif. Intell. Law* **25**(1), 107–126 (2017). <https://doi.org/10.1007/s10506-017-9197-6>
9. Zhang, P., Koppaka, L.: Semantics-based legal citation network. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, ICAIL 2007, pp. 123–130. ACM, New York (2007). <http://doi.acm.org/10.1145/1276318.1276342>



Textual Entailment in Legal Bar Exam Question Answering Using Deep Siamese Networks

Mi-Young Kim^{1,3(✉)}, Yao Lu³, and Randy Goebel^{2,3}

¹ Department of Science, Augustana Faculty, University of Alberta, Camrose, AB, Canada
miyoung2@ualberta.ca

² Department of Computing Science, University of Alberta, Edmonton, AB, Canada
rgoebel@ualberta.ca

³ Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB, Canada
yaol@ualberta.ca

Abstract. Every day a large volume of legal documents are produced, and lawyers need support for their analysis, especially in corporate litigation. Typically, corporate litigation has the aim of finding evidence for or against the litigation claims. Identifying the critical legal points within large volumes of legal text is time consuming and costly, but recent advances in natural language processing and information extraction have provided new enthusiasm for improved automated management of legal texts and the identification of legal relationships. As a legal information extraction example, we have constructed a question answering system for Yes/No bar exam questions. Here we introduce a Siamese deep Convolutional Neural Network for textual entailment in support of legal question answering. We have evaluated our system using the data from the competition on legal information extraction/entailment (COLIEE). The competition focuses on the legal information processing required to answer yes/no questions from legal bar exams, and it consists of two phases: legal ad-hoc information retrieval (Phase 1), and textual entailment (Phase 2). We focus on Phase 2, which requires “Yes” or “No” answers to previously unseen queries. We do this by comparing the extracted meanings of queries and relevant articles. Our choice of features used for the semantic modeling focuses on word properties and negation. Experimental evaluation demonstrates the effectiveness of the Siamese Convolutional Neural Network, and our results show that our Siamese deep learning-based method outperforms the previous use of a single Convolutional Neural Network.

Keywords: Legal question answering · Recognizing textual entailment
Siamese network · Convolutional neural network

1 Introduction

Every day a large volume of legal documents are produced, and lawyers need support for the analysis of the big documents, especially in corporate litigation. Typically corporate litigation has the aim of finding evidence for or against the litigation claims. Identifying the critical legal points within large volumes of legal text is time consuming and

costly, but recent advances in natural language processing and information extraction have provided new enthusiasm for improved automated management of legal texts and the identification of legal relationships.

We believe that developing tools to automatically or semi-automatically confirm entailment relationships between legal texts is fundamental to legal text understanding. We are interested in developing question-answering tools that help users obtain information about their questions in law; an initial important step is the development of a question answering system for Yes/No questions from bar exams. Yes/No question answering is significantly easier than general legal question answering, but we still need to develop tools for determining the semantic relationships amongst legal texts.

Our approach to Yes/No legal question answering requires a number of intermediate steps. For instance, consider a question such as *“Is it true that a special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on his/her own for a third party, the seller is not released of warranty?”* It is clear that a system must first identify and retrieve relevant documents, typically legal statutes (Phase 1), and subsequently identify most relevant segments in the statutes, appropriate for responding to the input query. Finally it must compare the semantic connections between question and the relevant segments, and determine whether an entailment relation holds (Phase 2).

Here we focus on Phase 2, and based on previous work, augment our accumulated developments with a Siamese Deep Neural Network model to support textual entailment. The Deep Neural Network (DNN) is a technology that has recently demonstrated dramatic success in several areas, including image classification, language modeling, and speech feature extraction and recognition [4, 5, 9].

Previous work that incorporated convolution and subsequent pooling into a neural network gives rise to a technique called a Convolutional Neural Network (CNN) [12]. CNNs have performed well in image and speech recognition [9], and many studies have been proposed applying CNN in natural language processing [4, 5]. In addition, so-called Siamese networks [1] have been used in image processing to predict whether persons illustrated in an input image pair are the same [2, 3]. These Siamese networks have also been used to detect semantic similarity in texts [13–15].

We have constructed a Siamese Network to support textual entailment for the Competition on Legal Information Extraction/Entailment (COLIEE)¹. The COLIEE competition focuses on two aspects of legal information processing related to answering yes/no questions from legal bar exams: Legal document retrieval (Phase 1), and Yes/No Question answering for legal queries (Phase 2).

The goal of Phase 2 is to construct Yes/No question answering systems for legal queries, by confirming entailment from the relevant articles. The answer to a question is typically determined by measuring some kind of heuristically computed semantic similarity between the question text and the relevant statutes, to determine the Yes/No answer. While there are many possible approaches, we here consider augmenting our previous approaches with a Siamese neural network-based distributional sentence model, because of their success in approximating semantic similarity [13–15]. As a

¹ <http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2017/>.

consequence of this success, it appears natural to approach textual entailment using similar techniques. We demonstrate the performance of a Siamese neural network-based sentence model for the task of textual entailment.

Like all systems based on the development of a classification model, our prototype system relies on appropriate identification and annotation of text features to inform the training process. Since our task is to provide Yes/No answers, we first extract word and negation features, then train a supervised model to identify relationships between questions and corresponding articles.

For feature extraction, we align heuristically relevant segments from a question to similar segments within the corresponding statute, following Kim et al. [6], and then, to reduce noise, we extract features only from the related segments. We then employ a Siamese convolutional neural network algorithm using the extracted features, and compare its performance with a previously constructed single Convolutional Neural Network.

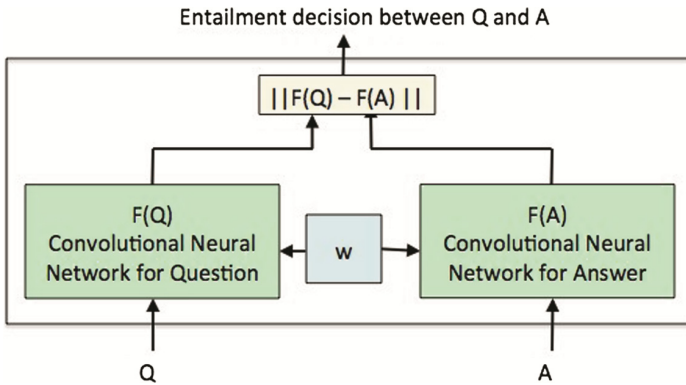


Fig. 1. Siamese network architecture

2 Siamese Network

We construct a Siamese network model based on our choice of entailment attribute measures. Siamese networks were first introduced in the early 1990s by Bromley et al. [16] to solve the signature verification problem conceived as an image-matching problem. A Siamese neural network consists of twin networks that accept distinct inputs, but are connected by a distance function (cf. Fig. 1). This function computes a metric between the highest-level feature representation on each side. The weights between both networks are shared, and we choose a contrastive loss function, described in [17], as the loss function for our Siamese network.

The motivation to use Siamese architectures is as following: Sharing weights across sub-networks means fewer training parameters, which in turn means less data required and less tendency to overfit. Because the COLIEE data size is not big, we expect good performance. In addition, Siamese networks have shown state-of-the-art performance in the text similarity task [20], which is similar to the textual entailment task. To our

knowledge, our system is the first to use Siamese networks in textual entailment, so this study will be a start in determining what features and network structures are appropriate for the use of Siamese networks in textual entailment.

Selected feature vectors are used as input to the neural network. Figure 1 shows a Siamese network, where Q represents a legal bar question text and A represents the corresponding law statute text. W is the shared weight between two neural networks. In our application, the function $\|F(Q) - F(A)\|$ is a heuristic measure of the semantic entailment relatedness between a legal bar question (Q) and a legal statute (A). Here, $F(Q)$ is a function with set of parameters W . $F(Q)$ is assumed to be differentiable with respect to W . The Siamese network seeks a value of the parameter W such that the symmetric similarity metric is small if Q is entailed from A , and large if Q is not. The Euclidean distance of the vectors is used to compute contrastive loss. The goal is to minimize the distance in the semantic space of two similar text pairs, and maximize for non-similar pairs. The contrastive loss can be given by following equation as shown in [13]:

$$L = Y \|F(Q) - F(A)\|^2 + (1 - Y) \max(0, 1 - \|F(Q) - F(A)\|),$$

where Y is 1 if Q can be entailed from A , 0 otherwise.

3 Our System

3.1 Model Description

The problem of answering a legal yes/no question can be viewed as a binary classification problem. Assume a set of questions Q , where each question $q_i \in Q$ is associated with a list of corresponding article sentences $\{a_{i1}, a_{i2}, \dots, a_{im}\}$, where $y_i = 1$ if the answer is ‘yes’ and $y_i = 0$ otherwise. We simply treat each data point as a triple (q_i, a_{ij}, y_i) . Therefore, our task is to learn a classifier over these triples so that it can predict the answers of any additional question-article pairs $(q_i, a_{ij}, ?)$.

Our solution to this problem assumes that correct answers are based on high semantic similarity between articles (statutes) and questions. We model questions and answers as vectors that comprise extracted words and negation information, and evaluate the relatedness of each question-article pair in a shared vector space.

We have employed a Convolutional Neural Network (CNN) for this task, which are a biologically-inspired variant of a multi layer perceptron. A CNN uses two techniques to create a model: (1) restrict the network architecture through the use of local connections known as receptive fields; and (2) constrain the choice of synaptic weights through the use of weight-sharing. Most CNNs include a so-called max-pooling layer that reduces and integrates the neighbouring neurons’ outputs. CNNs also exploit spatially-local correlation by enforcing a local connectivity pattern between neurons of adjacent layers. CNN-based models have been proved to be effective in semantic similarity detection [15, 18]. Our use of CNNs exploit linguistic and negation features with one convolutional layer and one pooling layer.

3.2 Architecture

Our experimental Siamese system consists of a pair of deep Convolutional Neural Networks (CNN) with convolution, max pooling and rectified linear unit (ReLU) layers, together with a fully connected layer at the top (see Fig. 2). Each CNN gives a non-linear projection of the question and answer vectors in the semantic space. Those semantic vectors are connected to a layer that measures distance or similarity between them. The contrastive loss function combines the distance measure and the label. The gradient of the loss function (with respect to the weights and biases shared by the sub-networks) is computed using back-propagation. A stochastic gradient descent method is used to update the weights.

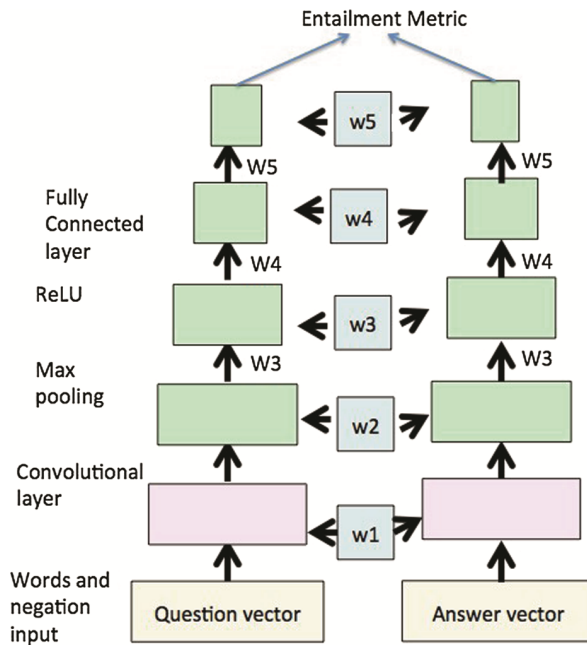


Fig. 2. Layers of deep convolutional neural networks

To extract word features from the aligned question and statute segments, we first remove stop words and perform stemming. Each question-answer pair consists of words and negation information. Subsequently, in the three layer CNN (convolutional layer, pooling layer, fully connected layer), the convolutional layer is applied to the question answer vectors.

Each question-statute pair is (q_i, a_i) such that $\bar{q}_i, \bar{a}_i \in \mathbb{R}^{n+2}$, where n is the total number of unique words in the training data. The convolutional layer is applied on the question answer vectors by convolving a filter with weights $c \in \mathbb{R}^{h \times w}$, where h is the filter height and w is the filter width. A filter, consisting of a layer of weights, is applied to a small window of a vector to get a single unit as output. The filter is slid across the length of

vector such that the resulting connectivity looks like a series of overlapping receptive fields, with output of width w . Max Pooling performs a kind of non-linear down-sampling, which helps reduce the tendency to overfit. This method splits the filter outputs into small non-overlapping grids, and takes the maximum value in each grid as the value in the output of reduced size. The max pooling layer is applied on top of the output given by a CNN, to extract the highest contributing local features to form a fixed-length feature vector [15].

When a neural network is trained on a small training set, it typically performs poorly on test data. This “overfitting” is greatly reduced by randomly omitting some of the feature detectors on each training case, which is a process called “dropout”. Dropout prevents complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors. Random dropout has been used to achieve improvements on many benchmark tasks, and has set new standards for speech and object recognition [11].

As the last component of our prototype adoption of CNNs, we also employ so-called Rectified Linear Units. Neural networks with rectified linear unit (ReLU) non-linearities have been highly successful for computer vision tasks and have been shown to train faster than standard sigmoid units [8]. Intuitively, an ReLU is a neuron that uses a rectifier instead of a hyperbolic tangent or logistic function as an activation function. *Rectifier* $f(x) = \max(0, x)$. The terminal layer of the CNN sub-network is a fully connected layer which converts the output of the last ReLU layer into a fixed-length semantic vector for input to the sub-network [18].

3.3 Input Features

We use features that are directly related to the legal sentence structure. We analyzed the training data, and found that a legal statute consists of ‘general condition’, ‘conclusion’, ‘exceptional case’, and ‘conclusion for the exceptional case’. The query can belong to either ‘general condition’ or ‘exceptional case’. We filter the cases in the statute that are not related to the input query, by determining where the input query belongs. We use condition/conclusion/exception detection rules in Kim et al. [6] as follows:

$conclusion := segment_{last}(sentence, keyword),$

$condition := \sum_{i \neq last} segment_i(sentence, keyword),$

$condition := sub_condition \text{ [or] } condition$

$condition := sub_condition \text{ [and] } sub_condition$

$exception_conclusion := segment_{last}(sentence, exception_keyword),$

$exception_condition := \sum_{i \neq last} segment_i(sentence, exception_keyword),$

$exception_condition := exception_condition \text{ [or] } exception_condition$

$exception_condition := sub_exception_condition \text{ [and] } sub_exception_condition$

The *keywords* of the condition are as follows: “*in case(s)*,” “*if*,” “*unless*,” “*with respect to*,” “*when*,” and “*,* (comma).” They used the symbol Σ to denote the concatenation of the segments. The *exception_keyword* is “... *this shall not apply, if (unless)*” [9].

As described in the previous sections, the types of language features we use are simple word lemma, and negation. We determine if a query belongs to the general case or exceptional case by computing the proportion of shared words between query and ‘condition’, and query and ‘*exception_condition*’. We collect words from the detected *condition* (*exception_condition*) and *conclusion* (*exception_conclusion*), and also determine a negation value for each *condition* (*exception_condition*) and *conclusion* (*exception_conclusion*). The negation level (*neg_level(segment)*) is computed as follows: if [negation + antonym] occurs an odd number of times in the segment, its negation level is 1. Otherwise if the [negation + antonym] occurs an even number of times, including zero, its negation level is 0.

Each question-answer pair is written as $(\vec{q_i}, \vec{a_i})$ where $\vec{q_i}, \vec{a_i} \in \mathbb{R}^{n+2}$. Here, n is the total number of unique words in the training data. The additional two feature values are the negation values of condition (*neg_level(condition)*) and conclusion (*neg_level(conclusion)*).

4 Experiments

4.1 Training

Our system is trained using questions and relevant articles, which have been confirmed to have semantic entailment relation pairs as positive samples, together with questions and relevant articles that do not have semantic entailment relations, as negative samples.

Training the network with a shared set of parameters not only reduces number of parameters (so saving computation) but also ensures consistency of the representation of questions and answers in the semantic space. As previously mentioned, the shared parameters of the network are learned with the aim to minimize the semantic entailment distance between the answers and entailed questions, and maximize the semantic entailment distance between the question and non-entailed questions.

The loss function is minimized such that question-answer pairs with label 1 (question-answer pairs confirmed to have a textual entailment relation) are projected nearer to each other, and those with label 0 (question-answer pair which do not have textual entailment relation) are projected far away from each other. The model is trained by minimizing the overall loss function in a batch.

COLIEE provides training and test set separately for the competition. Each question in the dataset contains ID, description, corresponding statute, and semantic entailment relation (‘Yes’ or ‘No’). Similarly, the validation dataset contains question answer pairs without semantic entailment answer. The hyper-parameters of the network are tuned on the validation dataset. We used the following parameters for our network: dropout rate = 0.2, epochs = 25, learning rate = 0.01, and momentum = 0.05.

We evaluated the performance of our models by measuring answer accuracy.

4.2 Testing

The term vectors of the question pairs are words and negation features, fed to the twin sub-networks. The trained shared weights of our system projects the question vectors in the semantic space. The similarity between the pairs is calculated using the similarity metric learnt during the training. So our system outputs a value of distance measure (score) for each pair of questions. The threshold is set to the average similarity score of the positive examples in training data.

4.3 Experimental Results

For comparison with previous work, we used the same data as in [18]: the COLIEE 2014 training (dry run) data for training, and the COLIEE 2015 test (formal run) data for validation. We checked that test data and training data are not overlapping. There is a balanced positive-negative sample distribution (55.87% yes, and 44.13% no) for a dry run of COLIEE 2014 dataset, so we consider the baseline for true/false evaluation is the accuracy when returning always “yes,” which is 55.87%. Our data for our dry run has 179 questions. Table 1 shows the experimental results on the dry run data of COLIEE 2014. We used this data for training, and achieved 64.25% accuracy.

Table 1. Experimental results on dry run data of COLIEE 2014

Our method	Accuracy (%)
Baseline	55.87
Our siamese network	64.25
A single convolutional neural network [18]	63.87

Table 2 shows the experimental results using formal run data of COLIEE 2015. The formal run data size of COLIEE 2015 is 66 queries for Phase 2 from the bar exam of 2013. Our performance of textual entailment is 68.18%, which achieved better performance than the Rank No. 1 system [18] in the COLIEE 2015 competition. The system that was ranked No.1 showed 66.67% of accuracy.

Table 2. Experimental results on formal run data of COLIEE 2015

Our method	Accuracy (%)
Our siamese network	68.18
A single convolutional neural network [18]	66.67

From unsuccessful instances, we classified the error types as shown in Table 3. We could not identify the errors arising from the Neural Network architecture or embedding vectors, so we just classified the errors into 7 cases. The biggest error arises, of course, from the paraphrasing problem. The second biggest error is because of complex constraints in conditions. As with the other error types, there are cases where a question is an example case of the corresponding article, and the corresponding article embeds another article.

There have been also errors in the case that a question is not correctly classified if it belongs to a *condition* or an *exception_condition* of the relevant legal statute.

Table 3. Error types

Error type	Accuracy (%)	Error type	Accuracy (%)
Specific example case	6.06	Paraphrasing	37.88
Incorrect detection of the <i>condition</i> (<i>exception_condition</i>) that query belongs to	9.09	Complex constraints in condition	28.79
Incorrect detection of <i>condition</i> , <i>conclusion</i> <i>exception_condition</i> , <i>exception_conclusion</i>	6.06	Reference to another article	3.03
Unclassified	9.09		

The following three examples show our system's error case and correctly answered case.

<pair id = "H25-6-E">

<statute>

Article 174-2

(1) The period of prescription of any right established in an unappealable judgment shall be ten years even if any period of prescription shorter than ten years is provided. The same shall apply to any right which is established in a settlement in a court proceeding or conciliation, or any other action which has the effect equivalent to that of the unappealable judgment.

(2) The provision of the preceding paragraph shall not apply to any claim which is not yet due and payable yet at the time when the judgment becomes unappealable.

Article 724

The right to demand compensation for damages in tort shall be extinguished by the operation of prescription if it is not exercised by the victim or his/her legal representative within three years from the time when he/she comes to know of the damages and the identity of the perpetrator. The same shall apply when twenty years have elapsed from the time of the tortious act.

</statute>

<query>

Even if a right to demand compensation for damages due to torts is established in a settlement in a court proceeding, and due date is fixed one year after the settlement, such right shall be extinguished by the operation of prescription if it is not exercised by the victim within ten years from the time when such settlement had been mentioned in the record of settlement.

</query>

</pair>

<pair id = "H25-8-5">

<statute>

Article 243

If two or more movables with different owners are so joined to each other that they can no longer be separated without damaging the same, ownership of the composite Thing shall vest in the owner of the principal movables. The same shall apply if excessive expense would be required to separate the same.

Article 244

If the distinction of principal and accessory cannot be made between the joined movables, the owner of each movable shall co-own the composite Thing in proportion to the respective price current at the time of the accession.

</statute>

<query>

If two or more movables with different owners are so joined to each other that they can no longer be separated without damaging the same, the owner of secondary movable shall co-own the composite Thing in proportion to the respective price current at the time of the accession.

</query>

</pair>

<pair id = "H25-4-U">

<statute>

Article 109

A person who manifested to a third party that he/she granted certain authority of agency to other person(s) shall be liable for any act performed by such other person(s) with third parties within the scope of such authority, unless such third parties knew, or were negligent in not knowing, that such other person(s) were not granted the authority of agency.

Article 110

The provision of the main clause of the preceding Article shall apply mutatis mutandis to the case where an agent performs any act exceeding its authority and a third party has reasonable grounds for believing that the agent has the authority.

Article 112

Termination of the authority of agency may not be asserted vis-a-vis a third party without knowledge provided, however, that, this shall not apply to the cases where such third party was negligent in not knowing such fact.

</statute>

<query>

In the case where a child who was not authorized by his/her father sells the real property of the father as an agent of the father, if the other party believe that the child was authorized and has reasonable grounds for believing that the agent has the authority, apparent authority shall be effected.

</query>

</pair>

In the example "H25-6-E", the bold sentence was chosen as the most relevant condition and conclusion of the query, and then our Siamese network's output was 'no', while the baseline system's answer was 'yes', which is correct. This error was caused by the incorrect detection of condition and conclusion. The example "H25-8-5" shows an error

caused by paraphrasing. Some words are paraphrased between query and statute and our Siamese network’s answer was ‘no’ because the machine could not detect the semantic relation and entailment, but the real answer is ‘true’.

The last example “H25-4-U” shows the case that our system’s output was ‘no’ which is correct, but the baseline’s system result was ‘yes’.

A significant limitation of this method is that we have no idea about how we can adjust the system architecture or parameters/weights to fix a certain errors of the output. An obvious approach would be to revise or expand training data and then re-train our system from the beginning step inefficiently. However, this is a symptom of a bigger challenge: that of instrumenting deep neural networks to provide explanations, and then creating some companion mechanism for direct adjustment corresponding to “being told.” At the very least, we will need to identify more legal features from the legal structures by deeply analyzing the legal inference, not relying on simply neural network architectures.

5 Related Work

Many researchers have applied deep learning to question answering [4, 7, 10]. Relevant work includes Yih et al. [7], who constructed models for single-relation question answering with a knowledge base of triples. In a similar fashion, Bordes et al. [10] used a type of Siamese network for learning to project question and answer pairs into a joint space. Finally, Yu et al. [4] selected answer sentences, which includes the answer of a question. They modeled semantic composition with a recursive neural network. These previous tasks differ from the work presented here in that our purpose is not to make a choice of answer selection in a document, but to answer “yes” or “no.”

A textual entailment method from Bdour and Gharaibeh [2] provided the basis for a Yes/No Arabic Question Answering System. They used a kind of logical representation, which bridges the distinct representations of the functional structure obtained for questions and passages. This method is also not appropriate for our task: if a false question sentence is constructed by replacing named entities with terms of different meaning in the legal article, a logic representation can be helpful. However, false questions are not simply constructed by substituting specific named entities.

There have been several deep learning approaches for textual entailment [18, 19, 21–32]. Kim et al. [18] constructed a CNN for the COLIEE competition to detect textual entailment, and achieved Rank No. 1 in the COLIEE 2015 competition. Lyu et al. [19] built a joint Restricted Boltzmann Machines (RBM) layer to learn the joint representation of the text-hypothesis pairs. Based on the approach of Bowman et al. [21], Rocktaschel et al. [22] used attention weights in a word-by-word neural attention mechanism to improve the performance of LSTM-based recurrent neural network. Yin et al. [23] then presented a general Attention Based Convolutional Neural Network (ABCNN) for modeling a pair of sentences. Liu et al. [24] introduced two coupled ways to model the interdependences of two LSTMs, and Vendrov et al. [25] proposed a method for learning ordered representations. In addition, Mou et al. [26] applied a Tree-based CNN which captures sentence-level semantics, Wang and Jiang [27] proposed a match-LSTM to

perform word-by-word matching of the hypothesis with the premise, Liu et al. [28] utilized a sentence’s representation to attend words appeared in itself, and Cheng et al. [29] proposed a machine reading simulator which processes text incrementally from left to right and performs shallow reasoning with memory and attention. To free the model from traditional parsing process, Bowman et al. [30] introduced a Stack-augmented Parser-Interpreter Neural Network (SPINN). Parikh et al. [31] used attention to decompose the problem into sub-problems that can be solved separately. Sha et al. [32] proposed to use the intensive reading mechanic, which means to re-read the sentence according to the memory of the other sentence for a better understanding of the sentence pair using LSTM-RNN. To our knowledge, there has been no previous work that used Siamese networks in textual entailment.

For the task of similar question detection, Goyal [13] ranked similar questions by using Siamese Network with an LSTM. Muller and Thyagarajan [14] also presented a Siamese adaption of the LSTM network to assess semantic similarity between sentences. Das et al. [15] constructed Siamese Networks using CNN for similar question retrieval in discussion forum. All results that used Siamese Networks showed improved performance in detecting semantic similarity in texts. Muller and Thyagarajan [14] further experimented with semantic entailment detection using their Siamese Network and achieved state-of-the-art performance.

As further research, we intend to apply our method to the SemEval textual entailment task and Stanford SNLI corpus, and investigate our method’s adaptability on other domain data besides law.

6 Conclusion

We have proposed a method to answer yes/no questions from legal bar exams related to civil law statutes. We used a Siamese convolutional neural network model using word and negation features for input, and extracted features only from the relevant segment in the statute with the query sentence by analyzing the legal sentence structures. The Siamese convolutional neural network shows good performance in these tasks because (1) it requires a fewer parameter to train for, which in turn means less data required and less tendency to overfit, and (2) it makes more sense to use similar model to process similar input pairs. We showed the improved performance over the system using a single convolutional neural network in the COLIEE competition.

Acknowledgements. This research was supported by Alberta Machine Intelligence Institute (www.amii.ca). We are indebted to all our colleagues who have participated in the COLIEE competition over the last 5 years, especially Ken Satoh (NII), Yoshinobu Kano (Shizuoka U), and Masaharu Yoshioka (Hokkaido U).

References

1. Kim, M.-Y., Kang, S.-J., Lee, J.-H.: Resolving ambiguity in inter-chunk dependency parsing. In: Proceedings of 6th Natural Language Processing Pacific Rim Symposium, pp. 263–270 (2001)
2. Bdour, W.N., Gharaibeh, N.K.: Development of yes/no arabic question answering system. *Int. J. Artif. Intell. Appl.* **4**(1), 51–63 (2013)
3. Nielsen, R.D., Ward, W., Martin, J.H.: Toward dependency path based entailment. In: Proceedings of the Second PASCAL Challenges Workshop on RTE (2006)
4. Yu, L., Hermann, K.M., Blunsom, P., Pulman, S.: Deep learning for answer sentence selection. arXiv preprint [arXiv:1412.1632](https://arxiv.org/abs/1412.1632) (2014)
5. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of ACL (2014)
6. Kim, M.-Y., Xu, Y., Lu, Y., Goebel, R.: Question answering of bar exams by paraphrasing and legal text analysis. In: Kurahashi, S., Ohta, Y., Arai, S., Satoh, K., Bekki, D. (eds.) *JSAI-isAI 2016. LNCS (LNAI)*, vol. 10247, pp. 299–313. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61572-1_20
7. Yih, W., He, X., Meek, C.: Semantic parsing for single-relation question answering. In: Proceedings of ACL (2014)
8. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Proceedings of Acoustics, Speech and Signal Processing (ICASSP), pp. 8609–8613 (2013)
9. Deng, L., Abdel-Hamid, O., Yu, D.: A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: Proceedings of Acoustics, Speech and Signal Processing (ICASSP), pp. 6669–6673. IEEE (2013)
10. Bordes, A., Chopra, S., Weston, J.: Question answering with subgraph embeddings. In: Proceedings of EMNLP (2014)
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint [arXiv:1207.0580](https://arxiv.org/abs/1207.0580) (2012)
12. Kouylekov, M., Magnini, B.: Tree edit distance for recognizing textual entailment: estimating the cost of insertion. In: Proceedings of the Second PASCAL Challenges Workshop on RTE (2006)
13. Goyal, N.: LearningToQuestion at SemEval 2017 task 3: ranking similar questions by learning to rank using rich features. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 310–314 (2017)
14. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: *AAAI*, pp. 2786–2792, February 2016
15. Das, A., Yenala, H., Chinnakotla, M., Shrivastava, M.: Together we stand: siamese networks for similar question retrieval. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 378–387 (2016)
16. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. In: *Advances in Neural Information Processing Systems*, pp. 737–744 (1994)
17. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: *IEEE Computer Vision and Pattern Recognition*, pp. 539–546 (2005)

18. Kim, M.-Y., Xu, Y., Goebel, R.: Applying a convolutional neural network to legal question answering. In: Otake, M., Kurahashi, S., Ota, Y., Satoh, K., Bekki, D. (eds.) JSAI-isAI 2016, pp. 282–294. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-50953-2_20
19. Lyu, C., Lu, Y., Ji, D., Chen, B.: Deep learning for textual entailment recognition. In: IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 154–161 (2015)
20. Neculoiu, P., Maarten, V., Mihai, R.: Learning text similarity with siamese recurrent networks. In: Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 148–157 (2016)
21. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint [arXiv:1508.05326](https://arxiv.org/abs/1508.05326) (2015)
22. Rocktaschel, T., Grefenstette, E., Hermann, K.M., Kocisky, T., Blunsom, P.: Reasoning about entailment with neural attention. arXiv preprint [arXiv:1509.06664](https://arxiv.org/abs/1509.06664) (2015)
23. Yin, W., Schutze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. arXiv preprint [arXiv:1512.05193](https://arxiv.org/abs/1512.05193) (2015)
24. Liu, P., Qiu, X., Huang, X.: Modelling interaction of sentence pair with coupled-LSTMs. arXiv preprint [arXiv:1605.05573](https://arxiv.org/abs/1605.05573) (2016)
25. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. arXiv preprint [arXiv:1511.06361](https://arxiv.org/abs/1511.06361) (2015)
26. Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R., Jin, Z.: Natural language inference by tree-based convolution and heuristic matching. In: Proceedings of the Conference on Association for Computational Linguistics (2016)
27. Wang, S., Jiang, J.: Learning natural language inference with LSTM. arXiv preprint [arXiv:1512.08849](https://arxiv.org/abs/1512.08849) (2015)
28. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint [arXiv:1605.09090](https://arxiv.org/abs/1605.09090) (2016)
29. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. arXiv preprint [arXiv:1601.06733](https://arxiv.org/abs/1601.06733) (2016)
30. Bowman, S.R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C.D., Potts, C.: A fast unified model for parsing and sentence understanding. arXiv preprint [arXiv:1603.06021](https://arxiv.org/abs/1603.06021) (2016)
31. Parikh, A.P., Tackstrom, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. arXiv preprint [arXiv:1606.01933](https://arxiv.org/abs/1606.01933) (2016)
32. Sha, L., Chang, B., Sui, Z., Li, S.: Reading and thinking: re-read LSTM unit for textual entailment recognition. In: Proceedings of COLING: Technical Papers, pp. 2870–2879 (2016)

SKL2017

4th International Workshop on Skill Science

Tsutomu Fujinami

Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi,
Ishikawa 923-1292, Japan

1 Aims and Scope

Human skills involve well-attuned perception and fine motor control, often accompanied by thoughtful planning. The involvement of body, environment, and tools mediating them makes the study of skills unique among researches of human intelligence. The symposium invited researchers who investigate human skill. The study of skills requires various disciplines to collaborate with each other because the meaning of skills is not determined solely by efficiency, but also by considering quality. Quality resides in person and often needs to be transferred through the master-apprentice relationship. The procedure of validation is strict, but can be more complex than scientific activities, where everything needs to be described by referring to evidences. We are keen to discussing the theoretical foundations of skill science as well as practical and engineering issues in the study.

2 Topics

We invited wide ranges of investigation into human skills, from science and engineering to sports, art, craftsmanship, and whatever concerns cultivating human possibilities. Fourteen pieces of work were presented at the workshop, including one invited lecture. Two selected pieces of work are included in the issue from our workshop.

The article titled “A study on intellectual tasks influenced by the embodied knowledge”, written by Itsuki Takiguchi and Akinori Abe, reports their findings of skill transfer between different tasks. Their result also identifies what can be termed as skill by considering its transferability.

The workshop organizer is honored to present the report and hopes that the reader will find it interesting and will be stimulated to look into the field of Skill Science.



A Study on Intellectual Tasks Influenced by the Embodied Knowledge

Itsuki Takiguchi¹(✉) and Akinori Abe²

¹ Graduate School of Humanities and Studies on Public Affairs,
Chiba University, Chiba, Japan
moonbow.shooting@gmail.com

² Chiba University, Chiba, Japan

Abstract. I have an assumption that knowledge of the known intellectual task will similarly influence on the new one. By using origami performances, it was verified the existence of embodied knowledge of the known intellectual task made the performance of unknown similar tasks better. In this paper, I defined as embodied knowledge is an advanced technique that the body has learned by experience, and it is skilled to the extent that it does not require assistance from the outside when executing it. Experiments were carried out as the origami performance of folding cranes and phoenixes. The performance of folding phoenixes consists of the common part of folding cranes and folding phoenixes and the unique part of folding phoenixes. Because of comparing the execution time of the folding cranes with that of folding phoenixes, the following three observations were obtained. (1) If they had the embodied knowledge of folding cranes, they could finish the task of folding phoenixes more quickly than those who do not have the embodied knowledge. (2) Significant differences due to the presence or absence of the embodied knowledge were observed only in the performance of the common part. (3) Once if they have experienced to fold cranes, it was possible to complete the task of folding phoenixes even if they did not have the embodied knowledge of folding cranes. As shown in the above results, the embodied knowledge of folding cranes influenced only on the common part of folding cranes and folding phoenixes. In the common part of folding cranes and folding phoenixes, only differences due to the presence or absence of experiences were observed, and no difference was found due to the proficiency inexperience. The reason for the increase in the efficiency of the new intellectual task like the known intellectual task by the embodied knowledge is that only efficiency was increased because the efficiency of their common part was increased. Thus, we cannot conclude that experiences have played some roles in the unique part. In addition, as shown in the above results, once they have experienced to fold cranes, they will be able to obtain the knowledge of how to fold the cranes.

Keywords: Embodied knowledge · Intellectual task · Origami performance

1 Introduction

When we look at the various actions from our morning getting up to sleep at night, it can be said that they are various kinds of task and accumulation of actions. In such tasks, even if it were intellectual tasks that are somewhat complicated, such as cooking, sports, creative activities, if we are always doing them, we can perform their intellectual task without any problems. Because we have knowledge gained as experiences for those intellectual tasks, that is, intuition and feeling in a task, movement, hand working, etc. by experiencing something.

The technique of the motion obtained by such experiences is called embodied knowledge. In this paper, I defined as embodied knowledge is an advanced technique that the body has learned by experience, and it is skilled to the extent that it does not require assistance from the outside when executing it.

On the other hand, when executing a new task, it is impossible to task as it is because it does not have that experience. To solve this problem and execute a new task, we think that we are promoting understanding of new task by using knowledge of known task like that. Therefore, it can be said that existing experiences are applied to for understanding and executing a new task. When working on a completely new task, there are many scenes that require external assistance to achieve it. However, there are parts that do not require assistance and parts that can finish work earlier than others. It is believed that embodied knowledge was used for that part.

In Maruyama's study, that purposed the elucidation of image formation process of folding using Origami "Yakkosan of hanging display" which is a transforming of "Yakkosan" (Maruyama 2015). The sample photograph of "Yakkosan of hanging display" was presented, and the participant performed a task of folding the same as that. And she observed the process of folding until the task was completed. However, the study did not focus on the influence of skills and knowledge on the tasks.

In this paper, we focus on the influence of embodied knowledge on new tasks using intelligent task called Origami.

2 Experiment 1

2.1 Purpose

We examined the influence of embodied knowledge on intellectual task using the intellectual task of folding origami.

2.2 Method

(1) Participant

16 college students (6 males, 10 females) were participated. Both were undergraduates enrolled at the Faculty of Literature, Chiba University.

(2) Procedure

We asked the participants to fold two types of origami, crane, and phoenix, and photographed and visually recorded the situation from the front with a video camera. After that, we output the captured image of the action underway to the personal computer and asked questions while confirming with the participant. The experiment time was 70 min. The method of recording the experiment and setting of the experiment time were made with reference to the task of folding the “Yakkosan of hanging display” of Maruyama (2015). Tasks were conducted in the order of crane and phoenix. Presenting the sample (Figs. 1 and 2) to each participant, participants folded the same origami as the sample. When they could not fold the origami, they would be presented the hints (Fig. 3) on the request from them. The hints would be shown according to the stage of the folding the origami. The hint was created based on the descriptions in Origami Daizenshu (Fukami 2000).

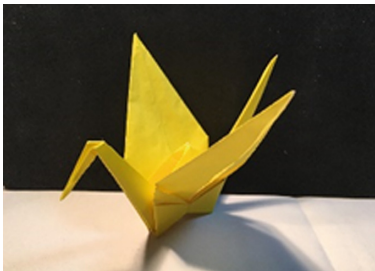


Fig. 1. Crane sample



Fig. 2. Phoenix sample

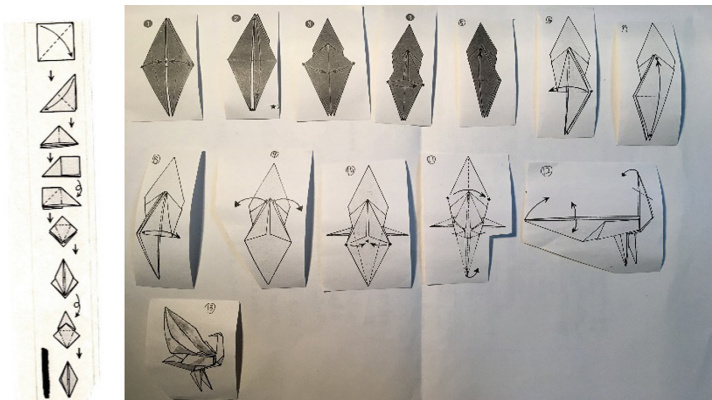


Fig. 3. An example of a hint

2.3 Result

Based on the shot image, the Crane is the task time from the beginning of folding to the completion, Phoenix starts from the folding stage to the stage of “Tsuru no Kiso” (see Fig. 4) which is a common part between crane and phoenix (Hereinafter referred to as “process α ”), the stage from “Tsuru no Kiso” to completion (Hereinafter referred to as “process β ”), the total working time from the beginning of creation to completion, were measured.

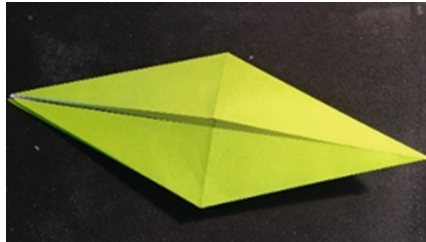


Fig. 4. Tsuru no Kiso

We divided the result of measurement into a group that knows how to fold a crane (hereinafter referred to as group A) and a group that did not know how to fold a crane (hereinafter referred to as group B). The classification was carried out based on the answer to the question after the end of the experiment. Participants who had folded cranes in the past and who knew how to fold were categorized to the group A. And the others were categorized to the group B. We classified result into each process as follows It is a graph (Fig. 5).

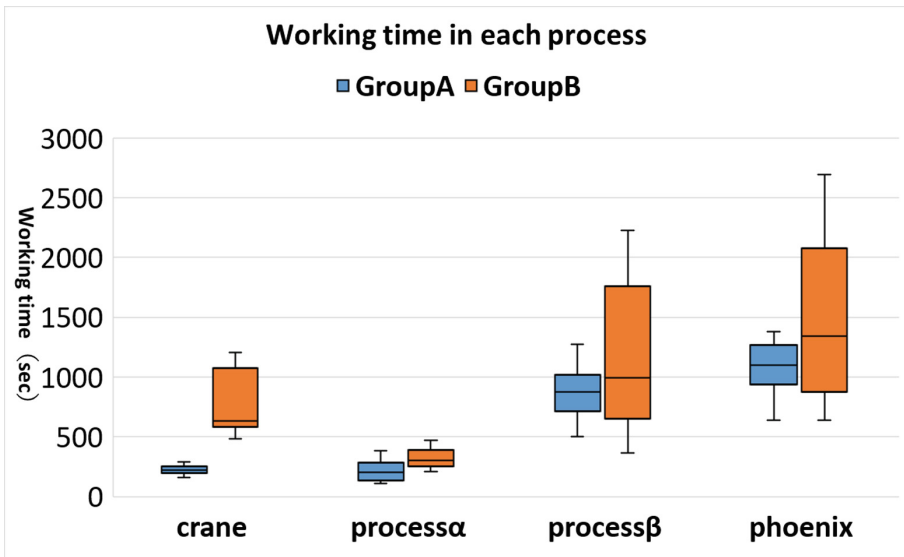


Fig. 5. Working time in each process of Experiment 1

Based on the obtained results, Wilcoxon ranked sum test of crane and phoenix task hours for Groups A and B showed a significant difference between Group A and Group B only in the crane task hours. ($Z = -3.254$, $p\text{-value} = 0.0002498$)

Also, in the question after the end of the assignment, there were differences in answers among the groups on questions about cranes, such as “Where are the difficulties in folding a crane?” “Was there a part you care about folding a crane?” Regarding the question “What is the difficulty in folding a crane?”, all the participants in group A responded that “there were no difficulties”, whereas in group B, “it was difficult to form a crane head and tail”, “It was difficult to grasp the whole form”, “I did not know how to fold itself”, etc.

For the question “Was there a part you care about folding a crane?”, Group A answered, “I did not care about it” and “I folded carefully as closely as possible to the sample.” For the same question, Group B answered, “I was careful not to make a wrong fold”.

For the question about phoenix, there were no characteristic differences in responses among the groups

The dot plot of the data obtained in Experiment 1 is shown below (Fig. 6).

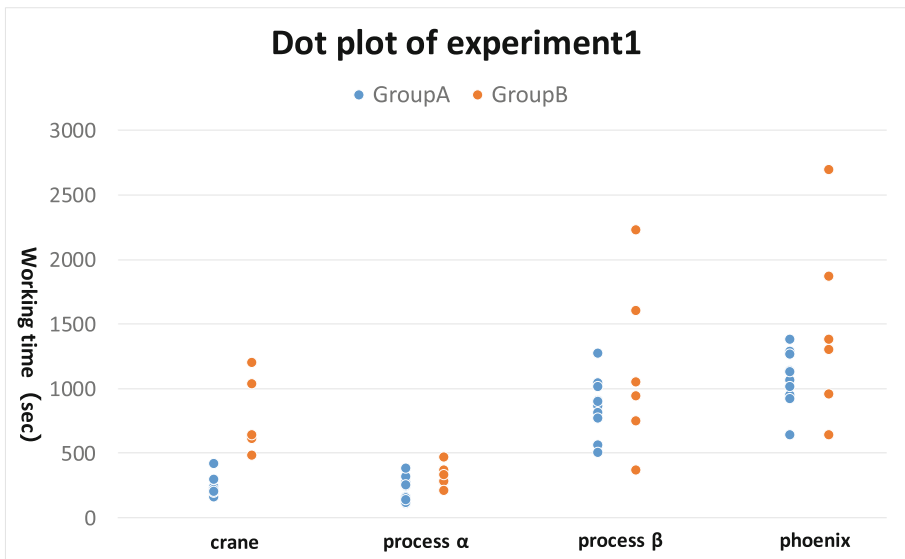


Fig. 6. Dot plot of Experiment 1

As seen from Fig. 6, it is seen that the variation of data varies greatly between Group A and Group B, even if there is no significant difference between the groups. From this, it is suggested that group A can perform more stable performance than group B.

2.4 Discussions

From Experiment 1, it was possible to obtain a result that the significant difference was observed between Group A and Group B only in the working hours of crane. From this result, it can be said that group A completed the task of crane significantly more quickly than group B. On the other hand, this also indicates that there was no significant difference in task time between groups A and B except for the working time of crane. Paying attention to the Crane and process α , the task contents of them were almost equivalent. However, despite the significant difference in the working time of the Crane, it is understood that the significant difference is not seen in the process α . Regarding questions after the end of the assignment, differences were found between Group A and Group B in questions about Crane, but differences were not found between the groups on the question about Phoenix.

Based on these facts, it seems that during the experiment, after the completion of the cranes task of Group B, there seems to be an influence that changed from the state before task execution. To verify the influence, looking at the group B in Fig. 4, we can see that the task time of the process α is shorter than the working time of the crane. It seems that task time has been shortened to the extent that there is no difference. Because of the task presentation order, they already experienced folding a crane at the time of starting process α in group B. Therefore, in Experiment 1, it is suggested that all participants became having experience of cranes at the time of the task of phoenix.

2.5 Further Issues

From the analysis of the results obtained in Experiment 1, it was suggested that all participants had embodied knowledge of crane at the time of performing the phoenix task. It is necessary to have participants who do not have cranes experiences perform the task of phoenix without having to acquire embodied knowledge. Therefore, experiments are carried out using similar participants, and the tasks are carried out in the order of Phoenix cranes rather than Crane, Phoenix in order. Since it is thought that all participants can perform the task of Phoenix without acquiring new experiences, it is necessary to perform a new experiment in which the order of experiment 1 and the task are exchanged.

3 Experiment 2

3.1 Purpose

In experiment 1, because of performing tasks in the order of cranes and phoenix, all the participants experienced folding the crane at least once at the beginning of folding phoenix. In other words, it is thought that all participants had acquired experience of cranes. We would like to verify the influence of existing experience of cranes on Phoenix without changing the order of tasks to acquire new experiences.

3.2 Method

(1) Participant

Twenty college students (8 males, 12 females) participated in. Both were undergraduates enrolled at the Faculty of Literature, Chiba University.

(2) Procedure

We asked the participants to fold two types of origami, crane, and phoenix, and photographed and visually recorded the situation from the front with a video camera. After that, we output the captured image of the action underway to the personal computer and asked questions while confirming with the participant. The experiment time was 70 min. The method of recording the experiment and setting of the experiment time were made with reference to the task of folding the “Yakkosan of hanging display” of Maruyama (2015).

The order of presenting the assignment is in the order of phoenix and crane. Tell us about presenting the sample (Figs. 1 and 2) to each participant and folding the same thing, presenting the hint (Fig. 3) stepwise if there is a request from the participant, we tackled the issue.

The difference from Experiment 1 is that the order of presenting the tasks was changed from the order of crane, phoenix to phoenix, crane in the order, and the rest is the same as Experiment 1.

3.3 Result

As in Experiment 1, the process α of the phoenix, the process β , and the total working time were measured. As for cranes, because there were many participants who were unable to carry out the task due to the experiment time, we did not use it for this analysis. We divided the result of measurement into a group that knows how to fold a crane (hereinafter referred to as group C) and a group that does not know how to fold a crane (hereinafter referred to as group D). The classification was carried out based on the answer to the question after the end of the experiment. Participants who had folded cranes in the past and who knew how to fold were group C. And the other was group D. We classified it into each process as the following graph (Fig. 7).

Based on the obtained results, Wilcoxon rank sum test was conducted for each working process of Phoenix against Groups C and D. As a result, a significant difference was observed in Step α . ($Z = -3.0237$, $p\text{-value} = 0.0004128$)

There was also a significant difference in the working time of the entire phoenix. ($Z = -2.9292$, $p\text{-value} = 0.0008256$)

In the question after the end of the assignment, in group C, we got responses to mention similarities between phoenix and crane. In Group D, only answers about phoenix.

The dot plot of the data obtained in Experiment 2 are shown below (Fig. 8).

As seen from Fig. 8, it is seen that the variation of data varies greatly between Group C and Group D, even if there was no significant difference between the groups. From this, it is suggested that group C can perform more stable performance than group D.

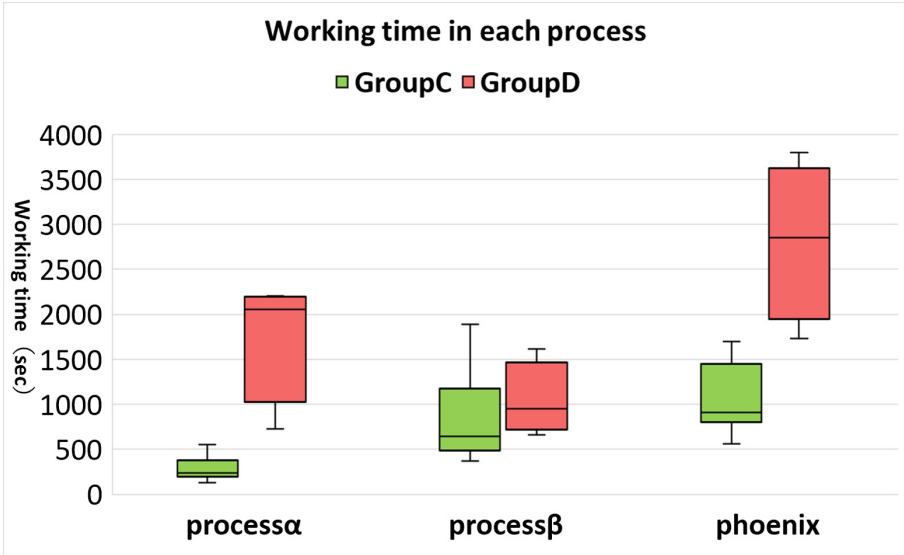


Fig. 7. Working time in each process of Experiment 2

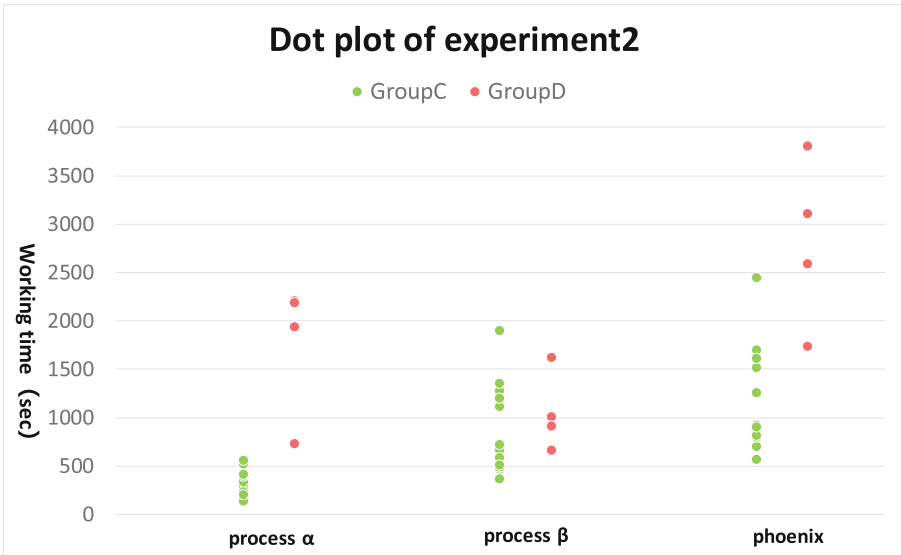


Fig. 8. Dot plot of Experiment 2

3.4 Discussions

From Experiment 2, significant differences were found in the working time of process α , phoenix in Groups C and D. From this result, it can be said that Group C completed the process α , phoenix significantly more quickly than Group D. This means that the hypothesis that experiential knowledge functioned well for similar tasks, even if it was a new task. The less time it takes to complete the task as the more task experiences like that task have been performed.

Also, in the question, after the task was completed, in group C, responses were mentioned referring to the similarity between crane and phoenix, whereas answers to phoenix were not obtained in group D at all in response to crane only the result was obtained. From this result, it is possible to point out the phoenix has a common part of the crane in case of having experience of cranes. Having experience of cranes helps to understand phoenix. This also supports the hypothesis that it is effective to task a new task to have experiences like a new task.

4 Discussions

Based on the data obtained in Experiment 1 and Experiment 2, comparison of data between experiments was conducted for group A and group C which are groups knowing how to fold crane, group B and group D which are groups does not know how to fold crane. Wilcoxon rank sum test was conducted for each task time among groups for each, and significant differences were found in the working time of process α between groups B and D. ($Z = -2.5584$, $p\text{-value} = 0.009524$) From this result it can be said that group B was able to finish the task significantly more quickly than group D. The above result shows that the time to completion of the process α , which is a common part of phoenix and crane, was significantly faster group B than group D. For this reason, the reason there was no significant difference between Group A and Group B in the phoenix task of Experiment 1 is that the participants in Group B were acquiring cranes experience at the beginning of phoenix's task. It can be said that the consideration in Experiment 1 to be reinforced. Based on the results of comparison between the above experiments and the results obtained respectively in Experiment 1 and Experiment 2, the results obtained in this study are summarized as follows. (1) If they had the embodied knowledge of folding cranes, they could finish the task of folding phoenixes more quickly than those who do not have the embodied knowledge. (2) Significant differences due to the presence or absence of the embodied knowledge were observed only in the performance of the common part. (3) Once if they have experienced to fold cranes, it was possible to complete the task of folding phoenixes even if they did not have the embodied knowledge of folding cranes. All these results support the hypothesis in this research that it is effective to task a new task to have experiences like those for a new task.

On the other hand, as shown in (2), the experiences of cranes functioned effectively only in parts like cranes in the task of phoenix, and in the part with low relevance to the crane, the existence of cranes experience did not give a significant difference. In other words, existing experiences are affecting only the part of the new task which is like the

existing task, which can be executed significantly by the existing experience, and similarity with the existing task will be diminished at all. It can be said that existing experiences do not have a significant influence at all. Even if it seems that having embodied knowledge of a certain task functions effectively for another task, this task effectively for similar parts in the task, so task can also be said to be effective and it can be thought that it is not that experiential knowledge was applied to the whole task, but experiential knowledge only affects the corresponding one.

In addition, as shown in (3), the influence of existence or nonexistence of experiential knowledge on task time is significant, but there is a big difference in task time among people with experience there was no significant difference in task time between those who were considered experienced at the time of the experiment and those who had experienced before. This is because the presence or absence of experience is the most important factor for the task time of the intellectual task of folding origami, how much knowledge has experience when to acquire experience such experiences. It can be thought that elements in intellectuals may not have much influence.

If it is assumed that only the presence or absence of experience has influence on the working time of origami, regarding the mastery of the movement touched in Suwa (2015), we have accumulated the experience of how to fold how much task time, it can be said that there is no effect even if it gets embodied knowledge. Then, what part of the influence due to the difference in experiences appears? Considering elements other than the task time of folding an origami, verify the difference in the part relating to the completeness of the task such as the politeness of folding or the small degree of reworking. It is thought that it can be done. In such parts, there may be differences among people with experience.

In Suwa (2015), as an ideal form of meta-cognition of the body, “Relationship of equality, neither language nor body are the main,” “Linking the stable feeling of ourselves to the feeling.” And by creating the relationship between the words and the words by yourself, we will spin our original words to drive the body. Focusing on this “spinning your own original language to drive the body”, even those who have embodied knowledge that did not see a significant difference in working time, there is a possibility that a significant difference may appear in terms of verbalization towards.

In this study, the participants verbalized by subjects are only self-assessment of tasks, and they do not verbalize parts such as awareness about actions and understanding of work processes. Also, there was no difference in self-evaluation among participants in self-assessment of tasks by participants. However, there was a difference such as the fact that the participants were nearly equal in the completeness of the tasks, which one was doing well, which was beautifully done, and it is certain that comparing the tasks It can be said that there is a difference in its perfection degree.

In this study, since the evaluation to the task was only the self-assessment of the participant himself or herself, we did not externally evaluate the completeness of the task, but from this it is possible to externally objective by evaluating, it can be considered that differences in empirical knowledge between each participant can be confirmed numerically as a difference in evaluation.

The phoenix task is almost the same as the crane task from the beginning to the middle. Therefore, it is thought that it was a task to fold the crane in the middle and then to fold the individual part of phoenix. As a result, individual part of the phoenix

became a completely separated part from the crane, and it seems that the application of embodied knowledge, generalization could not be verified in that part.

5 Conclusions

In this study, for a certain new intellectual task, we examine the influence of knowledge of existing intellectual task like that on new intellectual task, using intelligent task called origami task time, and obtained the following two conclusions.

- (1) If you have embodied knowledge of existing intellectual task like that for a new intellectual task, you could do the task significantly more quickly than if you did not have experience, but embodied knowledge influence only correspondence part.
- (2) Regarding working time in an intelligent task of folding origami, the presence or absence of experience is the most important factor, and the elements in embodied knowledge have no influence on working time.

For the further development of this research based on the above conclusion, the following problems can be considered.

First, there is an improvement of the hint of the folding method used in the experiment. In the experiment, we presented hints in the form of presenting hints in order as requested from participants, but the meaning of hint presentation is to present to guide the next stage to present task. However, in presenting hints, it is possible that the hint provided the participant more information than guiding the next step. I cannot completely deny the possibility that participants themselves hindered their task because of misinterpretation of presented hints. To solve this problem, it is conceivable to propose experiments that do not use hints when performing similar experiments. When using hints, we devised a hint that gives participants information other than information on guidance to the next stage, so that seeing hints will not affect unnecessarily the performance of the participants alternatively, rather than doing the presentation of hints at the request of the examinee, it is necessary to control the influence of the hint by presenting in order by time task.

Second, in this study, experiments were conducted on a single experience and a single new intellectual task, but in actual daily scenes, there are a plurality of tasks like a certain task, because there is embodied knowledge, we think that expansion of the object is necessary to conduct research on the experience as knowledge and its influence on intellectual task. With respect to the extension of the object, it is given to each task for multiple experiences considered to be related to a certain task, for each task in the case where a single experiential knowledge is affecting a plurality of task a study of the influence that can be considered.

In the case of targeting multiple experiences, it can be said that it is necessary to verify which part of the task affects each of the experiences and verify each other's influence on the experiences. When there are overlapping parts between experiences in multiple knowledge experiences, it is thought that as for the overlapping part, more experience is gained than in the case where each experience has the knowledge, it is thought that further development will be given to this research by conducting research such as verifying that fact.

Finally, it is important to verify the influence other than task time on intellectual task by experiential knowledge. In this study, experiments were carried out focusing on task time only on the influence on experience intellectual task given by experienced knowledge, but no differences in experiential knowledge were found among experienced persons in working hours. However, if it is origami, there are differences in experiential knowledge in terms of the completeness of the task, such as the precision of folding, politeness, or the skill of the task itself of folding origami, the awareness, and understanding of the task. In addition, this is described in Suwa (2015). It seems that there is a great correlation with the promotion of proficiency in behavior with comprehension by the connection between the experience in the metacognition method and the word (concept).

From these facts, to verify the influence of experiential knowledge on intellectual task, it is necessary to focus on experiential knowledge itself and to look at the difference within experienced knowledge in more detail. Therefore, to extend this research, it is effective to focus on embodied knowledge themselves and verbalize to embodied knowledge. For example, comparing self-evaluation of intellectual tasks, letting participants explain the process of intellectual tasks, and asking subjects for the explanation about work points, and so on.

The phoenix task is almost the same as the crane task from the beginning to the middle. Therefore, depending on the participants, there is a possibility that it was not a task of cranes and phoenix but a task of cranes and deformation of cranes. Therefore, when conducting similar experiments in the future, it is necessary to change the contents of the task so that all participants can perform tasks under the same conditions.

References

- Suwa, M.: The essence of knowledge that can be found because of first-person research, Recommendation of first-person research, the new trend of intelligence research. Kindai Kagaku Sha Co., Ltd., Tokyo (2015)
- Maruyama, M.: How do we fold Origami?(12): process analysis of “folding” image formation from finished product, Annual convention of the Japanese Association of Educational Psychology(57), Japanese Association of Educational Psychology (2015)
- Crawley, M.J.: Statistics: An Introduction Using R, Trans. by K. Nomakuchi and Y. Kikuchi. Kyoritsu Shuppan Co., Ltd. (2008)
- Suwa, M.: Metacognitive verbalization as a tool for acquiring embodied expertise. J. Japan. Soc. Artif. Intell. **20**(5), 525–532 (2005)
- Fukami, E.: Origami Daizenshuu, Seibido Shuppan (2000)
- Murakami, H.: Total Analysis Standard Nonparametric Method. Asakura Publishing Co., Ltd., Tokyo (2015)

AI-Biz2017

Artificial Intelligence of and for Business (AI-Biz2017)

Takao Terano¹, Hiroshi Takahashi², and Setsuya Kurahashi³

¹ Tokyo Institute of Technology, Japan

² Keio University

³ University of Tsukuba

1 The Workshop

In AI-Biz2017 held on November 14, one excellent invited lecture and eleven cutting-edge research papers were presented with a total of 27 participants. The workshop theme focused on various recent issues in business activities and application technologies of Artificial Intelligence to them.

The invited lecture was “Deep learning and South-East Asian Issues” by Prof. Kiyota Hashimoto of Prince of Songkla University Phuket Campus, Thailand. In his presentation, It was reported that the Deep learning research is actively conducted in Southeast Asia, mainly natural language processing, and also discussed the current state of university education in Southeast Asia.

The AI-Biz2017 was the second workshop hosted by the SIG-BI (Business Informatics) of JSAI and we believe the workshop was successful, because of very wide fields of business and AI technology including for evaluating signage system, ambiguous information, stock price prediction, medical insurance market, Korean market studies, a portfolio selection algorithm, patent classification, and so on.

2 Papers

18 papers were submitted for the workshop, and eleven papers were selected to be presented in the workshop. After the workshop, they were reviewed by PC members again and seven papers were finally selected. Followings are their synopses.

Eriko Shimada, Shohei Yamane, Kotaro Ohori, Hiroaki Yamada and Shingo Takahashi analyse a signage system installed in large-scale facilities such as airport passenger terminals and stations in order to make facility users feel good comprehensively. Agent-based models of airport terminals have been proposed to represent the behavioral characteristics of passengers and the essential features of signs. This study provides some evaluation results focusing on information message and location arrangement, which are part of the essential components.

Takaya Ohsato, Kaya Akagi and Hiroshi Deguchi address the task of data conversion of enterprise export/import ratio using text mining. An input-output table has been generated based on corporate data and information by using an algorithm. In this work, they have structured the data, and apply text mining to the information described

by free descriptions of import/export ratios of direct trade companies to sales. They simulate using structured data and propose an algorithm to construct an input-output table that includes foreign transactions.

Kei Nakagawa, Mitsuyoshi Imamura and Kenichi Yoshida propose a method to predict future stock prices with the past fluctuations similar to the current. As the levels of stock prices differ depending on the measured period, they have developed a scaling method to compensate for the difference of price levels and the proposed new method; specifically, they propose indexing dynamic time warping (IDTW) to evaluate the similarities between time-series data. To demonstrate the advantages of the proposed method, they analyze its performance using major world indices.

Ren Suzuki, Yoko Ishino and Shingo Takahashi have developed an agent-based model of consumer's behavior in purchasing medical insurance products and analyzed the characterization of consumers behavior to establish effective marketing strategies for the products. The information propagation model of purchasing behavior has difficulty estimating the values of parameters only from ordinary marketing surveys. To tackle this problem, they have developed a method of estimating the probability parameters of agent's behavior using Bayesian network based on questionnaire survey data, and then evaluated the effectiveness of the method by applying it to the actual insurance market.

Sungjae Yoon, Aiko Suge and Hiroshi Takahashi analyze the influence of news articles on Korean stock markets with high-frequency trading data. The news is an important source of information for investment decision-making. Especially, they focus on analyses of the relationship between news articles and financial markets. They also analyze differences in market reactions according to language (English or Korean) of news articles and present three case studies.

Kazunori Umino, Takamasa Kikuchi, Masaaki Kunigami, Takashi Yamada, and Takao Terano have examined the "short-term Mean Reverting Phenomenon" from two aspects. First, they clarified that excess return can be obtained by using the short-term Mean Reverting Phenomenon for the On-Line Moving Average Reversion (OLMAR) method, which is a portfolio selection algorithm and reportedly exhibits high performance. They have examined why the method was able to maintain superiority over the long term.

Masashi Shibata and Masakazu Takahashi provide the technology structure analysis and the technology field clustering through the analyses of patent classification codes with link mining method. Knowledge extraction from patent information has been made thus far, but conventional patent analysis methods depend on personal heuristic knowledge. They focus on classification codes in the patent, and have assigned to capture the technology fields of the patent.

3 Acknowledgment

As the organizing committee chair, I would like to thank the steering committee members, The members are leading researchers in various fields:

Hiroshi Deguchi (Tokyo Institute of Technology, Japan)
Reiko Hishiyama (Waseda University, Japan)
Manabu Ichikawa (National Institute of Public Health, Japan)
Yoko Ishino (Yamaguchi University, Japan)
Hajime Kita (Kyoto University, Japan)
Hajime Mizuyama (Aoyama Gakuin University, Japan)
Masakazu Takahashi (Yamaguchi University, Japan)
Shingo Takahashi (Waseda University, Japan)
Takashi Yamada (Yamaguchi University, Japan).

The organizers would like to thank JSAI for financial support. Finally, we wish to express our gratitude to all those who submitted papers, PC members, reviewers, discussant and attentive audience.



Agent-Based Simulation for Evaluating Signage System in Large Public Facility Focusing on Information Message and Location Arrangement

Eriko Shimada¹, Shohei Yamane², Kotaro Ohori²,
Hiroaki Yamada², and Shingo Takahashi¹(✉)

¹ Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

² Fujitsu Laboratories LTD., 4-1-1 Kamiodanaka, Nakahara-ku, Kawasaki-shi,
Kanagawa 211-8588, Japan
shingo@waseda.jp

Abstract. A signage system has been installed in large-scale facilities such as airport passenger terminals and stations in order to make facility users feel good comprehensively. Agent-based models of airport terminals as a typical large-scale facility have been proposed to represent the behavioral characteristics of passengers and the essential features of signs. They, however, are not enough to evaluate signage systems, since models have not described properly pedestrian agents' characteristics especially including their view to get information from the signs. This paper develops the simulation system that includes the model to represent a relationship between pedestrian agent's view and its information search behavior. The primary purpose of the simulation system is to support the facility manager's decision to design the signage system before it is actually implemented. This paper provides some evaluation results focusing on information message and location arrangement, which are part of the essential components.

Keywords: Signage system design · Agent's view · Agent based model

1 Introduction

Signage systems are introduced typically in large-scale facilities such as airport passenger terminals or shopping malls to provide helpful information about services or shops in the facility with facility users to find an easier way to their destinations. A signage system is defined in the literature as the whole system of various signs and the contents of the signs mutually connected in the unified way to provide systematically an efficient guidance of information about the facilities [1].

The requirements of a signage system are basically the following [1].

1. It consists of multiple signs.
2. It is planned so as to provide in a mutually complementary way necessary information as a whole of signs.
3. It is planned so as that facility users continuously get information during traveling in the target space.

4. The signs are classified into appropriate categories, and the signs of the same category have a unified format.
5. There are principles or rules of the terminology and the display method concerning the signs constituting the signage system.

If these are not sufficiently satisfied, the quality of service to facility users will deteriorate. Therefore, in order for the facility users to grasp fully the information while moving, a manager of the facility should create a system in which the content of the information, the format of the expression and the location of the signs are appropriately interrelated.

A signage system has been evaluated practically from the qualitative point of view mainly using a checklist or the like based on the above five requirements. This also makes it possible to judge whether the signage system meets a certain level. However, even though the evaluation of individual items is good, the user's comfort is not sufficient as a comprehensive evaluation of the entire facility. In fact, in planning a signage system based on these qualitative requirements, problems can be often found after an actual operation. Once a signage system is introduced, it is usually difficult to change it from a physical and budgetary point of view. Therefore, from the dynamic viewpoint that the user moves based on the sign information, it is necessary to develop a simulation system that can quantitatively evaluate the signage system before implementation.

The signs consisting of the signage system are classified into two types: signs for public information and signs for regulation indication. In public transportation, it is sufficient to have three types of Direction sign (indicating the direction to the target object), Indication sign (notifying the location of the facility) and Illustrated sign (showing correlated contents with diagrams) from signs for public information and a minimum of signs for regulation indication [2].

There are Information Message, Expression Form and Location Arrangement as important attributes when developing and evaluating the signage system [1]. Information Message consists of two elements: Content and Code. Content is a display item, and code represents a term or a symbol. Expression Form consists of two elements: Mode and Style. Mode represents a display method or an illumination method, and Style indicates an appearance such as a shape, a layout or a color. Location Arrangement consists of two elements: Position and Placement. Position indicates the height of the posting and the orientation of the display surface, and Placement indicates the placement point and the placement interval on the area.

The pedestrian agent is an important component in this paper. From the viewpoint of pedestrian agent simulation, Kaneda [3] pointed out the following three usages as effective from the practical viewpoint.

1. preliminary consideration of prediction and measures of crowd accident risk
2. consideration of emergency evacuation measures in large-scale facilities
3. design of pleasant walking space in commercial facilities or in bustling space.

Then the large-scale facility of our research target is a subject that can be modeled using pedestrian agent simulation. In particular, the signage system can be evaluated effectively by using agent models that autonomously make their decisions based on information from that signage system.

The guideline issued by the Ministry of Land, Infrastructure and Transport [4] classifies the equipment for information guide into two categories: the equipment for visual display and the equipment for helping visually handicapped persons. The equipment for visual display is further classified into the signage system and the variable information display device. Then the guideline defines the signage system as the whole of the equipment for visual display of the information guidance that provides appropriate information to facility users walking around by settling the signs along the flow line of the users.

The pedestrian's view is one of the essential factors of the mechanism behind the walking behavior. For example, elderly pedestrians, than younger ones, walk slower with poorer sense of balance, cannot search information well, or behave more dangerously. These difficulties of elderly people principally come from their aging of sensing, perceiving, physical and understanding abilities. Hence, we need to consider walking characteristics of an agent to build a pedestrian agent model for evaluating the signage system.

Agent's wayfinding behavior plays a central role in a pedestrian agent model. The agent's wayfinding behavior can be classified into the behavior in selecting the route before departing and the behavior in moving along the selected route. A route selection is efficient if a pedestrian does not move a longer way than necessary, or reaches the destination avoiding obstacles [5]. Even though a map is used in selecting a route, it would be more difficult for elderly people than young to select the optimal way. These show that elderly people easily make mistakes and need more time to select routes than young [6]. Though the critical for a pedestrian is the ability not to lose the destination and direction in moving along the route selected. Aging would make reduced such ability that the pedestrian recognizes his/her current location and select the route [7]. This kind of reduction of the ability of elderly people causes the problems concerning memorizing the selected route, learning necessary attributes of the signs, understanding the special ordering of the signs or the like. Hence, we should notice that it is not easy for a pedestrian to get information from the sign while moving and to head to the destination. We see that a pedestrian easily forgets the route even though he/she recognizes it. Walking and navigating are highly visual [8]. With aging, the speed of the brain and nervous system conveying the sensation slows down. This change shows the decline of executive functions, which include the capacity for updating and monitoring information in working memory, inhibiting inconsistent or useless information, and shifting [9].

Consequently, in pedestrian behavior, wayfinding and view have an essential relationship. At large public facilities such as airports, pedestrians need maps to facilitate route planning. After route planning, there is the use of a sign as a means to mitigate forgetting routes when traveling along a selected route.

Visibility, readability, legibility, discrimination, attractiveness and vigilance are involved in the visual appearance of everyday life [10]. Visibility is a threshold value that can perceive the existence of the object when gazing. Things located remarkably far away cannot be admitted with the naked eye and can be perceived only when it approaches to some extent (visibility threshold). When approaching further, color can be perceived (color threshold) and the form can be discriminated only after closer approach (form threshold). The accuracy of discrimination generally differs according

to the distance. It is thought that a pedestrian who forgets or cannot obtain information overlooks far to seek information and takes action approaching signs.

The purpose of this paper is to develop a simulation system to evaluate qualitatively from a dynamical point of view the signage system planned before its installation, and to support a decision on designing and implementing the signage system. The simulation systems developed so far [11, 12] did not take pedestrian walking behavior into account in a proper way, and was not sufficient to evaluate the signage system in some scenarios on the location of the signs. Hence particularly focusing on a view of the pedestrian agent model, which would affect the evaluation result of the signage systems, we develop a pedestrian agent model considering the relationship of the agent's view with information search behavior, which are based on the requirements for designing the signage systems.

This paper focuses on modeling Information Message and Location Arrangement from among the three attributes Information Message, Expression Form and Location Arrangement. As for Information Message, "Content" meeting the information needs is modeled. For Location Arrangement, modeling is made for easy-to-view display surface orientation, and placement point and placement interval where information needs are generated. Also, we deal with "Content" suitable for the placement point, which represents a problem for composition of Information Message and Location Arrangement. This paper currently shows a basic framework of the evaluation model and confirms its effectiveness. Though Expression Form is not treated in this paper, the current basic structure of the evaluation model can be applied to model Expression Form on the extension line of the proposed model in this paper.

Our simulation system is composed of the environmental model representing the large-scale facility pedestrians use, the sign model representing the characteristics of the signs, and the pedestrian agent model representing the walking behavior of agents in the facility. In this paper, supposing a virtual airport terminal as a large-scale facility, we will conduct experiments of four scenarios on the design condition of the signage system to evaluate the design plans represented as the scenarios. Then the reasons why such evaluation results are shown will be considered analyzing micro dynamics by using behavior logs of agents in the simulation results.

2 Model

In this paper, an agent-based model consists of an environmental model, a signage system model and a pedestrian agent model. A pedestrian agent is generated from a given specific node, then he/she gets information from the signage system and plans to use facilities, check in for a flight and so on.

2.1 Environmental Model

Environmental model is composed of cells that pedestrian agents move over, nodes that show pedestrians' destinations when they move over cell and have specific functions,

and edges which are connections between nodes and pedestrians use to avoid obstacles (Fig. 1).

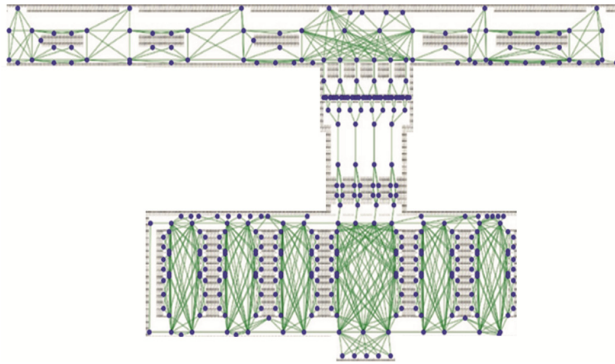


Fig. 1. Environmental model: points are nodes and lines are edges.

Cell. Cell consists of $\text{cell}(x, y, \text{floor})$, cellType , cellArea , where $\text{cell}(x, y, \text{floor})$ shows coordinate, cellType shows whether a pedestrian agent moves over the cell and cellArea means area in the large-scale facility. The variable cellarea has 5 types: Before-check in Area, After-check in Area, inspection Area, After-inspection Right Area and After-inspection Left Area.

Node. Node consists of $\text{node}(x, y, \text{floor})$, nodeType , nodeFid , nodeTime , nodeCapacity , nodePfid , nodeWaitingnum and nodeServicenum , where $\text{node}(x, y, \text{floor})$ shows coordinate, nodeType the type of the node, nodeFid the number of facility, nodeTime the number of steps that a pedestrian agent needs to use a facility, nodeCapacity the number of agents the node can accommodate, nodePfid the number of sub-facilities, nodeWaitingnum the number of pedestrian agents who are queuing to use a facility and nodeServicenum the number of pedestrian agents using facilities. Node has 6 types: Waypoint, Procedural facility, Sub-facility, Commercial facility, Goal point and Start point. Each type of node describes a different function of the node, the function of which is specified by a combination of the parameter values.

Edge. Pedestrian agents could not go straight on from the current position to the destination because some entity on the environmental model might obstruct pedestrian agents' movement and view. So operating routes is necessary for them. In our model, we define operating routes as edges which pedestrian agents use to move between facilities. Edge consists of $\text{edge}(\text{node } u, \text{node } v)$ that is a combination of nodes and $\text{edgeWeight}(\text{node } u, \text{node } v)$ that gives the distance between node u and node v .

2.2 Signage System Model

A pedestrian agent can get information from a sign within the pedestrian agent's view, if the pedestrian agent exists inside the information delivery range of the sign. Sign model has 2 types: areaSign and FacilitySign.

Area Sign. areaSign consists of sign(x, y, floor), signFid, signDirection(sightx, sighty), signTheta, signR, signR2, signTheta2, signCategory, signArea, signTime, and signRouteinfo, where sign(x, y, floor) means coordinate, signFid the number of the sign, signDirection(sightx, sighty) the direction of the sign, signTheta the angle at which the sign can pass information, signR the distance to which information can be passed, signR2 the distance at which a pedestrian agent can recognize the sign, signTheta2 the angle at which a pedestrian agent can recognize the sign, signCategory a category of facilities existing in the area sign, signArea an area the sign shows, signTime the number of steps the pedestrian agent can store information and signRouteinfo routes information the sign have.

Facility Sign. facilitySign consists of sign(x, y, floor), signFid, signDirection(sightx, sighty), signTheta, signR, signR2, signTheta2, signCategory, areaSign, signArea, signFacility, signWaitingtime, signTime, and signRouteinfo, where sign(x, y, floor) means coordinate, signFid the number of the sign, signDirection(sightx, sighty) the direction of the sign, signTheta the angle at which the sign can pass information, signR the distance to which information can be passed, signR2 the distance at which a pedestrian agent can recognize the sign, signTheta2 the angle at which a pedestrian agent can recognize the sign, signCategory a category of facilities existing in the areaSign, signArea an area the sign shows, a facility the sign shows, signWaitingtime the number of waiting steps, signTime the number of steps a pedestrian agent takes to store information and signRouteinfo routes information the sign have.

2.3 Pedestrian Agent Model

If a pedestrian agent does not have any route information and a sign is recognizable, the agent approaches the sign. A pedestrian agent can get information from a sign within the pedestrian agent's view, if the agent is inside the information delivery range. The pedestrian agent decides the facility to go based on the information the agent gets and walks toward the decided facility. If a pedestrian agent does not have any information, the agent walks randomly to search information.

Pedestrian agent model consist of agent(x, y, floor), agentCurarea, agentTheta, agentR, agentSchedulelist, agentCategorylist, agentGoalcategory, agentAreainfo, agentGoalarea, agentFacilityinfo, agentFacilitylist, agentGoalfacility, agentCategoryrecallset, agentFacilityutilityset, agentFacilitypreferenceset and agentRouteinfo, where agent(x, y, floor) means coordinate, agentCurarea the agent's current location area, agentTheta agent's viewing angle, agentR agent's cognitive ability, agentSchedulelist the list of the schedule, agentCategorylist the list of categories an agent wants to go, agentGoalcategory the category an agent selects as the destination, agentAreainfo information of the area an agent has, agentGoalarea the area an agent selects as the

destination, `agentFacilityinfo` the information of facility an agent has, `agentFacilitylist` the list of facilities an agent wants to go, `agentGoalfacility` the facility an agent selects as the destination, `agentCategoryrecallset` the probability to select each category, `agentFacilityutilityset` the utility value of each facility, `agentFacilitypreferenceset` the preference of each facility and `agentRouteinfo` the route information an agent has.

3 Simulation Model

3.1 Initial Settings

We assume that 1 execution step of the simulation stands for 1 s and do simulation for 9000 steps. First, the environmental model is generated. This is composed of generating cells, generating nodes and generating edges. From the input files, values are assigned to the parameters of each cell, each node and each edge. Next, the signage system model is generated. From the input files, values are assigned to the parameters of each sign. Then, the pedestrian agent model is generated. After generations, a flight schedule type is set at random and according to the flight schedule type, 4 schedules: emergence, check-in, security inspection/departure examination, boarding, are registered in the list. After that, a pedestrian agent's type is defined based on the given probability.

3.2 Pedestrian Agent's Information Updating

Information updating process consists of forgetting information and information updating/acquisition. In information forgetting process, if the steps that a pedestrian agent takes to store route information are elapsed, the agent erases the route information.

A pedestrian agent gets information by "cognizing" the environment. If a sign is located inside the pedestrian agent's view (Fig. 2), the agent can see the contents of the sign. Inside the sign's visual recognition range (Fig. 3) a pedestrian agent can see a sign, even if the agent cannot receive the contents of the sign. The sign's visual recognition range is composed of a sign's placement point, range of visibility and visible angle.

The agent's view is composed of their current position, viewing angle and cognitive capacity. If some entity obstructs the pedestrian agents' view, the agent cannot recognize the sign even inside their view (Fig. 4).

Information delivery range (Fig. 2) is defined by a range where a pedestrian agent can get information from a sign. The information delivery range is composed of a sign's direction, transmission range and transmission angle.

If a pedestrian agent is inside an information delivery range and a sign is inside the agent's view, the agent can update and get information from the sign. If a pedestrian agent does not have any route information and is inside the sign's visual recognition range, the agent moves toward the sign and tries to get information.

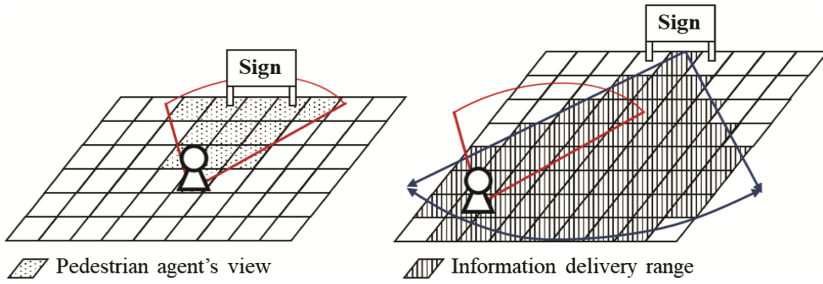


Fig. 2. The agent's view and information delivery range

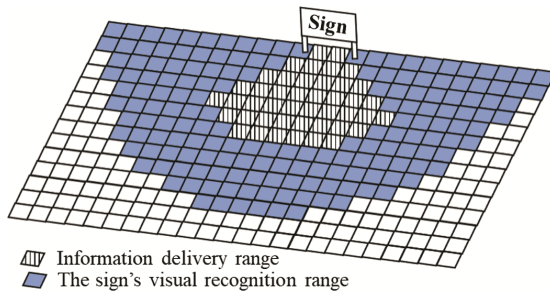


Fig. 3. Two types of sign's range

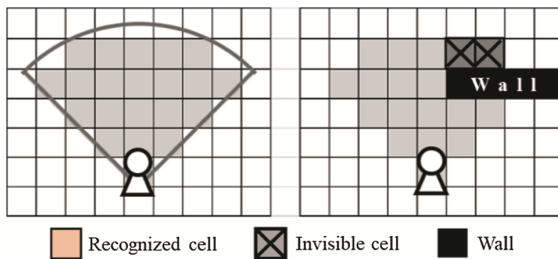


Fig. 4. A range that pedestrian agents can perceive when the wall is in sight

3.3 Pedestrian Agents' Decision Making

A pedestrian agent's decision making is composed of two processes: selecting facilities concerning boarding and selecting category/area/commercial facilities.

If the simulation steps taken from the current position of a pedestrian agent to the node of the check-in facility is expected to exceed the time scheduled to utilize the node of the facility for boarding, the agent forcibly heads to the facility for boarding.

In the category/area/commercial facility selecting process, a pedestrian agent makes decisions in multiple stages. One category is randomly selected from the category recall set and assigned to the goal category. A pedestrian agent randomly selects one area that satisfies the goal category from the acquired area information and sets it in the goal area.

If a pedestrian agent has the facility information that satisfies the goal category in the goal area, and if the total time estimated from the time spent at a facility in the facility information set and the time to move from the current position to the facility is not expected to exceed the scheduled time, that facility is added into the facility recall set. The utility value is calculated from the Eq. (1) using the preference for facility and the moving time from the current position to the facility. Then a commercial facility is selected from the facility recall set by using the multinomial logit model (2).

$$U(i) = \alpha_{hi} + \beta \cdot time_{hi} \tag{1}$$

α_{hi} represents a preference for the commercial facility i of the pedestrian agent h , $time_{hi}$ the number of steps taken to move from the current position of agent h to the commercial facility i , and β the weight for the move time (Fig. 5).

$$p(i) = \frac{\exp U(i)}{\sum_{n \in X} \exp U(n)} \tag{2}$$

$$X = \{n | facility_n \in agentFacilitylist_n\}$$

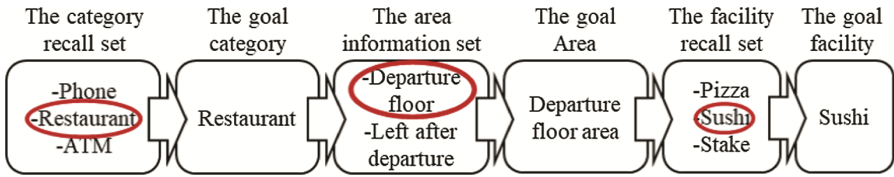


Fig. 5. Decision making process

3.4 Pedestrian Agents’ Walking Behavior

The walking behavior is composed of two processes: moving on cells and determining the target node.

In the process of moving on cells, a pedestrian agent searches 8 cells around the current position and updates the agent’s coordinate to the cell that the agent can enter and is of the shortest distance to the destination node. The walking behavior depends on the way of determining a target node, the route information, the goal facility and the current position status.

In this model a pedestrian agent has 5 types of walking behavior. If the agent has some route information, the agent does “walking according to the route.” If an agent arrives in the goal area but does not have any route information, the agent does “random walking in an area.” If the agent does not have any route information and are inside the sign’s visual recognition range, the agent does “walking to get information.” If the agent doesn’t have route information, the agent does “random walking”. If the agent uses all planned facilities, the agent does “no purpose walking” (Fig. 6).

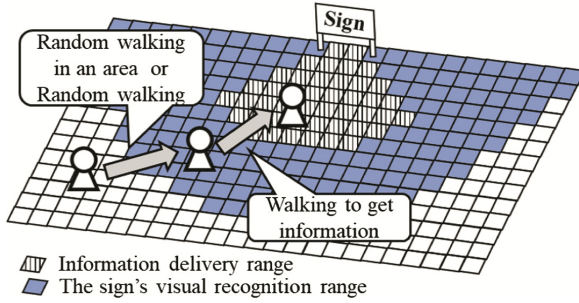


Fig. 6. How to do “walking to get information”

4 Evaluation Experiment of a Signage System

In the simulation, a virtual airport passenger terminal with 255 cells in length and 570 cells in width (one cell as 0.5 m × 0.5 m) is assumed as environmental model. The virtual airport passenger terminal is divided into five areas: before-departure-area, departure-floor-area, security-inspection/departure-examination-area, right-after-departure-area and left-after-departure-area. In this experiment, a signage system is placed in three areas of before-departure-area, departure-floor-area and security-inspection/departure-examination-area. There are totally 18 facilities in the environmental model. The commercial facilities that a pedestrian agent may recall are divided into seven categories: restaurant, phone, Exchange, ATM, bookstore, souvenir and insurance. The population of pedestrian agents is 100. We conduct the simulation of 9000 execution steps.

We first validate the proposed model of the agent’s view by evaluating Scenario 4. This scenario was evaluated in the previous study [12] and showed the problem that agent’s random walking steps often happen. Then we conduct four scenarios (Scenario 1 to 4) with different positions of signs. White squares represent sign with area information, black squares represent sign with facility information (Fig. 7).



Fig. 7. Scenarios used in this paper

4.1 Evaluation Index of a Signage System

Based on qualitative measures used to measure the quality of traffic service (LOS) [13] and the indicators of the comfort of the airport [14], we select three indices: “number of moving steps,” “waiting time,” and “achievement rate.”

The index of “number of moving steps” shows the total time taken for a pedestrian agent to arrive at the destination. Then it is possible to measure how much a pedestrian agent got lost. The index of “waiting time” shows the time a pedestrian agent was queued from arrival at the goal facility until receiving services and procedures. It can express dispersion effect of congestion by a signage system. The “achievement rate” shows the rate at which a pedestrian agent received the service he wanted.

In the pedestrian agent model, a pedestrian agent moves randomly while searching for route information. This means that even if the agent has a facility he wants to visit, the agent gets lost because the agent has no information. “Random walking” also includes situations where a pedestrian agent cannot get information from the signage system well. Hence as an evaluation index we also use the ratio of the number of “random walking” steps to the total walking time.

4.2 Experimental Results

We conducted 10 trials for each evaluation index. Each point represents one trial result.

First, we made a comparison of no agent’s view model [11, 12] with the proposed agent’s view model (Fig. 8). If there is no agent’s view model, the pedestrian agent’s search behavior increases, then the pedestrian agent has too many “random walking” steps.

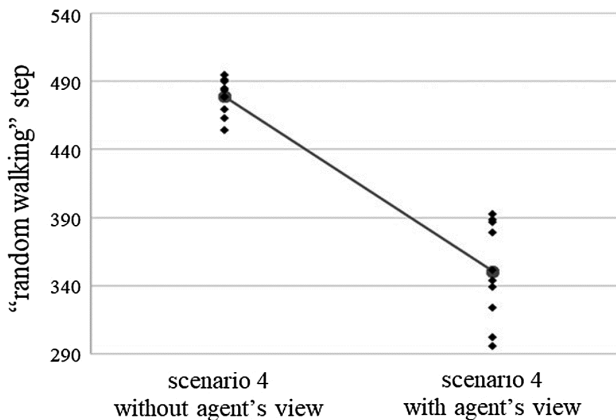


Fig. 8. Landscapes of each value of random walking step

The number of random walking steps decreases in all trials by introducing the agent’s view model. This shows that while agent’s walks without a purpose are reduced, the

number of steps of “walking according to routes” has increased. As a result, it became easier to get to a sign and get the route information.

Next, we compare the four scenarios by four indicatess.

From the evaluation of “number of moving steps”, in the case where the number of installed signs is reduced (scenario 1), the number of agents who cannot use any commercial facilities increased, since these agents repeated the behavior of information acquisition and oblivion and walked around between signs and facilities.

From the evaluations of “waiting time”, “achievement rate”, when the number of installed signatures is reduced (scenario 1), a pedestrian agent gets fewer information acquisition times, and often loses sight of the direction while moving. Hence the agent cannot arrive at a goal facility easily. As a result, the congestion degree of the facility decreases.

From the evaluations of “number of moving steps” and “achievement rate,” when a sign with facility information was installed at both sides of the facility (scenario 2, scenario 3 and scenario 4), it took for a pedestrian agent only a short time to determine the next destination after the pedestrian agent uses a commercial facility. As a result, “number of moving steps” has decreased and “achievement rate” has increased (Figs. 9, 10, 11 and 12).

From the evaluations of “achievement rate” and “random walking steps,” when the number of installed signs increased (scenario 3), a pedestrian agent is easy to get information.

Hence it becomes easy to arrive at the facility, and the opportunity to get lost in the way decreases and the number of facility use increases. Hence it becomes easy to arrive at the facility, and the opportunity to get lost decreases and the number of facility use increases, a pedestrian agent needed a long time to do “random walking” toward a sign. Hence the ratio of the number of “random walking steps” increased.

When signs with area information are located away from signs with facility information (scenario 1 and scenario 4), a pedestrian agent did not get information quickly, then “random walking steps in the area” increased.

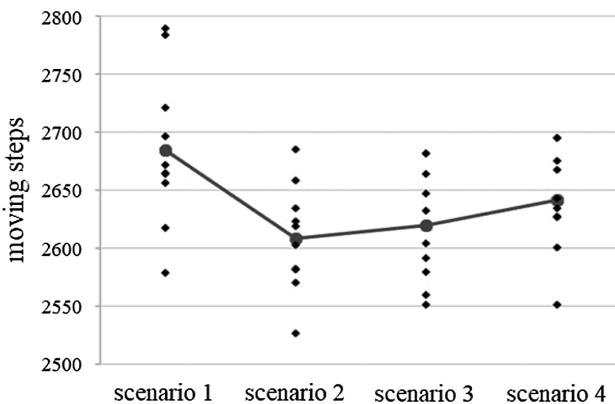


Fig. 9. Landscapes of each value of moving steps

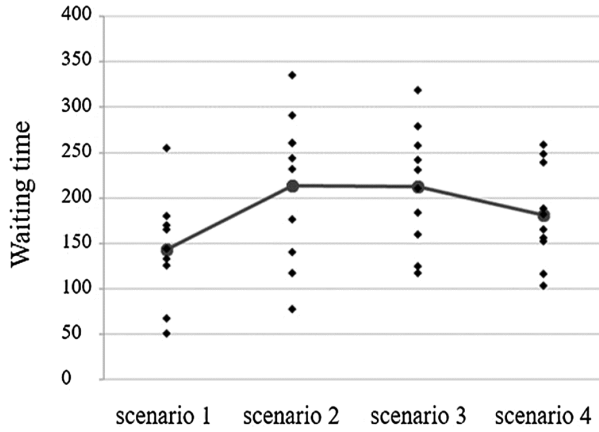


Fig. 10. Landscapes of each value of waiting time

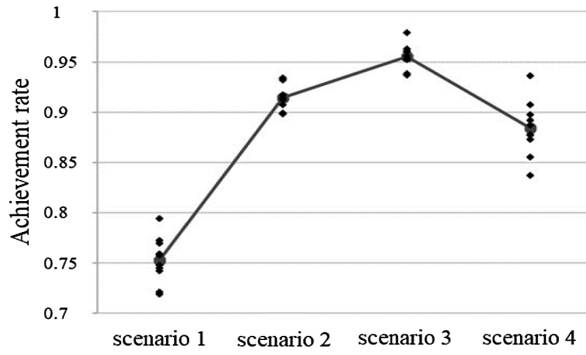


Fig. 11. Landscapes of each value of achievement rate

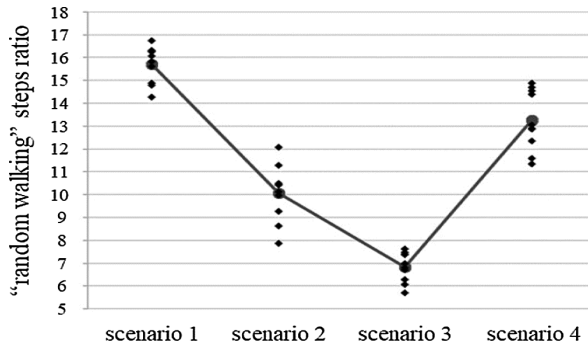


Fig. 12. Landscapes of each value of random walking steps ratio

5 Discussion

Depending on the number and arrangement of commercial facilities, the size of facilities and the attributes of passengers, it turned out that there was an optimum way of installing signs. For users of all attributes, it is important to set up signs within the range that they can move without forgetting information. When considering what kind of signage system is required in the aging society, we change the variables related to age such as the ease of forgetting information, the number of information storable at one time or walking speed.

When there are many users who could not acquire information, the degree of congestion in commercial facilities will decrease. If all the routes to the destination are indicated by signs, it is considered that the commercial facility is crowded as the users who saw the same sign, follow the same route and head to the destination. From the viewpoint of congestion degree, it is not appropriate that there are too many signs and information on signs is too detailed.

At Narita airport, to provide a new form of guidance tailored to the needs of each customer, Narita airport and NTT DATA develop interactive digital signage and provide services to provide maps and route guidance to the destination facilities in the airport [15, 16]. With such a system, it can be assumed that users with various attributes can easily acquire information. But, in a public facility with many users, it is essential to have signs on the assumption that they will be installed all the time. It also leads to the idea of universal design [17] which is the design of products and environments to be usable all people, to the greatest extent possible, without the need for adaptation or specialized design.

In the previous study, the number of moving steps was large in the scenario where no sign was set near the entrance. It was believed that if there was no sign near the entrance it would take time to get the first information. However, by introducing a model of the relationship between the information search behavior and the pedestrian's view, even if users cannot find a sign right away, they can approach sign. Therefore, it is not likely that the user can obtain information or forget it and the number of moving steps was reduced.

In this paper, only Information Message and Location Arrangement are considered from Information Message, Expression Form, and Location Arrangement which are elements when designing a signage system. As a future task, there is development of a simulation system focusing on Expression Form of the signage system.

It is important to consider the concepts of simplicity, clarity, consistency, continuity and systemicity in Expression Form of the signage system. Simplicity and clarity express readability and understandability of the information described in the sign. Consistency, continuity and systemicity represent the way how the relationship between signs is easy to understand how the signs are easy to recognize. Public transport passenger facilities guidelines [2] list pictograms, multilingual notation, design, font size, lighting, etc. as items to be considered when setting up signs. Also, conditions related to visibility. These conditions have mutual influences include brightness, contrast, size and time. These conditions have mutual influences on visibility [9]. Hence, they can only focus on the simplicity and clarity of the elements of Expression Form.

As a problem related to Expression Form, it is desirable to develop the following two systems. First, it is desirable to develop a system that distinguishes the same symbols color-coded or considers the continuity of symbols and colors. Second, we want to quantify the ease of perception, conspicuity and ease of eye attraction of multiple signs in the view and develop a system that can deal with cases where pedestrian agents have different probabilities of getting information from signs or when the amount of information that can be acquired differs.

In order to realize these systems, it is necessary to construct a quantitative model for each. These quantitative models would make possible to incorporate an expression style model into the system developed in this paper. Although the evaluation system currently proposed realizes only two of the three attributes of the sign, we should notice that we are going to model the third attribute: Expression Form as well. In this paper, incorporating all the attributes of the sign is not the main purpose. It provides a basic framework for the evaluation of the signage system that can incorporate and use the signage system attributes as needed.

6 Conclusion

In this paper, we proposed an agent-based model considering a pedestrian agent's vision. We developed an agent-based simulation tool in a large-scale facility that is possible to quantitatively evaluate a signage system in advance. It will be helpful in designing and implementing a signage system. We assumed a virtual airport passenger terminal and conducted a scenario analysis. By refining a pedestrian agent's behavior, we confirmed that the time to acquire the route information was improved, the lost before reaching the destination was reduced, and the comfort was improved.

References

1. Akase, T.: Sign System Planning: Public Space and System of Sign. Kajima Institute Publishing (2013)
2. Foundation for Promoting Personal Mobility and Ecological Transportation: Public Transport Passenger Facility Sign System Guidebook, TAISEI-SHUPPAN CO., LTD. (2002)
3. Kaneda, T.: Pedestrian Agent Simulation Start with ArtiSoc, Kozo Keikaku Engineering Inc. (2010)
4. Ministry of Land, Infrastructure, Transport and Tourism, Movement facilitation improvement guideline Regarding Passenger Facilities of Public Transportation (2013). (in Japanese)
5. Salthouse, T.A., Siedlecki, K.L.: Efficiency of route selection as a function of adult age. *Brain Cong.* **63**, 279–286 (2007)
6. Sanders, C., Schmitter-Edgecombe, M.: Identifying the nature of impairment in planning ability with normal aging. *J. Clin. Exp. Neuropsychol.* **34**, 724–737 (2012)
7. Kléncklen, G., et al.: What do we know about aging and spatial cognition? Reviews and perspectives. *Ageing Res. Rev.* **11**, 125–135 (2012)
8. Shinar, D., Schieber, F.: Visual requirements for safety and mobility of older drivers. *Hum. Fact.: J. Hum. Fact. Ergon. Soc.* **33**, 507–519 (1991)
9. Ohmi, G.: Model Psychology, Fukumura Shuppan Inc. (1984)

10. Salthouse, T.A., et al.: Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *J. Exp. Psychol.: Gener.* **132**, 566–594 (2003)
11. Utsumi, S., Takahashi, S., Ohori, K., Anai, H.: Agent-based analysis for design of signage systems in large-scale facilities. In: *Proceedings of the 2015 Winter Simulation Conference* (2015)
12. Ohori, K., Yamane, S., Anai, H., Utsumi, S., Takahashi, S.: An agent-based analysis of the effectiveness of signage system in a large-scale facility. In: *Social Simulation Conference*, vol. 32 (2016)
13. Fodness, D., Murray, B.: Passengers' expectations of airport service quality. *J. Serv. Mark.* **21**(7), 492–506 (2007)
14. Ushiro, M., Ueshima, K.: A basic study on evaluation for passenger usability in the airport terminals. *Technical Note of NILIM*, No. 313 (2006)
15. Narita International Airport Official Website: Next Generation Interactive Signage infotouch™ Now in Terminal 1! <https://www.naa.jp/jp/20171012-infotouch.pdf>. Accessed 7 Jan 2018
16. NTT DATA: “High precision indoor digital map system” to Narita International Airport. http://www.nttdata.com/jp/ja/news/services_info/2017/2017101201.html. Accessed 7 Jan 2018
17. NC State University: The Center for Universal Design. https://projects.ncsu.edu/ncsu/design/cud/about_ud/about_ud.htm. Accessed 7 Jan 2018



Developing an Input-Output Table Generation Algorithm Using a Japanese Trade Database: Dealing with Ambiguous Export and Import Information

Takaya Ohsato^(✉), Kaya Akagi, and Hiroshi Deguchi

Tokyo Institute of Technology, Tokyo, Japan
Osato.t.aa@m.titech.ac.jp

Abstract. In this study, we undertook the task of data conversion of enterprise export/import ratio using text mining. In addition, an input-output table was generated based on corporate data and information by using an algorithm. We also address the lack of input-output tables pertaining to individual companies in the reports released by various government ministries and agencies. Moreover, sales information and cost information of individual companies include the sales obtained through overseas exports and costs due to overseas imports. Therefore, in order to accurately calculate what is produced in one country, it is necessary to distinguish and separate the sales revenues and costs incurred from overseas trade. In this work, we structured the data. Text mining was applied to the information described by free descriptions of import/export ratios of direct trade companies to sales. Then, we simulate using structured data and propose an algorithm to construct an input-output table that includes foreign transactions.

Keywords: Input-output tables · Corporate data · SNA · Business transactions
Text mining · Overseas exports/imports

1 Introduction

Input-output Table is an economic statistic that systematically describes the input structure, calculation structure, and interdependency relationships between production departments produced in one country. This technique was developed by Leonchev [1]. Presently, it is at the core of System of National Accounts (SNA), which is an accounting system that captures the circulation in the economy of a country. In addition, in Tokyo, the input-output table is being used to calculate the net economic effect due to the Tokyo Olympic Games in 2020; it is estimated the event will produce about 14 trillion yen in the Tokyo region. However, the input-output tables used by the national and prefectural governments, which are announced by their respective ministries and agencies, are calculated for a period of five years. In addition, it takes about three years from the survey implementation to the launch, it is not suitable for analysis in real time.

Therefore, studies were conducted to solve the problem; the endogenous department of the input-output table was constructed using corporate transaction information and financial information from Teikoku Databank, Ltd. (TDB). In previous studies, input-output tables were constructed for 600,000 enterprises, taking account of business establishments. However, while the sales and cost of goods sold used for estimating the transaction amount include exports and imports abroad, the information in TDB only comprises data from domestic companies. Therefore, Overseas exports should be excluded from sales, while the cost of sales should exclude overseas imports. This is because the transaction information may include transactions of overseas exports and overseas imports.

The Size of the Small and Medium Business Administration Agency publishes the Input-output Table, which contains export/import ratios. However, the ratios are calculated for different industries and do not reveal the ratios for individual companies. Moreover, in the trade statistics of the Ministry of Finance, the proportion of individual companies has not been reported. On the other hand, TDB publishes the items of export/import implementation belonging to companies directly exporting or importing to overseas companies. However, the export/import ratio does not exist as an item. There are cases where descriptions on exports and imports of sales are made in free texts such as “business contents of companies” and “characteristics of companies”.

Therefore, in this work, we construct export/import ratios by applying textual mining on “business contents” and “characteristics of company” of companies conducting direct trade in TDB. The data export/import ratios are extracted from the information of individual companies; we attempt to model and the obtained ratios.

We also construct an input-output table that takes into account the exports and imports and formulates a calculation of endogenous departments confined to Japan.

1.1 Previous Research

In previous studies [2, 3], input-output tables were constructed using private data.

In [2], an algorithm for constructing an input-output table is developed that utilizes transaction data between TDB companies; the algorithm takes into considerations the mathematical implications due to the theory of exchange algebra. The supplier information in TDB includes items generated by each purchase and sale in the form of free descriptions. The items are classified according to industries in Japan based on “2 Side Classification Algorithm”, which allocates items described in the free descriptions by using the maximum likelihood estimation technique. In this paper, we describe a novel transaction amount proportioning algorithm that estimates transaction costs in companies in order to balance the cost of sales of ordered enterprises and sales of ordered companies on intercompany trading networks.

In [3], the cost of sales ratio by industry is incorporated as a calculation target for companies whose cost of sales is not known; this was done to improve the comprehensiveness of companies to be calculated. In addition, a construction algorithm is proposed with enhanced coverage.

In addition, in [3], calculations are based on individual head office units. Each unit is a summary unit of TDB from the business base based on the business place, information on business establishments owned by TDB, and the number of employees in a 500 m × 500 m mesh, which is published by government agencies. Moreover, the number of employees at each establishment is estimated by using the information of the establishment.

In this research, we will further develop these algorithms; in addition, we will redefine the transaction height estimation algorithm by considering export/import data.

2 Data Mining of Export/Import Ratio by Text Mining

2.1 Usage Data

TDB publishes a report on the credibility of companies called Corporate Credit Report (CCR) based on customer requests and field surveys. In this paper, overseas enterprises are selected and their ratios are extracted using “supplier column” and “current status and outlook column” the described in CCR.

First, for the selection of overseas enterprises, we identify companies that are engaged in direct trade. To that end, we use the items that describe export/import transaction, which also specify if the trade is direct or indirect; these items are provided in the supplier column of CCR. In this paper, the credit report conducted from 2011 to 2017 is used, in which comprises the latest surveys and statements of 13,236 companies; the items of “Exported (direct trade)” or “Imported (direct trade)” are considered.

As regards the information required for extracting ratios, there are cases where the export/import ratio is described in the business contents of the “current conditions” and the outlook column.

It can be seen that “current status and outlook” is stated in free form; moreover, there are not ratio items, and therefore it is necessary to extract the ratios from the free descriptions.

2.2 Overseas Ratio Extraction Algorithm and Extraction Results

To extract overseas ratios, 1,000 visits were conducted, and the features of the following description rules were confirmed.

[Rule gained by eye observation]

- If there is a ratio, any of the following terms, “overseas”, “import”, “country” or “direct trade”, must be listed
- In majority of the cases, ‘%’ is used as the notation; therefore, ‘Discount’ description is converted to ‘%’
- “Most” and “almost” usually refer to 90% of the business context; therefore, “Most” and “almost” are replaced by “90%”
- “Rare” and “little” often refer to “2%” in business context; therefore, “Rare” and “little” are replaced by “2%”.

Next, morphological analysis was performed using MeCab, wherein frequently occurring words are investigated; the following rule and trends were obtained.

[Rule Obtained by Morphological Analysis]

- From frequently occurring words, set overseas keywords, export keywords, and import keywords.
 - Overseas keywords: overseas, foreign, direct trade
 - Export keyword: export, sales, sales
 - Import keywords: export, purchase, procurement, ingredients
- When foreign keywords and export keywords are included, they are classified as “export group”, and when foreign keywords and import keywords are included, they are classified as “import group”.
 - ※ If all sentences are included in the sentence, discriminate to both.

[Trends Obtained by Morphological Analysis]

- There are many cases where the export/import ratio is stated before and after overseas keywords. Also, it is often described later.
- When four or more postpositions are present between overseas keywords and ratios; in several cases, ratios other than export/import ratios are listed.
- When there are two or more overseas keywords, the one with the shortest number of characters as the ratio is often the export/import ratio.

From the above, the following algorithm was developed for extracting overseas ratio.

[Extraction Algorithm]

- (1) “Overseas”, “Import” or “Country” is included
- (2) “Discount” description converted to “%”
- (3) “Most” and “almost” are often about 90% of the project content, so replace with “90%”
- (4) “rare” “little” is often “2%” in the business content, so replace with “2%”
- (5) When export keyword is included, it is discriminated as “export group”, when it contains import keyword, it is determined as “import group”
- (6) Extract keywords up to a certain ratio after overseas keyword + export/import
- (7) Exclude cases where there are 4 or more particles from the keyword to the ratio
- (8) When multiple keywords are included in the same document, use keywords with fewer characters as the ratio.

Furthermore. (6) If deleted under the following conditions, also do the following.

- (9) Extract keywords up to overseas keywords + export/import
- (10) Exclude cases where there are 4 or more particles from ratio to keyword
- (11) When multiple keywords are included in the same document, use keywords with fewer characters as the ratio.

[Extraction result]

- Among 7,880 export companies, 5,841 companies extracted (%)
- Of the 9,883 import companies, 6,139 companies extracted (%)

Import and export ratios of more than half of companies were extracted.

2.3 Accuracy Verification

In the initial survey, 1,000 visual observations were made. Then, the export/import ratios obtained by visual inspection were compared against the ratios extracted by the algorithm and the true positive rate was determined.

In the export, 154 cases were positive, in the range of $\pm 5\%$, compared to 198 correct answers obtained by visual inspection; the hit rate was 77.8%.

For imports, 53 cases were positive, in the range of $\pm 5\%$, compared to 74 correct answers data visually obtained; the hit rate was 71.6%.

It was found that less than 80% of the extracted ratios was close to the correct value; therefore, in this paper, the extraction ratio obtained here will be used.

3 Estimation of Export/Import Ratio

Given that there were companies whose ratios could not be extracted by the proposed algorithm, we investigate the variation in export/import ratios due to factors such as industry type and size and attempt to model this variation.

3.1 Ratio Trends and Modeling

First, we investigate the trends in the extracted export and import ratios. Table 1 shows the distribution in the companies by industry, wherein the export/import ratio is determined by the algorithm. In Fig. 1, export/import ratios and their frequency of occurrence are shown; the horizontal axis represents the export/import ratios and the vertical axis indicates the appearance density.

Table 1. Industry Oita classified export/import ratio found number

Industry classification	export		Import	
	number	Proportion	number	Proportion
Manufacturing industry	2350	40.2%	710	11.6%
Wholesale trade	3311	56.7%	5143	83.8%
Other	180	3.1%	286	4.7%

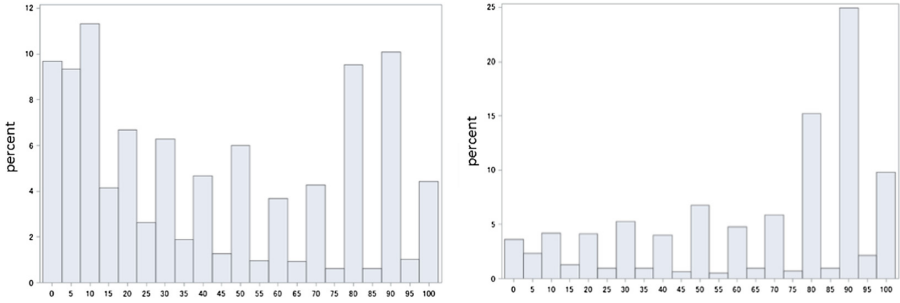


Fig. 1. Histogram of ratio (left: export, right: import)

Given that several ratios are described in terms of the numerical value of “X0%”, the distribution contains mountain valleys.

There are several companies in the manufacturing and wholesaling industries that are involved in direct export, this trend is also visible in the extraction results. As regards direct import, wholesale trade accounts for the majority share. Moreover, when comparing imports and exports, such as in Fig. 1, several companies with high import ratio are present. Overall, we find that the trends of exports and imports are clearly distinguishable.

Next, we test the following hypothesis.

[Hypothesis and verification method]

- Is the distribution different for different industry types?
 - ⇒ confirmed separately by “manufacturing industry”, “wholesale business”, and “other” (Fig. 2).
- Different distribution depending on scale.

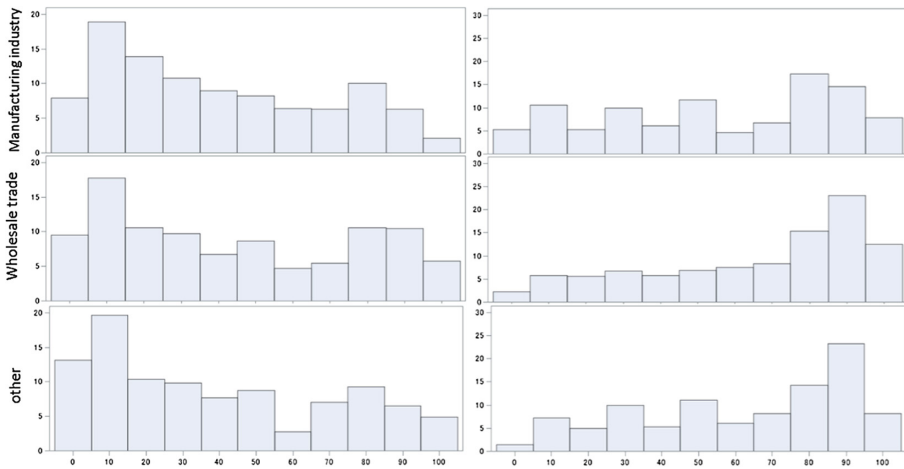


Fig. 2. Industry Oita classification extraction ratio (left: export, right: import)

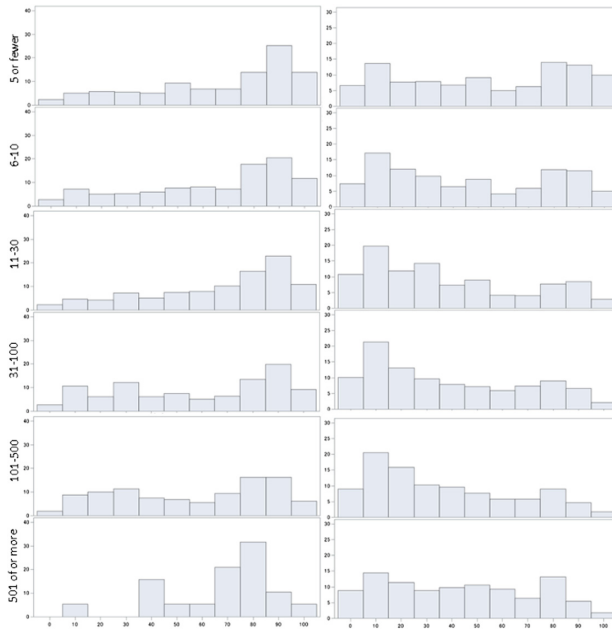


Fig. 3. Probability density by employee size (left: export, right: import)

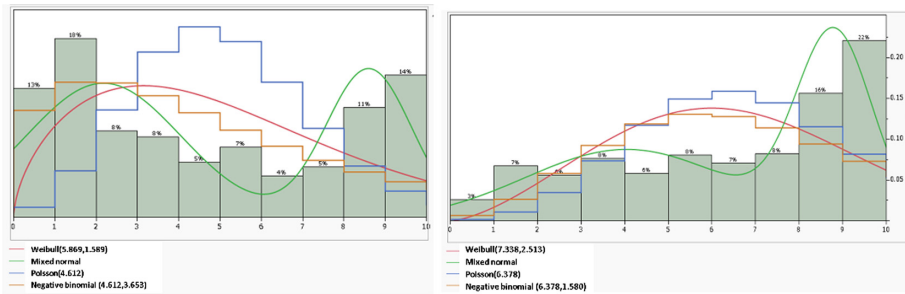


Fig. 4. Fitting of multiple distributions (left: export, right: import)

⇒ Confirmed by employee size. The classification of the scale is “5 or fewer”, “10 people”, “30 people”, “100 people”, “500 people”, “500 people or more” conducted (Fig. 3).

- Is it possible to replace with a parametric distribution for estimation modeling?
 ⇒ Confirmed by continuous distribution and discrete distribution

[inspection result]

- Is the distribution different depending on the industry?
⇒ There is no big difference
- Different distribution depending on scale.
⇒ There is no big difference
- Is it possible to replace with a parametric distribution for estimation modeling?
⇒ It does not apply to a mixed distribution, which is semi-nonparametric (Fig. 4).

Because modeling is difficult, we calculate the probability density according to each industry and employee scale and estimate each export/import ratio by Monte Carlo simulation.

3.2 Estimation by Simulation

For companies that were involved in direct exports and direct imports in 2011, simulation was carried out using the probability density divided by industry and employee size. By multiplying the obtained probability and the sales figures, we obtained total the export/import figures (Fig. 5).

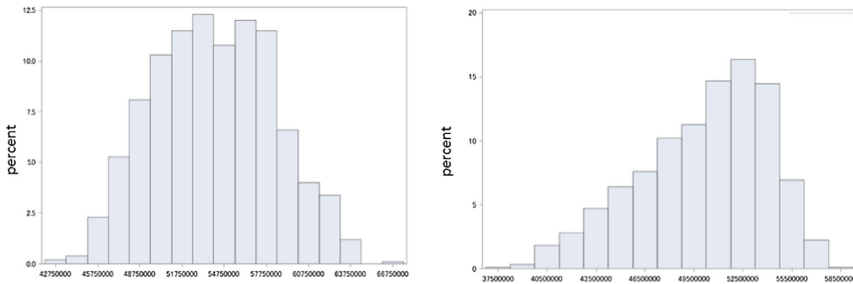


Fig. 5. Calculation result of export/import amount by simulation (left: export, right: import)

These values were then compared with the trade statistics published by the Ministry of Finance; it was found that the values calculated using simulation were almost in agreement, as shown in the following table (Table 2).

The total import and export value obtained by simulation is about 1,100 trillion yen, which is recorded as the total sales by TDB. On the other hand, the economic census provides a figure that is close to 1,300 trillion yen; therefore, this discrepancy can be ascribed to the bias of tabulation. Accordingly, we adjust for the 10% increase in total sales.

Table 2. Comparison with published figures

2011 year	Ministry of Finance	Estimated value (median, sales bias corrected)	Difference
Export	65.5 trillion yen	65.1 trillion yen	-0.4 trillion yen
Import	68.1 trillion yen	61.2 trillion yen	-6.9 trillion yen

4 Proposal of a Transaction Estimation Algorithm Considering Export/Import

In this paper, we adopt the exchanging algebra form (Deguchi [5]) and the AADL (Algebraic Accounting Description Language), which is specialized for its calculation as in previous research.

In general, exchange algebra can be represented as

$$\text{value}\langle \text{name}; \text{unit}; \text{time}; \text{subject} \rangle$$

It expresses in what form (name), what unit (unit), when (time), who (subject), how much (value) processed in the form. Each base (name, unit, time, subject) has a base set, and the base of exchange algebra consists of each element. Such an algebraic expression makes it feasible to maintain mathematically guaranteed robustness in information processing.

As regards the generation of the input-output table in this paper, we extend the basis of exchange algebra and deal with bookkeeping elements (such as sales and cost of sales) and treat data groups related to multiple subjects (such as transactions between companies) as the basis. In order to process such bases consistently, it is important to extract and process retroactively from the composed overall input-output table for each base (for example, companies, products, and industry classification).

In order to formulate the algorithm, the following definition of the variable is used.

$$\text{Ordered firm} : R = \{r_i | i = 1, 2, \dots, n_1\},$$

$$\text{Ordered company domestic sales amount} : SD = \{sd_i | i = 1, 2, \dots, n_1\},$$

$$\text{Order company Overseas export value} : SF = \{sf_i | i = 1, 2, \dots, n_1\},$$

$$\begin{aligned} \text{Ordered company sales} : S &= \{s_i | i = 1, 2, \dots, n_1\} \\ &= \{sd_i + sf_i | i = 1, 2, \dots, n_1\} \end{aligned}$$

$$\text{Ordering company} : O = \{o_j | l = 1, 2, \dots, n_2\},$$

$$\text{Order Company Domestic Cost Amount} : OD = \{od_j | j = 1, 2, \dots, n_2\},$$

Order company Overseas import amount : $OF = \{of_j | j = 1, 2, \dots, n_2\}$,

$$\begin{aligned} \text{Order company Cost of sales : } C &= \{c_j | l = 1, 2, \dots, n_2\} \\ &= \{cd_j + cf_j | l = 1, 2, \dots, n_2\} \end{aligned}$$

$$\text{Deal set : } E = \{e_{ij} | <r_i, o_j > | r_i \in R, o_j \in O\}$$

Since the endogenous department of the input-output table is the total domestic transaction volume, the transaction height estimate in the inter-company network is calculated using domestic transactions SD and CD, as follows.

$$\text{Proportional division cost of sales : } wc_{ij} = cd_j \frac{sd_i}{\sum_{i \in e_j} sd_i}$$

$$\text{Estimated sales proportional distribution transactions : } ws_{ij} = sd_i \frac{wc_{ij}}{\sum_{j \in e_i} wc_j}$$

However, since the sum of the transaction volumes exceeds the cost of goods sold by the ordering company, which is not established accurately in account terms, the following adjustment is added.

Cost adjusted transaction volume estimation:

$$\begin{aligned} ws_{ij} &= ws_{ij} \times \frac{cd_j}{\sum_i ws_{ij}}, & \sum_i ws_{ij} &> cd_j. \\ &ws_{ij}, & \sum_i ws_{ij} &\leq cd_j. \end{aligned} \tag{1}$$

4.1 Accuracy of Transaction Amount

In this subsection, we describe the verification process of the accuracy of the transaction amount obtained by Eq. (1).

In TDB, since the transaction value of about 65,000 transactions, which represents 1% of all transactions, is known, it is compared with the transaction estimate calculated by this algorithm. Figure 6 shows a box-and-whisker plot of the logarithmic transformation of the transaction amount with 1% up and down. The line indicates a value for which the estimated value and the actual value are identical.

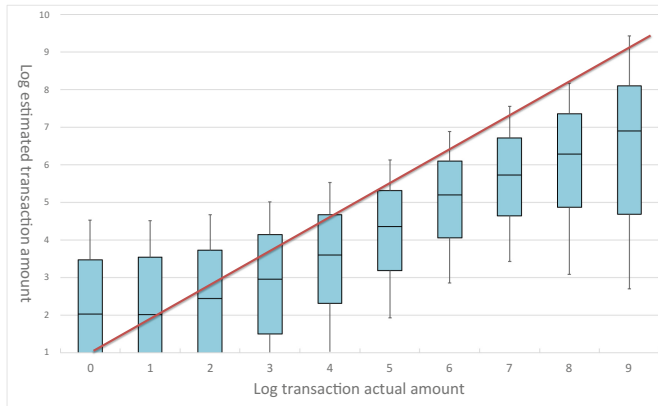


Fig. 6. Comparison of transaction actual value and transaction estimate

In Fig. 6, the log transaction value of 7 or more deviates from the top 25% of the actual amount. However, when the actual amount is 7 or more after the logarithmic transformation, the number of transactions is about 6,000, which is approximately 8% of the total transactions.

When the post-log transformation transaction amount, which accounts for 92% of the transaction amount, is 6 or less, it corresponds to a box of a box-whisker plot; the transaction amount is considered to be reasonable if this condition is satisfied.

Researchers including Tamura [7] have conducted studies to estimate intercompany transactions using private data. In this research, we do not use private data because the calculation conditions for constructing the input-output table, such as directly estimating overseas import/export and not exceeding the sales of enterprise, are different. However, since the transaction value estimate, as obtained by Tamura [7], is the transaction amount that is also utilized by the Regional Economy and Society Analyzing System (RESAS), the comparative verification will be a subject for future work.

4.2 Construction of Algorithm Considering Production Base

Kikukawa [6] estimated optimal inter-branch transactions using corporate office location information. In Ohsato [4], the number of employees at the establishment were estimated and the estimated transaction amount was distributed proportionally among business-to-business transactions at the head offices of the company. In this work, we formalize the process to apportion transaction height estimates by considering overseas import/export transactions.

$$\text{Production base of Receive company : } RB = \{rb_{ij} | i = 1, 2, \dots, n_1, j = 1, 2, \dots, r_i\}$$

$$\text{Production base of Order company : } OB = \{ob_{lk} | l = 1, 2, \dots, n_2, k = 1, 2, \dots, o_k\}$$

Main trade : $M_{il} = \{m_{il} | m_{il} \in <0, 1>, 0 = \text{not main trade}, 1 = \text{main trade}\}$

Trading set : $Ex = \{<r_i, o_i, m_{il}> | r_i \in R, o_i \in O, m_{il} \in M\}$

The distance between companies : $Dis = \{d_{ijk}\}$

The shortest distance : $SD = \{\text{Min}(d_{ijk})\}$
 $= \{sd_{ijk}\}$

Then, the algorithm can be defined, as described by Kikukawa [6], as follows,

$$\begin{aligned} \text{Best trading : } BT &= md_{i1lk} && , r_i = 1 \\ &md_{ijkl} && , r_i \geq 2, o_k = 1, m_{il} = 0 \\ &md_{i1lk}, md_{i2lk}, \dots, md_{isilk}, (r_i \geq 2, m_{il} = 1) \cup (r_i \geq 2, o_k \geq 2) \end{aligned}$$

There is a strong correlation between the sales of a company and the number of employees. A company with several employees generates significant sales volume. According to TDB company outline data, the correlation coefficient for sales and employees is approximately 0.65. Therefore, we conjecture that a production base with a large number of employees makes profit; in addition, we assume that the proportional division is appropriate for the number of production base employees.

Number of Receiver business establishment Employees:

$$RE = \{re_{lk} | l = 1, 2, \dots, n_2, k = 1, 2, \dots, o_k\}$$

Trading Amount with multiple receiver business establishments:

$$\begin{aligned} TM &= tm_{ijkl} \\ &= ws_{ij} \times re_{lk} / \sum_k^{1,2,\dots,o_k} re_{lk} \end{aligned} \tag{2}$$

The amount to be distributed between regions in Japan’s statistics was estimated using information obtained from hearings and mail surveys to large-scale business establishments. Such surveys are, however, expensive and time consuming. Consequently, transaction between regions as estimated from business establishment transactions are useful and indispensable to the construction of interregional input-output tables.

By proportionally allocating estimated transaction amount to business establishments, it is possible to consider production sites that are already present in the area. Therefore, the transaction amount between areas can be calculated by adding the transaction amount (1) between business sites at business establishments/companies belonging to that area.

5 Conclusion

In this research, the ratios of overseas exports/imports to sales in individual companies were calculated. These were determined using data, the credit company survey report of TDB, using text mining, and by structuring the export/import ratio data. We developed an algorithm for constructing an input-output table by considering an overseas scale, which was indistinguishable in previous studies.

Initially, it was hypothesized that distribution of proportion may differ across industry types and scale; however, after the structuring of data, no significant change was observed in the distributions across industry types and scale. Moreover, it was found that the overseas ratio of imports was higher than overseas ratio for exports in Japan. Majority of the importing companies were wholesale businesses; however, even in comparison with wholesale export businesses, the number of companies importing was higher; however, it cannot be ascribed to the bias due to industry. We conjecture that many importing companies occupy most of their sales by imports.

Moreover, it was found that by simulation, it is possible to calculate the approximate export value/import values. However, this will require further attention, as it is based on the premise that there is no observable trend in the ratios for each year. This premise requires verification, and it is necessary to consider the import/export ratios in a state divided in time series in the future.

In addition, the simulation shows that the import amount is slightly reduced. This may be the case for this wholesale industry, this ratio will be lower. For example, in major wholesale businesses such as ITOCHU Corporation and Marubeni Corporation, it is one of the improvement measures to individually import import/export ratios from information such data.

Once the import/export ratios for individual companies are calculated, it is possible to calculate the magnitude of ripple effects caused by changes in overseas affairs such as foreign exchange. This suggested that Japan is not making money with trade, but it may be a trigger to verify this.

Although this study only proposes the construction of the algorithm, the calculation and verification of actual input-output tables using actual data will be considered in future work.

References

1. Leontief, W.: Environmental repercussions and the economic structure: an input-output approach. *Rev. Econ. Stat.* **52**(3), 262–271 (1970)
2. Akagi, K., Ohsato, T., Deguchi, H.: Input-output table constructed with private business data and its algebraic description. In: 2015 IEEE/SICE International Symposium on System Integration (SII). IEEE (2015)
3. Ohsato, T., Akagi, K., Deguchi, H.: To improve comprehensiveness of input-output table with private business data and considerations of changes in industrial linkage structure. In: SICE Society System Sectional Meeting (2016)

4. Ohsato, T., Akagi, K., Deguchi, H.: Input-output table constructed with private business establishment on company information data. In: JSAI International Symposia on AI 2016, Artificial Intelligence of and for Business Proceedings (2016)
5. Deguchi, H.: Economics as an Agent-Based Complex System. Springer, Tokyo (2004). <https://doi.org/10.1007/978-4-431-53957-5>
6. Kikukawa, Y., Tsutsumi, M.: Estimate dealings between the business establishment by private business data and utilization plan. In: Japan Society of Civil Engineers Conference 2015 Spring (2015)
7. Tamura, K., et al.: Estimation of flux between interacting nodes on huge inter-firm networks. *Int. J. Mod. Phys. Conf. Ser.* **16**, 93–104 (2012)



Stock Price Prediction with Fluctuation Patterns Using Indexing Dynamic Time Warping and k^* -Nearest Neighbors

Kei Nakagawa^{1,2(✉)}, Mitsuyoshi Imamura^{1,3}, and Kenichi Yoshida²

¹ Nomura Asset Management Ltd., Tokyo, Japan
kei.nak.0315@gmail.com

² Graduate School of Business Sciences, University of Tsukuba, Tsukuba, Japan

³ Department of Risk Engineering, University of Tsukuba, Tsukuba, Japan

Abstract. Various methods to predict stock prices have been studied. A typical method is based on time-series analysis; other methods are based on machine-learning techniques using cross-sectional data as feature values. In the field of empirical finance, feature values for prediction include “momentum”. The momentum strategy is simply based on past prices. Following the nearest trend, we buy current performers. From the different viewpoint from momentum, We’d like to challenge EMH. Our proposed method is following the similar trend. In other word, we look for past pattern similar to the current and predict from that. When predicting stock prices, investors sometimes refer to past markets that are similar to the current market. In this research, we propose a method to predict future stock prices with the past fluctuations similar to the current. As the levels of stock prices differ depending on the measured period, we develop a scaling method to compensate for the difference of price levels and the proposed new method; specifically, we propose indexing dynamic time warping (IDTW) to evaluate the similarities between time-series data. We apply the k^* -nearest neighbor algorithm with IDTW to predict stock prices for major stock indices and to assist users in making informed investment decisions. To demonstrate the advantages of the proposed method, we analyze its performance using major world indices. Experimental results show that the proposed method is more effective for predicting monthly stock price changes than other methods proposed by previous studies (Based on the comments received in Ai-Biz 2017, we clarified the differences from previous studies. And we added economic discussions about our proposed method such as differences from “momentum”, a challenge to Efficient Market Hypothesis and meanings as investment behavior).

Keywords: Dynamic time warping · Indexing dynamic time warping
 k^* -nearest neighbors · Weighted k -nearest neighbors

1 Introduction

Are the past stock price movements effective for price prediction? It's a frequently asked question in financial study. And there are many researches of both positive and negative. The representative study of negative opinion is Efficient Market Hypothesis (EMH) by [10]. According to EMH, future prices cannot be predicted by analyzing the past prices. On the other hand, the representative study of positive opinion is the momentum effect identified by [16]. The momentum strategy is simply based on past prices. Following the nearest trend, we buy current performers. From the different viewpoint from momentum, We'd like to challenge EMH. Our proposed method is following the similar trend. In other word, we look for past pattern similar to the current and predict from that. this method is reasonable as practitioners. When predicting stock prices, investors sometimes refer to past markets that are similar to the current market. For example, if the current market is bullish, investors refer to similar past bull markets. Investors then try to predict stock price changes based on movements that occurred in the past. In this research, we propose a method to predict future stock prices with the past fluctuations similar to the current.

First, we extract the past period in which the stock price fluctuation is most similar to the current fluctuation. To measure similarity between time-series data, we employ the dynamic time warping (DTW) method [15], which is used mainly in the field of speech recognition. As the stock price levels differ depending on the measured period, we also develop a scaling method to compensate for the difference of price levels and the proposed new method; specifically, we propose indexing DTW (IDTW) to evaluate the similarities between time-series data. In this method, the daily stock price fluctuation patterns are analyzed every month. Here, the fluctuation patterns are expressed as a ratio with the previous day.

Next, we forecast future stock prices based on similar price fluctuations extracted by IDTW. We use the k^* -nearest neighbors (k^* -NN) method [1] to select the k most similar fluctuations extracted by IDTW. Knn is the most natural algorithm to predict. The k^* -NN algorithm is an improvement of the k -NN algorithm, which is a non-parametric method for pattern recognition and machine learning. In our setting, the stock price fluctuation patterns are given as data points. The next month returns of the data points are given as labels. Given a new data point, we want to predict the label of the new data point. And the distances between data points are measured by IDTW. Then, knn uses the weighted average of the labels closest to the data points. The question of how to set the optimal number of neighbors k , as well as the optimal weights, has received much attention over the years. In fact, this remains an open research problem. Anava and Levy [1] proposed a simple approach to calculate the optimal weights and found the optimal number of neighbors for each data point whose value must be estimated.

To demonstrate the advantages of the proposed method, we analyze its performance using major stock indices. Experimental results show that the proposed method is more effective for predicting monthly stock price changes than other

methods proposed by previous studies. IDTW outperforms conventional DTW and its successor, derivative DTW (DDTW) [17], from the viewpoints of both accuracy and profitability. Furthermore, we show that k^* -NN outperforms conventional k -NN.

The remainder of this article is structured as follows. Section 2 reviews relevant literature, while Sect. 3 describes our research methods. Section 4 details experiments on stock markets, and finally, our conclusion is given in Sect. 5.

2 Related Works

Various methods to predict stock prices have been studied to challenge EMH. In a literature review of financial time-series forecasting, two types of methods can be observed: time-series and machine-learning methods. The first method is based on time-series analysis [13]. Time-series analysis emphasizes statistical approaches like Auto-Regressive Integrated Moving Average (ARIMA) [7] and Generalized Auto-Regressive Conditional Heteroscedasticity (GARCH) [5] models. The ARIMA model is used to predict stock prices or returns. The GARCH model is used to predict the standard deviation, also known as volatility, of stock prices. Past studies show that the conditional variance model adequately captures “volatility clustering”¹, i.e., the stock price shock is sustainable. Time-series analysis is a method with low arbitrariness in the sense that it models data with linear equations of past data.

Other methods used to predict stock prices are based on machine-learning techniques using cross-sectional data as feature values. Typical features for prediction in the field of empirical finance include “value” and “momentum” [3, 11, 12, 16]. Unlike time-series analysis, prediction by machine learning sufficiently selects features.

Recently, many researchers have attempted to develop computational methods to support decision-making in different financial markets. Cavalcante et al. [8] presented a review of the application of several computational methods in financial applications. In their survey and to the best of their knowledge, research that uses DTW or k -NN with DTW as the method to classify or predict financial time-series does not exist. Coelho [9] analyzed fluctuation patterns using return, volatility, and volume with DTW and k -NN. However, their method has room for improvement. For example, methods to tune various parameters must be developed. The results of Coelho [9] are not stable compared with the results reported in this paper.

From practical view points, our proposed method has similarity with so-called technical analysis. Many traders use them to try to predict future price trends by studying past price movements and charts. Traditionally, the approach to technical analysis is a manual one. However, it can be said that our proposed method has improved technical analysis so that it can systematically extract price patterns.

¹ Volatility clustering is a phenomenon by which a period of high (low) volatility continues for some time after the volatility rises (decreases) [19].

A preliminary version of this paper was presented at the 31st Annual Conference of the Japanese Society for Artificial Intelligence [14, 20]. Although [14, 20] use k -NN for prediction, this paper refines this method with k^* -NN.

3 Proposed Method: k^* -Nearest Neighbors with Indexing Dynamic Time Warping

In our method, we first refer to past markets that are similar to the current market. To measure the similarity between stock price fluctuations, we employ IDTW. Then, we predict stock price changes based on movements that occurred in the past. Intuitively, this procedure can be represented using k^* -NN.

In this section, we first explain conventional DTW and DDTW. Then, we propose IDTW, which applies DTW on scaled stock price fluctuations. In addition, we describe the k^* -NN algorithm, which is an improvement of the k -NN algorithm.

3.1 Dynamic Time Warping

DTW is a method used to evaluate similarity between time-series data. Many methods have been proposed to measure similarities between time-series data. Correlation coefficients and the Euclidean distance are often used as simple similarity measures between sequence data. However, they are not sufficient to evaluate similarities between financial time-series data. Correlation coefficients only capture the linear relationship between time-series data, and the Euclidean distance is very sensitive to small distortions in the time axis. Furthermore, it cannot be used when the lengths of two time-series are different.

The DTW distance is a measure that can be used to overcome these problems. It makes optimal correspondence with respect to two time-series by non-linearly expanding and contracting the time axis. This measure can be applied to a pair of time-series of different lengths. In addition, it fits human intuition. Figure 1 shows an example of correspondence of time-series data based on the Euclidean distance and DTW distance.

The DTW distance measures the similarity between time-series x and y using Algorithm 1. Here, $x[t]$ and $y[t]$ are the values of x and y , respectively, at time t ; N and M are the lengths of the time-series x and y , respectively, and d is a metric or distance function. In this paper, we use the absolute distance $d(x[i], y[j]) = |x[i] - y[j]|$.

A disadvantage of the DTW distance is that its calculation is of the order $O(MN)$. However, recently, high-speed computers have made this calculation more efficient.

3.2 Derivative Dynamic Time Warping

Keogh and Pazzani [17] proposed DDTW to solve the ‘‘singularities’’ problem. Keogh did not directly apply DTW to time-series data, but instead, applied

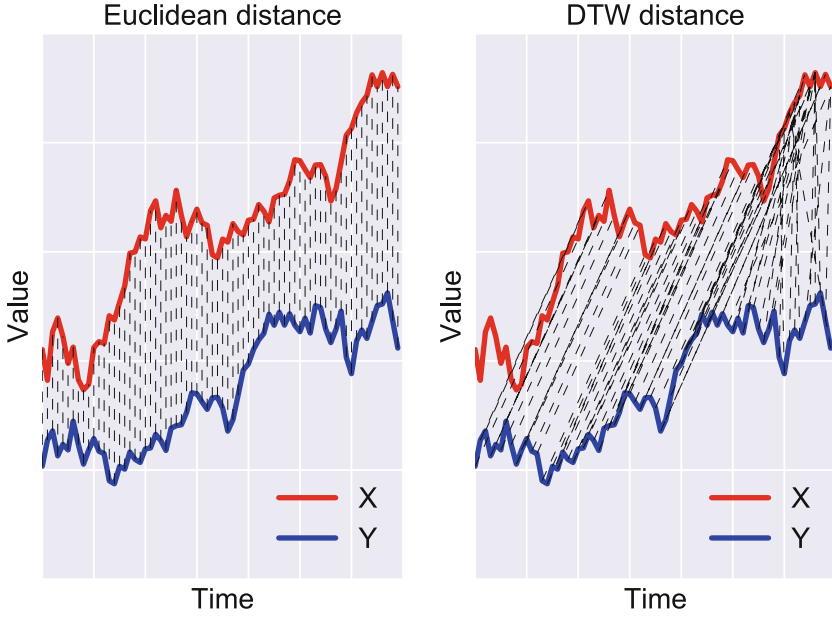


Fig. 1. Correspondence of time-series data.

Algorithm 1. DTW distance

```

1: procedure DTW( $x, y$ )
▷ Initialize matrix D
2:   Var  $D[N, M]$ 
3:    $D[1, 1] = 0$ 
4:   for  $i = 2$  to  $N$  do
5:     for  $j = 2$  to  $M$  do
6:        $D[i, j] = \infty$ 
7:     end for
8:   end for
▷ Calculate DTW distance
9:   for  $i = 2$  to  $N$  do
10:    for  $j = 2$  to  $M$  do
11:       $D[i, j] = d(x[i - 1], y[j - 1])$ 
 $+ \min(D[i, j - 1], D[i - 1, j], D[i - 1, j - 1])$ 
12:    end for
13:  end for
14:  return  $D[N, M]$ 
15: end procedure

```

DTW to the change of time-series data. Algorithm 2 outlines DDTW. Here, D_x and D_y are the differentials of the original time-series data.

Figure 2 shows the results of DDTW when it is applied to TOPIX data. Although DDTW has various advantages over DTW [17], it does not work well

Algorithm 2. DDTW distance

```

1: procedure DDTW( $x, y$ )
▷ Initialize vectors
2:   Var  $Dx, Dy$ 
3:   for  $i = 2$  to  $N - 1$  do
4:      $Dx[i] = \frac{(x[i] - x[i-1]) + ((x[i+1] - x[i-1])/2)}$ 
5:   end for
6:   for  $j = 2$  to  $M - 1$  do
7:      $Dy[j] = \frac{(y[j] - y[j-1]) + ((y[j+1] - y[j-1])/2)}$ 
8:   end for
▷ Apply DTW
9:   return  $DTW(Dx, Dy)$ 
10: end procedure

```

for financial time-series whose values rapidly change in a short period of time. In the next subsection, we propose an improvement of DTW for financial time-series data.

3.3 Indexing Dynamic Time Warping

As the level of a stock price differs depending on the measured period, we develop a scaling method to compensate for the difference of price levels.

Another important issue that must be considered is the measuring period. The seasonality of a stock price is widely recognized by investors. For example, “Sell in May” is a well-known type of seasonality. The seasonality of stock prices has been verified by [2, 6], and the “Sell in May” effect has been confirmed in global stock markets. In practice, the monthly return is the fundamental evaluation unit for fund and investment managers. In other words, investors are aware of stock price fluctuations by “month” periods. Thus, we calculate the DTW distance between monthly stock fluctuation patterns, each of which is composed of daily stock prices. Note that the number of days per month varies depending on the year and month. Thus, we cannot use measures that require the lengths of data to be the same such as the Euclidean distance and correlation coefficients.

We conjecture that investors pay little attention to the differential value of stock prices; the form of fluctuation is more important. The assumption behind the proposed method is that the pattern of stock price fluctuations is an important source of information for investors. In practice, when comparing two price fluctuations, it is natural to compare them by indexing rather than comparing with current series themselves or the rate of return. Similar patterns cause similar investor behaviors and result in similar changes in stock prices. Algorithm 3 of IDTW is designed to capture such pattern fluctuations in stock prices. Here, I_x and I_y are the scaled values of the original time-series data.

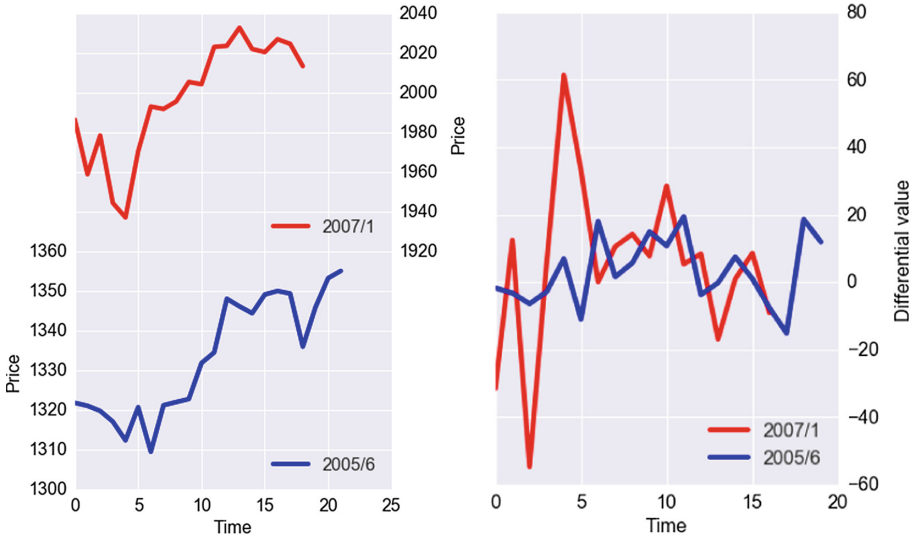


Fig. 2. Conversion of DDTW for TOPIX data.

Algorithm 3. IDTW distance

```

1: procedure IDTW( $x, y$ )
2:   Var  $I_x, I_y$  ▷ Scaling data
3:    $I_x[1] = 1, I_y[1] = 1$  ▷ Initialize  $I_x, I_y$ 
4:   for  $i = 2$  to  $N$  do
5:      $I_x[i] = I_x[i - 1] \frac{x[i]}{x[i-1]}$ 
6:   end for
7:   for  $j = 2$  to  $M$  do
8:      $I_y[j] = I_y[j - 1] \frac{y[j]}{y[j-1]}$ 
9:   end for
▷ Apply DTW
10:  return  $DTW(I_x, I_y)$ 
11: end procedure

```

Figure 3 shows an example of the IDTW results applied to TOPIX data. As above, the IDTW distance is calculated using the daily stock price within each month. Note that smaller distances indicate greater similarity between stock price fluctuations.

3.4 k^* -Nearest Neighbors

k -NN is a non-parametric method for pattern recognition and machine learning. It is considered to be a lazy learning method that does not build a model or function, but yields the closest k records to the training data set that have

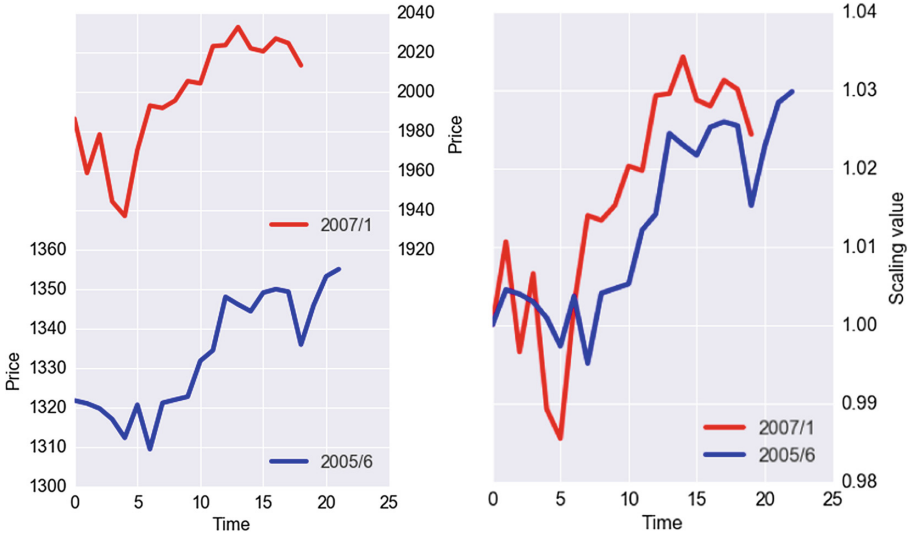


Fig. 3. Conversion of IDTW for TOPIX data.

the highest similarity to the test set. The (weighted) average of the k records is the predicted value of the k -NN. The question of how to set the optimal number of neighbors as well as the optimal weights has received much attention over the years. To the best of our knowledge, this remains an open research problem. Anava and Levy [1] proposed a simple approach for calculating the optimal weights and finding the optimal number of neighbors for each data point whose value must be estimated. In this section, we review the k^* -NN algorithm according to Anava and Levy

First, we are given n data points $x_1, \dots, x_n \in \mathbb{R}^d$ and n corresponding labels $y_1, \dots, y_n \in \mathbb{R}$. Assume that for any $i \in \{1, \dots, n\} = [n]$, it holds that $y_i = f(x_i) + \epsilon_i$, where $f(\cdot)$ and i satisfy the following:

- (1) $f(\cdot)$ is a Lipschitz continuous function: For any $x, y \in \mathbb{R}^d$, it holds that $|f(x) - f(y)| \leq Ld(x, y)$, where the distance function $d(\cdot, \cdot)$ is DTW, DDTW, or IDTW.
- (2) ϵ_i is a noise term: For any $i \in [n]$, it holds that $E[\epsilon_i | x_i] = 0$ and $|\epsilon_i| \leq b$ for given $b > 0$. In addition, it is assumed that for the given data points x_i , the noise terms ϵ_i are independent.

Given a new data point x_0 , our task is to estimate $f(x_0)$, where the estimator $\hat{f}(x_0)$ is restricted to be of the form $\hat{f}(x_0) = \sum_{i=1}^n \alpha_i y_i$; that is, the estimator is a weighted average of the given noisy labels. Formally, our goal is to minimize the absolute distance between our prediction and the ground truth $f(x_0)$, which translates into

$$\arg \min_{\alpha \in \Delta_n} \left| \sum_{i=1}^n \alpha_i y_i - f(x_0) \right| \quad (\text{P1})$$

where minimization is over the simplex $\Delta_n = \{\alpha \in \mathbb{R}^n \mid \alpha_i > 0 \text{ and } \sum_{i=1}^n \alpha_i = 1\}$. Decomposing the objective of (P1) into a sum of bias and variance terms, we arrive at the following relaxed objective:

$$\left| \sum_{i=1}^n \alpha_i y_i - f(x_0) \right| \leq \left| \sum_{i=1}^n \alpha_i \epsilon_i \right| + L \sum_{i=1}^n \alpha_i d(x_i, x_0).$$

By Hoeffding's inequality, it follows that $|\sum_{i=1}^n \alpha_i \epsilon_i| < C \|\alpha\|_2$ for $C = b\sqrt{2 \log(\frac{2}{\delta})}$ with probability at least $1 - \delta$. Thus, we arrive at a new optimization problem (P2) whose solution yields a guarantee for (P1) with high probability:

$$\arg \min_{\alpha \in \Delta_n} \left| C \|\alpha\|_2 + L \sum_{i=1}^n \alpha_i d(x_i, x_0) \right| \quad (\text{P2})$$

Without loss of generality, assume that the points are ordered in ascending order according to their distance from x_0 , i.e., $d(x_1, x_0) < d(x_2, x_0) < \dots < d(x_n, x_0)$. Also, let $\beta \in \mathbb{R}^n$ be such that $\beta_i = Ld(x_i, x_0)/C$. Then according to the Karush-Kuhn-Tucker (KKT) conditions, the optimal solution α^* for (P2) is of the following form:

$$\alpha_i^* = \frac{(\lambda - \beta_i) \times \mathbf{1}\{\beta_i < \lambda\}}{\sum_{i=0}^n (\lambda - \beta_i) \times \mathbf{1}\{\beta_i < \lambda\}}$$

Anava and Levy [1] proved that Algorithm 4 finds the exact solution of (P2). Notice that the optimal weights depend on a single parameter L/C , namely, the Lipschitz to noise ratio. As L/C increases, k^* decreases.

Ratio L/C is determined by n -fold cross validation (CV). CV [21] is one of the most widely used methods for assessing the generalizability of algorithms in classification and regression schemes. However, with respect to time-series predictions, the serial correlation of data, along with possible non-stationarities, render the CV be problematic as it does not account for such issues [4]. In addition, future data should not be used to predict past data. Therefore, to determine the k -NN and k^* -NN parameters, we instead use a usual out-of-sample evaluation.

4 Experiment on Stock Markets

To show the effectiveness of past stock price fluctuation patterns extracted by k^* -NN with IDTW for price predictions, we compared six methods combining DTW, DDTW, and IDTW with k -NN and k^* -NN. They are abbreviated by DTW+ k NN, DDTW+ k NN, IDTW+ k NN, DTW+ k^* NN, DDTW+ k^* NN, and IDTW+ k^* NN. Here, DTW+ k NN is examined in [9].

Algorithm 4. k^* -NN

```

1: procedure  $k^*$ -NN( $\beta, y, L/C$ )
▷ Initialize  $\lambda$  and  $k$ 
2:   Var  $\lambda[N]$ 
3:    $k = 0, \lambda[0] = 0$ 
4:    $\beta = L/C \times \beta$ 
5:   while  $\lambda[k] > \beta[k + 1]$  And  $k \leq N - 1$  do
6:      $k = k + 1$ 
7:      $\lambda[k] = \frac{1}{k}(\sum_{i=0}^k \beta[i] + \sqrt{k + (\sum_{i=0}^k \beta[i])^2 - k \sum_{i=0}^k \beta[i]^2})$ 
8:   end while
▷ Calculate weight parameter  $\alpha$ 
9:   Var  $\alpha[N]$ 
10:  for  $i = 1$  to  $N$  do
11:     $\alpha[i] = \frac{(\lambda[k] - \beta[i]) \times \mathbf{1}\{\beta[i] < \lambda[k]\}}{\sum_{i=0}^N (\lambda[k] - \beta[i]) \times \mathbf{1}\{\beta[i] < \lambda[k]\}}$ 
12:  end for
13:  return  $\sum_{i=0}^N \alpha[i] \times y[i]$ 
14: end procedure

```

We used the daily indices of TOPIX, S&P500, FTSE100, DAX30, and CAC40. These are the most commonly followed equity indices, and many consider them the best representations of the each country's stock market. Index data was acquired from Bloomberg terminal².

The data period of all indices was from January 1989 to August 2017. We used data from January 1989 to December 2005 for reference purposes. The test data were from January 2006 to August 2017. This is because we wanted to hold a test period over 10 years including the date of Lehman shock. But, we have to check the impact of reference period choice on performance for further study. In this experiment, several of the most similar stock price fluctuations in a month were searched from past data using the DTW, DDTW, and IDTW distances. The following month's returns were then predicted using k -NN or k^* -NN for each given distance, and monthly stock prices were repeatedly predicted. To determine the k -NN and k^* -NN parameters, we performed a 36 month out-of-sample evaluation. The most accurate parameter was chosen every month³. Our proposed method is as follows:

Step 1. Calculate the distance vector β with DTW, DDTW, and IDTW. β represents sorted distances between month t and all months from 1 to $t - 1$ of monthly price fluctuations (Fig. 4, left sub-figure).

Step 2. Choose n closest months, and predict the next month's return using k -NN or k^* -NN given y and β (Fig. 4, center sub-figure).

² The ticker codes are TPX Index, SPX Index, UKX Index, DAX Index, and CAC Index, respectively.

³ The parameter k ranged from 1 to 10 and parameter $L/C \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. These are the same setting as Anava and Levy [1].

- Step 3.** If the next month’s return is positive, buy one unit of TOPIX at the end of the month. If negative, sell one unit and calculate revenue for month $t + 1$ (Fig. 4, right sub-figure).
Step 4. Proceed to the next month: $t = t + 1$.

Table 1 lists the average accuracy for all years. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the six methods. Notice in Table 1 that IDTW+ k^* NN is the most accurate of all the methods.

Notice that for the total average, IDTW+ k NN outperforms DTW+ k NN and DDTW+ k NN in terms of accuracy. Likewise, IDTW+ k^* NN outperforms DTW+ k^* NN and DDTW+ k^* NN in terms of accuracy for the total average. These results show that IDTW is superior to DTW and DDTW in terms of accuracy. Likewise, k^* -NN outperforms k -NN; this result is confirmed with respect to accuracy.

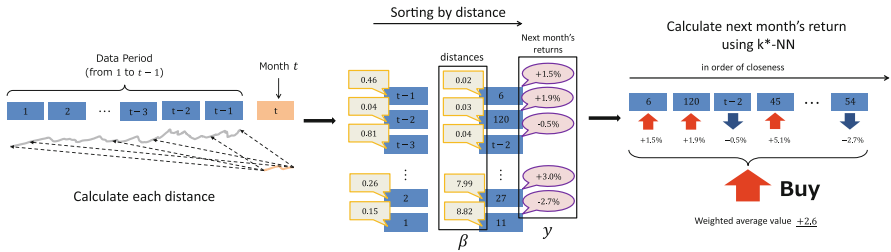


Fig. 4. Stock price prediction framework.

Table 2 shows the total returns of the six methods. Although the total return is proportional to the accuracy, the total return declines if predictions are missed when there are big fluctuations. However, IDTW+ k^* NN is the most profitable except for FTSE100. Again, IDTW is superior to DTW and DDTW, and k^* -NN is superior to k -NN in terms of profitability.

Figure 5 shows the change in cumulative returns for each index for DTW+ k NN, DDTW+ k NN, IDTW+ k NN, DTW+ k^* NN, DDTW+ k^* NN, and IDTW+ k^* NN. Notice that the variation in cumulative returns decreases in the order of DTW, DDTW, and IDTW; IDTW+ k^* NN exhibits an upward trend for all indices.

All of these results indicate that stock price predictions using IDTW+ k^* NN are superior to the other methods in terms of both profitability and accuracy. The total return by IDTW+ k^* NN exceeds all indices.

The prediction power shown above is competitive to that reported in [18], which uses text information. Although [18] requires additional information, i.e., the Nikkei Newspaper, the proposed method only requires stock price information. Its ease-of-use is an important characteristic of the proposed method.

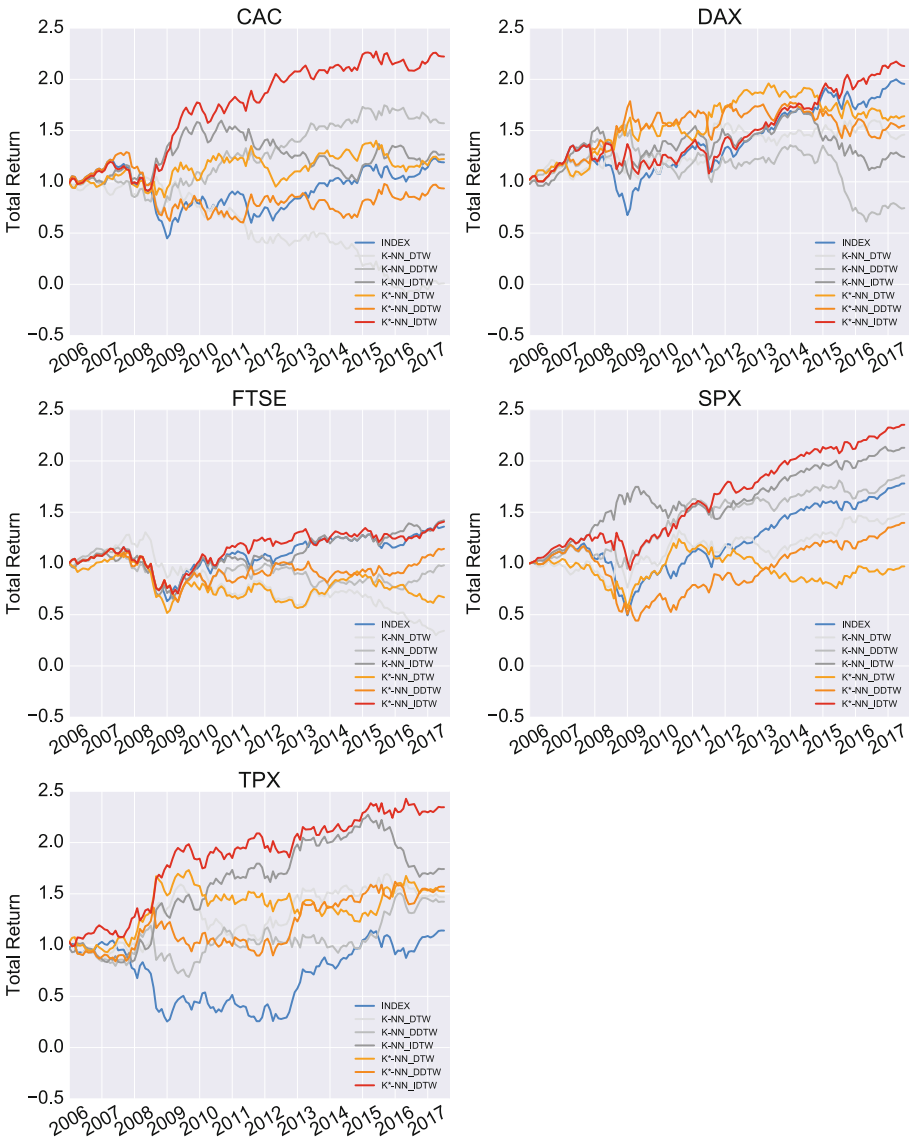


Fig. 5. Change in cumulative returns of each index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

Table 1. The average accuracy of all years for each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the six methods.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	<i>k</i> -NN	46.76%	51.80%	46.04%	52.52%	52.52%	49.93%
	<i>k</i> *-NN	52.52%	53.96%	50.36%	47.48%	50.36%	50.96%
DDTW	<i>k</i> -NN	51.80%	53.24%	51.80%	60.43%	51.08%	53.67%
	<i>k</i> *-NN	50.36%	56.12%	54.68%	58.99%	55.40%	55.11%
IDTW	<i>k</i> -NN	49.64%	53.24%	55.40%	61.87%	55.40%	55.11%
	<i>k</i> *-NN	57.55%	59.71%	57.55%	66.91%	60.43%	60.43%

Table 2. The total returns of each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the highest cumulative returns of the six methods.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	<i>k</i> -NN	0.98%	146.03%	34.32%	148.15%	146.79%	95.25%
	<i>k</i> *-NN	122.22%	164.03%	66.98%	97.12%	152.67%	120.60%
DDTW	<i>k</i> -NN	157.14%	74.38%	98.21%	185.76%	142.32%	131.56%
	<i>k</i> *-NN	93.53%	154.90%	114.29%	139.60%	156.98%	131.86%
IDTW	<i>k</i> -NN	126.59%	124.26%	142.10%	212.87%	174.20%	156.00%
	<i>k</i> *-NN	222.24%	212.91%	140.74%	235.29%	234.53%	209.14%

5 Conclusion

The contribution of this paper is two-fold. We proposed *k**-NN with IDTW, a highly accurate stock prediction method. We compared the performance of several prediction methods, including previous research method.(DTW+*k*NN, DDTW+*k*NN, IDTW+*k*NN, DTW+*k**NN, DDTW+*k**NN) to our proposed method.

An empirical analysis was conducted with major world indices and confirmed the following results:

- IDTW is superior to DTW and DDTW in terms of both profitability and accuracy.
- *k**-NN is superior to *k*-NN in terms of both profitability and accuracy.
- IDTW-*k**NN is the best prediction method in terms of both profitability and accuracy.

Our proposed method only requires price information and parameter L/C. This is a direct challenge to EMH and a remarkable advantage other than the accuracy of the proposed method. Our proposed method is following the similar trend. In other word, we look for past pattern similar to the current and predict

from that. This is reasonable as practitioners. Because this method approximates the way investors actually predict stock prices. We think investors pay attention to the form of fluctuation. Similar patterns cause similar investor behaviors and result in similar changes in stock prices. For further study, we have to check the impact of reference period choice on performance.

References

1. Anava, O., Levy, K.: k^* -nearest neighbors: from global to local. In: *Advances in Neural Information Processing Systems*, pp. 4916–4924 (2016)
2. Andrade, S.C., Chhaochharia, V., Fuerst, M.E.: “Sell in may and go away” just won’t go away. *Finan. Anal. J.* **69**(4), 94 (2013)
3. Asness, C.S., Moskowitz, T.J., Pedersen, L.H.: Value and momentum everywhere. *J. Finan.* **68**(3), 929–985 (2013)
4. Bergmeir, C., Benítez, J.M.: On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **191**, 192–213 (2012)
5. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **31**(3), 307–327 (1986)
6. Bouman, S., Jacobsen, B.: The halloween indicator, “sell in may and go away”: another puzzle. *Am. Econ. Rev.* **92**(5), 1618–1635 (2002)
7. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken (2015)
8. Cavalcante, R.C., Brasileiro, R.C., Souza, V.L., Nobrega, J.P., Oliveira, A.L.: Computational intelligence and financial markets: a survey and future directions. *Expert Syst. Appl.* **55**, 194–211 (2016)
9. Coelho, M.S.: Patterns in financial markets: dynamic time warping. Ph.D. thesis, NSBE-UNL (2012)
10. Fama, E.F.: Efficient capital markets: a review of theory and empirical work. *J. Finan.* **25**(2), 383–417 (1970)
11. Fama, E.F., French, K.R.: The cross-section of expected stock returns. *J. Finan.* **47**(2), 427–465 (1992)
12. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33**(1), 3–56 (1993)
13. Hamilton, J.D.: *Time Series Analysis*, vol. 2. Princeton University Press, Princeton (1994)
14. Imamura, M., Nakagawa, K., Yoshida, K.: Evaluation of financial market forecasting with using similarity of asset price fluctuation patterns (in Japanese). In: *The 31st Annual Conference of the Japanese Society for Artificial Intelligence*, p. 2D1-2 (2017)
15. Itakura, F.: Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust. Speech Sig. Process.* **23**(1), 67–72 (1975)
16. Jegadeesh, N., Titman, S.: Returns to buying winners and selling losers: implications for stock market efficiency. *J. Finan.* **48**(1), 65–91 (1993)
17. Keogh, E.J., Pazzani, M.J.: Scaling up dynamic time warping for datamining applications. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 285–289. ACM (2000)
18. Kuramoto, T., Izumi, K., Yoshimura, S., Ishida, T., Nakashima, A., Matsui, T., Yoshida, M., Nakagawa, H.: Analysis of long-term market trend by text-mining of news articles (in Japanese). *Trans. Jpn. Soc. Artif. Intell.* **28**(3), 291–296 (2013)

19. Mandelbrot, B.: The variation of certain speculative prices. *J. Bus.* **36**(4), 394–419 (1963)
20. Nakagawa, K., Imamura, M., Yoshida, K.: Stock price prediction using similarity of stock price fluctuation patterns (in Japanese). In: *The 31st Annual Conference of the Japanese Society for Artificial Intelligence*, p. 2D1-1 (2017)
21. Stone, M.: Cross-validators choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **36**, 111–147 (1974)



Characterization of Consumers' Behavior in Medical Insurance Market with Agent Parameters' Estimation Process Using Bayesian Network

Ren Suzuki¹, Yoko Ishino², and Shingo Takahashi^{1(✉)}

¹ Graduate School of Creative Science and Engineering, Waseda University, Tokyo, Japan
r-suzuki6@akane.waseda.jp, shingo@waseda.jp

² Graduate School of Innovation and Technology Management, Yamaguchi University,
Ube, Japan
ishino.y@yamaguchi-u.ac.jp

Abstract. In medical insurance market as well as other markets, it is not straightforward for an institution to develop effective marketing strategies because consumers' preferences and the environment surrounding consumers are constantly changing. This paper develops an agent-based model (ABM) of consumer's behavior in purchasing medical insurance products and analyzes the characterization of consumers' behavior to establish effective marketing strategies for the products. In general, the information propagation model of purchasing behavior has difficulty estimating the values of parameters only from ordinary marketing surveys, especially in the case of products that require a person to conduct advanced information processing, such as an insurance policy. To tackle this problem, this paper developed a method of estimating the probability parameters of agent's behavior using Bayesian network based on questionnaire survey data, and then evaluated the effectiveness of the method by applying it to the actual insurance market. In the analysis using ABM constructed, we mainly focus on the power of influence of the sales activity using word-of-mouth communication between consumers. As the result we obtained several key findings regarding marketing strategies that can be utilized in the real marketing of insurance products.

Keywords: Agent-Based Social Simulation · Bayesian network
Medical insurance market

1 Introduction

The primary purposes of this paper are to propose the whole steps of an agent modeling method that provides the way of constructing a valid agent model using Bayesian network, by which the probability parameter values of the agent behavior can be inferred. Then the effectiveness of the proposed method is evaluated by applying it to the actual medical insurance market to analyze the characteristics of the consumer behavior.

The insurance industry in Japan has undergone drastic changes since the new Insurance Business Law became effective in 1996, as aiming to loosen regulations on insurance companies. To date, a variety of medical insurance products have been launched in various sales channels such as sales representatives, the Internet and so on. Moreover, consumers' attitudes towards insurance products have greatly changed compared to before. Economic situations in Japan as well as in the world have also significantly changed in various ways. Accordingly, Japanese consumers came to put higher values on the insurance that enhances the medical treatment, the pension, and nursing, rather than the expensive life insurance against death. It has become more difficult to win new contracts for insurance institutions, because of changes in customer behavior as well as in environments. Insurance intuitions have been recently trying to win new contracts by strengthening the point of contact with existing contractors [1]. In order to make marketing strategies more effective, it is essentially meaningful to use a consumer's behavior model that can evaluate marketing strategies quantitatively. In this paper, we develop an agent-based model (ABM) by which we can deal with such diversities of consumers' behavior and environments. Then we characterize consumers' behavior which can be utilized in marketing strategies.

There have been several studies reported on the consumer's behavior of purchasing the insurance products. Ishino [2] analyzed the structure of customer-perceived value of insurance products relating to the medical treatment using a Bayesian network, then found that the word-of-mouth communication has a very large impact on the purchasing decision of the medical insurance product, and consumers' decisions are also affected by the way how to promote the features of medical insurance products. Miyazaki et al. [3] also examined the word-of-mouth communication suggested by Ishino [2] to find that the network topology of the social network which expresses the allocation of salespeople essentially affects the diffusion of the medical insurance products. Matsumoto et al. [4] analyzed the situation from the two perspectives: the mass media advertising and the activities of salespeople. It was found that the effectiveness of the sales promotion of marketing, especially the sales activity and the media advertisement, depends on the types of the medical insurance. Also it was found that the life stage of people affects the choice of the medical insurance product.

In general, identifying components of an agent model and estimating parameters of the model are the key in modeling consumers' behavior. In the market in our concern, consumers have diverse attributes and behavioral characteristics, and the product characteristics of medical insurance products should be essentially designed based on such consumer diversity. Hence, it is necessary to determine appropriately the components of a model that essentially affect consumers' behavior in purchasing medical insurance products. Furthermore, the internal parameters of a behavior model relating to consumer's product selection and information diffusion structure cannot be estimated straightforward in agent-based modeling. Hence we need to have a valid way to set these parameters.

In agent-based modeling parameter values that cannot be determined only by empirical knowledge or questionnaire data are usually calibrated so as to reproduce facts empirically accepted from empirical studies or problem situations to be analyzed. However, how the calibration goes well depends essentially on the skill of a modeler.

And the calibration does not work well to estimate simultaneously various types of parameters. Against these problems, some related works have been proposed: the method using Bayesian network as a behavior model [5], the inverse simulation [6], the virtual grounding method [7] and the estimation method using Bayesian network [4]. The method using Bayesian network as a behavior model [5] is one to provide definitely the cause-and-effect relationship of agent's behavior, though this method did not discuss how to determine parameter values in detail. The inverse simulation [6] is a method of searching and adjusting parameters using Genetic Algorithm so that the simulation results approach the given macro social indicator which represents the characteristics of the target society. This method does not work well to estimate simultaneously various types of parameters. The virtual grounding method [7] is a method to construct valid facsimile models in ABSS (Agent-Based Social Simulation) where the real world data is not available to build the behavioral model and to estimate the parameters, using a questionnaire survey in which participants are asked their behavior within the acceptable behavioral model with dynamically changing parameter values. This method can be applied only when a certain valid model has been already constructed and accepted, which cannot be used when you should make a model from scratch, like a case in this paper.

The estimation method using Bayesian network [4], which is our target method to be substantially improved, is a method of modeling agent's behavior to estimate the probability parameters using Bayesian network based on questionnaire survey. This method is expected to identify components of a model and estimate parameters of the model. This method developed so far, however, has not succeeded in constructing a valid model with consistent links among the behavioral nodes. The reason is that some unexpected relationships among the behavioral nodes are identified by Bayesian network. These unexpected relationships can be considered unnatural and do not play any role in the agent behavioral model because of the logical inconsistency contradicting the questionnaire survey. This paper provides an essentially improved method for it.

First in order to model the consumers' behavior in the medical insurance market, we propose a new method how to construct a valid model of agent's behavior, the probabilities of which can be inferred using Bayesian network basically identified from questionnaire survey data. Then this paper analyzes the characterization of the consumers' behavior using the constructed model. Miyazaki et al. [3] proposed a model of information diffusion model in medical insurance markets, though word-of-mouth communication was not modeled as the consumer's behavior. In addition, Miyazaki et al. [3] assumed only one kind of medical insurance products and could not analyze the impact of word-of-mouth communication on the insurance products. We develop a model describing word-of-mouth communication as the consumers' behavior and analyze the effectiveness of the word-of-mouth communication between consumers. Then we analyze the marketing strategies to strengthen the point of contact with existing contractors.

2 Proposed Method for Agent-Based Modeling

We propose a method that enables us to simultaneously construct the agent's behavior model and estimate the internal parameters within the model. The general procedure of the proposed method is as follows. Our method is different from the conventional way of modeling in the sense that a hypothetical behavior model is constructed in Step 1 and 2, and Bayesian network is revised to become consistent with the agent behavior. Then we apply this procedure to the medical insurance market described in the subsequent sections.

Step 0. Investigation of the Target Social System

We select the target social system, and roughly design the agents that constitute the system.

Step 1. Hypothesis Creation

About the selected system, we shed light on all of the factors that relate to an agent's behavior, using existing data and/or the experiences of experts. Then, we create the hypothesis of the agent's behavior model, in which the behavior is interpreted as cause-and-effect relationships of factors. In some cases, we can represent cause-and-effect relationships between some factors considered as causes of agent behavior. The hypothesis about the agent's behavior model can be expressed as a network structure of the factors.

Step 2. Agent-Based Model Design

We design the architecture of the ABSS targeted. Based on the created hypothesis about the agent's behavior model, we first design agent's behavior in detail using the specific parameter variables, and also design a system world (society). Determining variables relating to the agent's behavior is critical in the modeling, although the distribution of the variable values and the precise interactions between variables are unknown in this stage. In this stage we design the question items that should be asked to the respondents of the questionnaire.

Step 3. Questionnaire Survey

We design and perform the questionnaire survey in order to understand the behavior in the target system regarding the extracted factors. We should notice that the survey should meet the logical consistency. We extract appropriate respondents participating in the questionnaire so as to satisfy the surveyed conditions such as age, sex and so on.

Step 4. Validation and Update of Behavior Model

A Bayesian network is created from the survey data to be consistent with the hypothetical behavior model. The hypothetical behavior model is validated by evaluating the network structure using the statistical methods such as the Akaike's Information Criterion, Pearson's chi-square value and log-linear model.

Finally based on the validation result, the behavior model is updated. The relationships between factors assumed in the hypothesis but not considered valid should be deleted from the behavior model. On the other hand, the relationships between factors

even not assumed in the hypothesis but considered valid are included newly in the behavior model.

Step 5. Parameter Determination

The parameter values included in the agent’s behavior model are determined as the probabilities, which are inferred from the constructed Bayesian network.

3 Agent-Based Model in Medical Insurance Market

Applying the procedure described previously, we develop an agent-based model in a medical insurance market.

3.1 Hypothesis Creation About an Agent

As a result of analyzing consumers’ decision making of the purchase of the medical insurance products, it was found that there are two important factors: the current purchase status about the medical insurance, and the mood of having people think “it is time for me to purchase the medical health insurance” [8]. We call this mood “timely state of mind.” Moreover, the previous research indicated five motivations which solicit for the purchase of the medical insurance: changes in life stages, word-of-mouth communication, anxiety about health, the advertisement, and the insurance renewal [9]. The other previous research indicated the relationships between these motivations: the association of word-of-mouth communication and the insurance renewal and the association of anxiety about health and the advertisement [4]. Based on the foregoing knowledge, we created the hypothesis about the agent’s purchase behavior of the medical insurance, as shown in Fig. 1. An agent decides to make the purchase or the cancellation of the medical insurance going through some events including changes in its life stages, word-of-mouth communication, the advertisement, and so on.

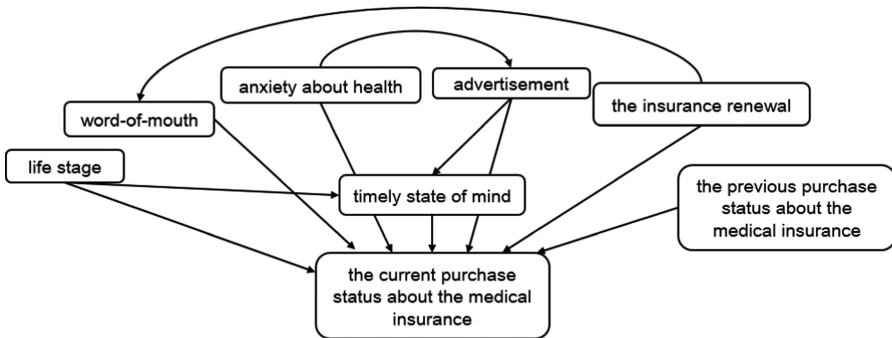


Fig. 1. Hypothetical agent’s behavior model of the medical insurance

3.2 Agent-Based Model Design

Figure 2 shows the conceptual scheme of the agent-based model that we constructed. An agent has five parameter variables: age, life stage, threshold of life event, event, and insurance status, as shown in Fig. 2. The age variable represents the agent's age, which increases according to the simulation steps.

When the agent's age reaches a certain threshold, a life event is generated and the agent's life stage alters. Seven kinds of life events are supposed: heading into the workforce, getting married, giving a birth to a baby, sending a child to college, buying a home, bringing up a child to be independent, and retiring from work. According to insurance company's surveys [9, 10], these life events are considered as significant for purchasing insurance products by each insurance industry. Since it is assumed that these seven kinds of life events happen in this order, eight life stages are set in total. The threshold of life event variable indicates the threshold value of the agent's age, and the life event occurs when the agent's age exceeds the threshold. Each agent has the unique value of the threshold of life event variable given at the initial step of the simulation. The event variable indicates a single-shot event other than the life events described above, which leads to the purchase of the medical insurance product. The event variable specifically consists of arrays of states of four kinds of incidents: word-of-mouth communication, anxiety about health, the contact to the mass media advertisement, and the insurance renewal. Finally, the insurance variable indicates the purchase status on eight kinds of the medical insurance described as follows. Since you are given a choice from four types of the medical insurance (general medical insurance, cancer insurance, specified disease insurance, and nursing insurance) and two modes of the contract agreement (the main contract and the special accessory contract), there are eight kinds of the medical insurance, i.e., four times two equals eight.

We constructed a consumer network of agents as a place or a society where agents interact with each other. Edges in a network indicate the close relationships between agents. For example, an event to tell a friend about what an agent experienced can be

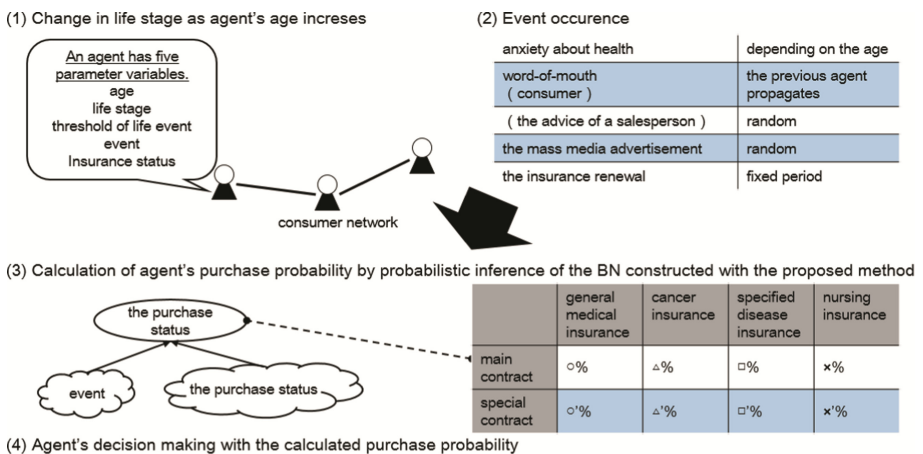


Fig. 2. Conceptual scheme of the model

realized using the edges. The consumer network topology is identified based on the survey data.

At each simulation step, agent's purchase probability of each insurance product is calculated by using probabilistic inference of the Bayesian network constructed with the proposed method, since the posterior purchase probability changes depending on the agent's life stage and the status of other events. Then, according on the calculated purchase probability, an agent makes a purchase decision.

3.3 Questionnaire Survey

Based on the agent's behavior model derived from the hypothesis, we designed the questionnaire survey. The previous research [4] revealed that there are the unexpected relationships between factors, and the logical inconsistency contradicting the questionnaire survey. Hence the current survey was designed to improve these problems. The number of the respondents was determined by the following equation.

$$n = \frac{N}{\left(\frac{CI}{2k}\right)^2 \times \frac{N-1}{p(1-p)} + 1} \quad (1)$$

where N is the parent population. As the number of the parent population, we used 103,357,000, the number of 20-year-old or more men and women as of Mar. 1, 2015. And C , k , p , and n represent the confidence interval coefficient, the reliability coefficient, the population proportion, and the needed sample number, respectively:

$$n = 666 \text{ when } C = 0.1, k = 2.58, p = 0.5.$$

Considering some respondents might not appropriately answer, the number of the samples was decided as 800. The allocation of respondents in terms of sex, age and purchase rate was made to represent the current Japanese insurance market situation. The questionnaire survey was performed through the Internet from Dec. 4 to Dec. 10 in 2015.

3.4 Validation and Update of Behavior Model

An information criterion such as Akaike's Information Criterion (AIC) is usually applied to construct a network of nodes. We should notice that if only AIC is applied to all the nodes at once, the resultant structure could include some invalid and unacceptable causal relations among the purchasing factors. Hence to improve such unacceptable results, the method this paper proposed classifies the relations of nodes and applies different statistical criteria to each class of nodes.

Each node described in the graphical model of the agent's behavior such as "the anxiety about health" expresses one item made in the questionnaire and has two states: the value of which is 1 if the agent is worried about its health, otherwise is 0. As a result of the survey design, two types of factors are identified: factors formed with only one

node and factors formed with more than one node. We call the factor in the latter case a group entity. For example, the life stage is a group entity consisting of seven nodes, including “heading into the workforce,” “getting married,” “giving a birth to a baby,” and so on.

Then the relations of nodes are classified into three types: the relations of nodes within a group entity, the relations of nodes belonging to different group entities, and the relations of nodes not belonging to any group entities. According to these types, the hypothetical behavior model is validated in order by the following procedures.

- (1) **For the relations of nodes within a group entity.** The connections of nodes of this type are made based on the Akaike's information criterion (AIC) value because there is no hypothesis of the cause-and-effect relationships of the group. Figure 3 shows the part of the network where AIC is applied. As the result of the evaluation, “getting married” and “giving a birth to a baby” are newly connected by an edge whose direction is from “getting married” to “giving a birth to a baby.”

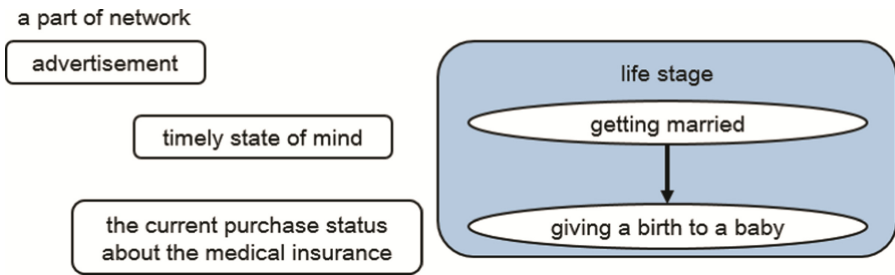


Fig. 3. Application of Akaike's Information Criterion (AIC) to group categories

- (2) **For the relations of nodes belonging to different group entities.** Pearson's chi-square testing is applied to evaluate the linkage strength of all the combinations of nodes of this type. If two nodes significantly correlate to each other, they are connected by an edge (10% significance level is used in this paper). The directions of the edges, which represent cause and effect, are already determined from the hypothesis. Figure 4 shows the part of the network where Pearson's chi-square testing is applied. As the result of the testing, for instance, “advertisement” and “the current purchase status about the medical insurance” are connected by an edge whose direction is determined by the hypothesis in Fig. 4.

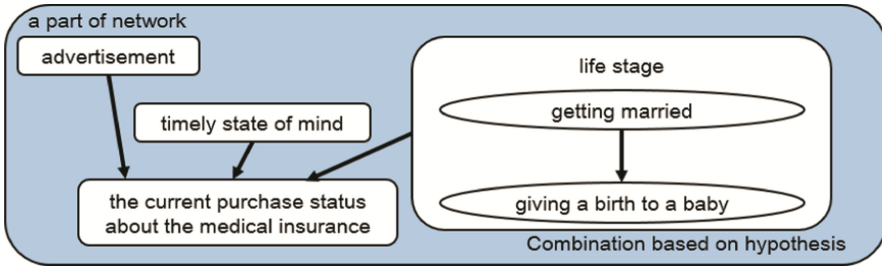


Fig. 4. Application of Pearson’s chi-square testing to all combinations of nodes belonging to different groups

(3) **For the relations of nodes not belonging to any group entities.** The log-linear model is applied to test the relations of nodes of this type. This method considers interactions between more than two nodes as well as two nodes because the whole structure of BN is not valid. If two nodes significantly correlate to each other and the connection can be reasonably explained, they are connected by an edge. Figure 5 shows the part of the network where the log-linear model is applied. As the result, “advertisement” and “timely state of mind” are connected by an edge in Fig. 5.

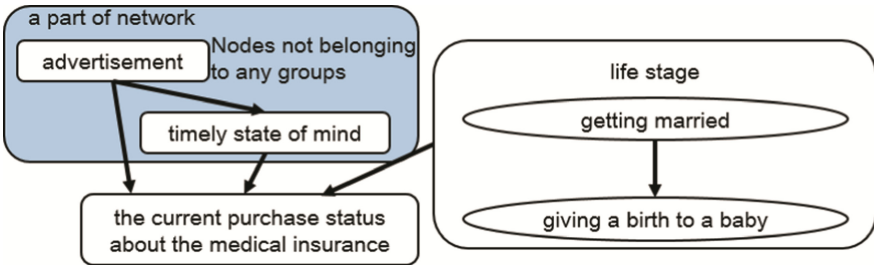


Fig. 5. Application of the log-linear model to the nodes not belonging to any groups

Based on the result of the validation of the behavior model, we need to modify the agent’s behavior. The factor of “heading into the workforce,” one of the life stage, is not associated with other factors. Hence “heading into the workforce” is removed from the life stage. Figure 6 shows the agent’s behavior model finally modified. “Life stage” consists of 6 nodes and 2 states, “the previous purchase status about the medical insurance” consists of 8 nodes and 2 states, and “the current purchase status about the medical insurance” consists of 8 nodes and 2 states. The consequent behavior model has several parts different from the hypothetical behavior model. The resultant Bayesian network is exhibited in Fig. 14 in Appendix.

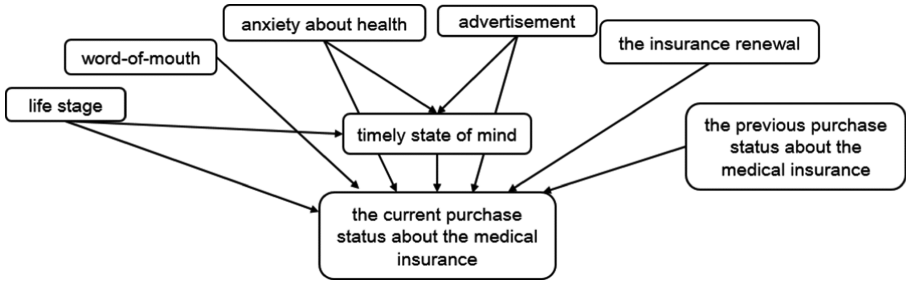


Fig. 6. Agent's behavior model modified

Subsequently, we estimate a consumer network topology based on the survey data. In the social scientific network analysis, the power exponent is roughly applied to analyze scale-free property because the analytical data include outliers or missing data [11]. In this paper, we create the degree distribution from the survey data and analyze scale-free property by confirming rough approximation of the power exponent. Then, we estimate the consumer network topology that reproduces the power exponent obtained from the survey data. To analyze the effectiveness of word-of-mouth communication, we construct a consolidated network. The degree distribution is created from the survey data excluding degree 0. The value of the power exponent γ calculated from the degree distribution is 1.4, which is smaller than common value that show scale-free property [12]. Since the CNN model [13] with small exponents can be constructed, we reproduce the same power exponent in the CNN model. Figure 7 shows a comparison in the log logarithm of the degree between the survey data and the CNN model. The connection probability of the CNN model is 0.15. Compared with the survey data, we could construct the CNN model with consistent degree distribution to some extent. In this paper, we adopt the CNN model as the consumer network.

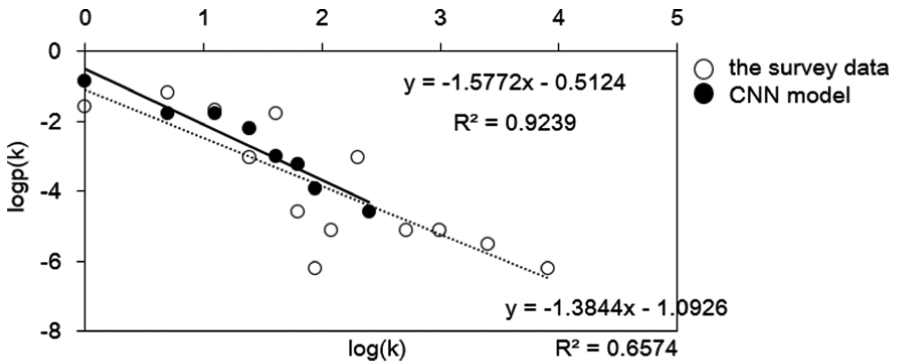


Fig. 7. Comparisons of log logarithm of degree

3.5 Parameter Determination

The parameter values to be obtained for executing the simulation are given as the probabilities to purchase each insurance product when an agent reaches a life stage. In the behavior model shown in Fig. 6, the number of nodes we can operate is 18, because the life stage and the current insurance finally have 6 and 8 nodes, respectively, and other 4 variables have only one node each. Since all these variables are binary, the total number of an agent’s state is 2^{18} . The parameter values are computed in all kinds of an agent’s state from the probabilistic inference of Bayesian network based on the obtained graph structure of the agent’s behavior model. In short, the objective variables, which are the values of the purchase probability of eight kinds of insurance products, are calculated under assumption that the state of 18 explanatory variables can be observed (Fig. 8).

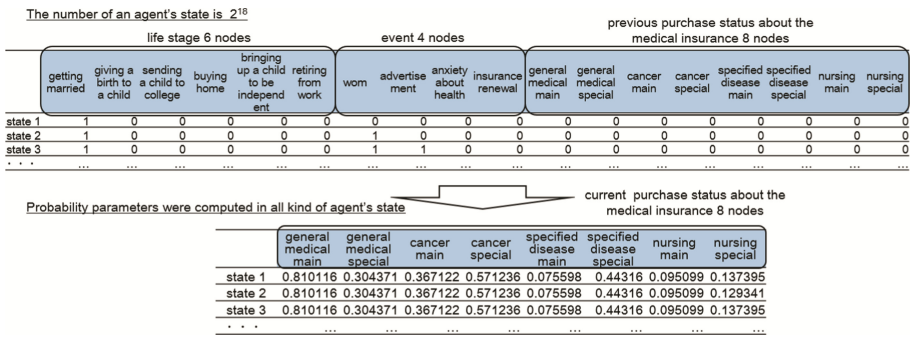


Fig. 8. Procedures of parameter determination

4 Computer Simulation Experiments

The simulation experiments are performed to analyze scenarios about the purchasing behavior. The population of agents are 100. An execution time step in the simulation experiment corresponds to six months in the real world. A trial of the simulation requires 10 time steps standing for five years. In each case, 100 trials, which can be expect to produce a sufficient variety of behavior, were performed.

4.1 Validation of Obtained Parameter Values

For the validation of the model, we should verify how well the model by the proposed method reproduced the questionnaire survey data. However, since there are a large amount of probability parameters, it is difficult to directly examine, in terms of a numerical way, the validity of the obtained parameter values. Hence we validate the model qualitatively by comparing the simulation results with the survey data.

The dots in Fig. 9 shows the ratio of agents that possessed the medical insurance products of 8 categories of the insurances: general medical main/special, cancer main/special, specified disease main/special and nursing main/special, when the simulation

finished. Consequently, we see that the model reproduced to the enough extent the ratio of agents that possessed each medical insurance product compared to the questionnaire survey data under the condition that maintains the structure with consistent links among behavioral nodes.

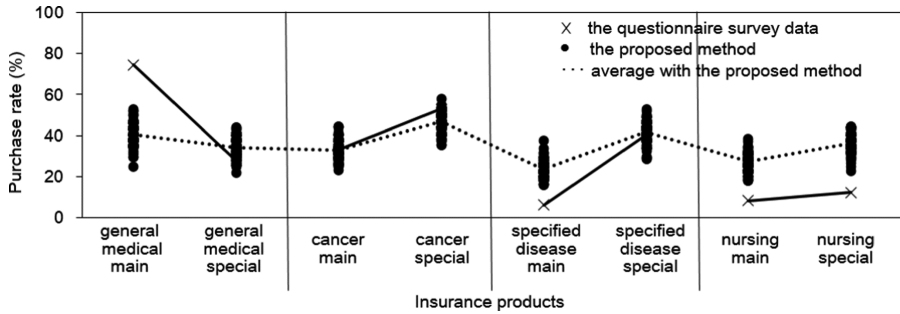


Fig. 9. Comparisons of simulations results and questionnaire survey data

4.2 Characteristic Analysis

The scenario analysis is conducted to analyze the characterization of the consumer's behavior. The analysis has been widely used as a means to produce information that will guide decision-making under a variety of alternative futures. We can find the possibility of changing the target system by performing the simulation under some scenarios composed of a combination of the several situation alternatives that would trigger the distinctive market configuration and several strategic options that would affect the market situation. We call them a situation alternative and a strategic option, respectively.

Firstly, we analyze the effectiveness of sales activity with the word-of-mouth communication between consumers. The above-mentioned six kinds of life events are set as situation alternatives. The ratio of word-of mouth communication are set as strategic options. We set two types of strategy: the sales activity with word-of-mouth communication, and the sales activity without word-of-mouth communication. In the case of the sales activity with word-of-mouth communication, we assume that consumers who purchased the products in the previous step tell the friends about what they experienced. The word-of-mouth event can be expressed as edges of CNN model. The event probability is determined as 0.296, which is calculated based on the questionnaire survey. There are 12 (= 6 × 2) scenarios in total for six situation alternatives and two strategies. For each scenario, 100 trials of the simulation were performed. The mean values of the purchase rate of each medical insurance product under each scenario are depicted in Fig. 10. From Fig. 10, we see that the purchase rate falls when word-of-mouth communication does not occur. The variations of rates vary depending on the kind of insurance products and the life stage. The model constructed shows that consumers have many opportunities for the consideration of purchasing insurance products because of the experience of word-of-mouth communication, and consumers who experience word-of-mouth communication tend to purchase insurance products

positively. This results suggest the possibility of winning new contacts using the referral marketing such as word-of-mouth communication. Then, we investigate the effectiveness of the social network topology by comparing other network topologies. We constructed other agent’s behavior models which include not the CNN model but a regular network. The simulation experiments were executed under the scenarios as stated using the model. Figure 11 depicts the differences of the mean values of the purchase rate of each medical insurance product obtained by the experiments of each models. The numbers of word-of-communications in each trial are depicted in Fig. 12. From Fig. 12, we see that the number of word-of-mouth communication in each trial is larger than in the case of the models including a regular network. From Fig. 11, we see that most of insurance products increase in the purchasing rates. In particular, the insurance products: specified disease special and nursing special are significantly effective on the word-of-mouth communication.

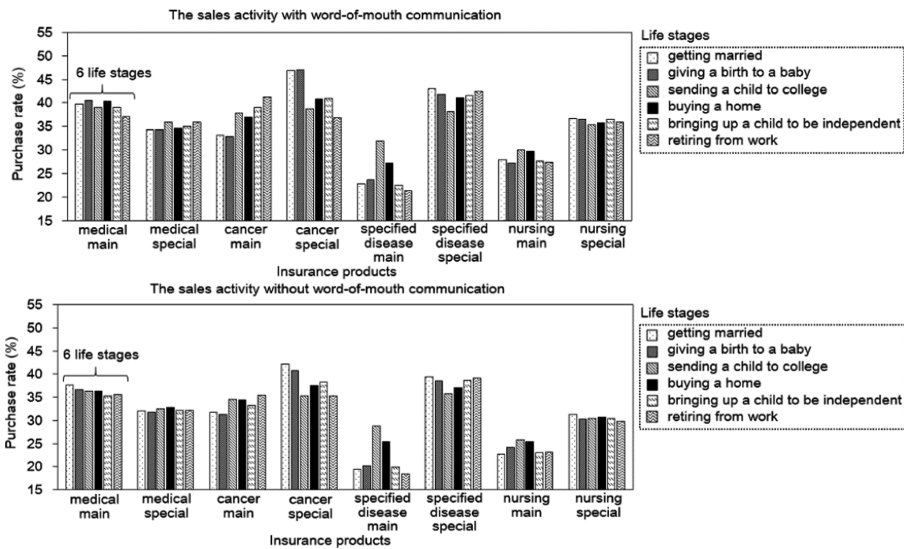


Fig. 10. Purchase rates of medical insurance products under scenarios

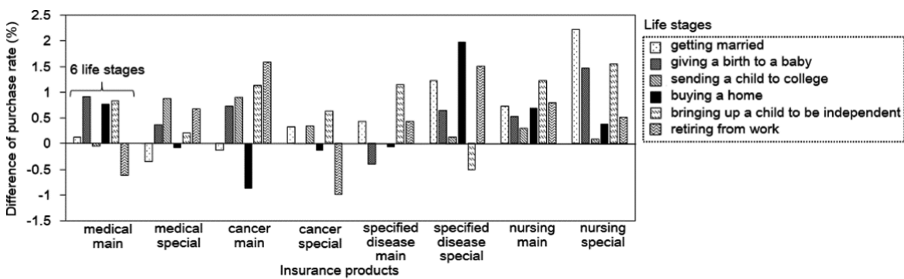


Fig. 11. Comparisons of difference of purchase rates by two models

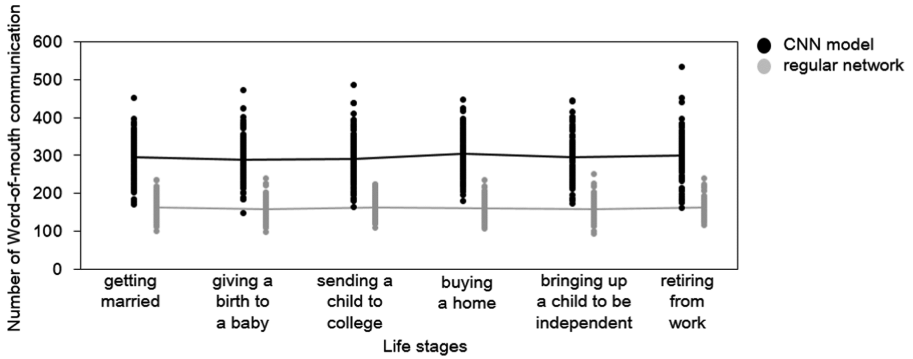


Fig. 12. Number of word-of-mouth communication in each trial

Subsequently, we investigate the effectiveness of the marketing strategies to strengthen the point of contact with existing contractors. The above-mentioned six kinds of life events are set as situation alternatives. The insurance statuses are set as strategic options. In particular, two types of strategies are experimented: the sales activity for consumers who have ever purchased the insurance products, and the sales activity for consumers who have never purchased the insurance products. There are 12 ($= 6 \times 2$) scenarios in total, since six situation alternatives and two strategies are prepared. For each scenario, 100 trials of the simulation were performed. Figure 13 compares the mean values of the purchase rates of each medical insurance product under each scenario. From Fig. 13, we see that there are some insurances whose purchase rates increase and others which are not when the sales activity for consumers who have ever purchased the insurance products is applied. In other words, the purchase rates of the main contract of the general medical insurance product and the special contract of the cancer insurance increase when the sales activity for consumers who have never purchased the insurance products is applied. On the other hand, the purchase rates of the not-mentioned insurance increase when the sales activity for consumers who have ever purchased the insurance products is applied. Consumers who have ever purchased consist of the existing contractors and new customers for insurance intuitions. Consumers tend to purchase the products when they acknowledge the credibility of sales representatives [13]. The number of the existing contractors who purchase new products are more than that of new customers who does because the sales representatives build a closer relationship with the existing contractors. Accordingly, we showed that the marketing strategies to strengthen the point of contact with existing contractors was effective. Also, we suggested the products to sell for acquiring new customers who have never purchased the products.

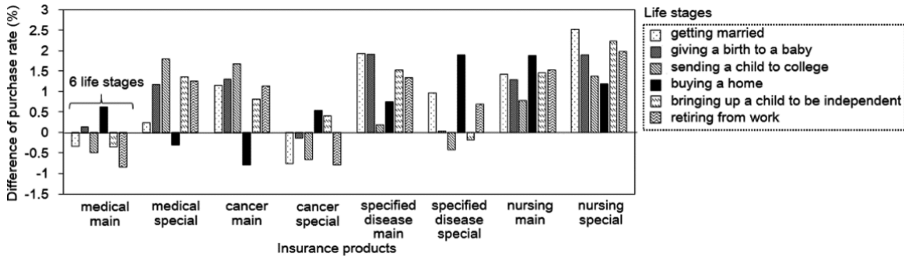


Fig. 13. Comparisons of difference of purchase rates under scenarios

5 Conclusion

In this paper, we proposed a method to model agent’s behavior in ABSS using Bayesian network based on questionnaire survey data. By applying the proposed method, we constructed a consumer behavior model that could analyze the characterization of consumers’ behavior. Moreover, we estimated the internal parameters by applying the probabilistic inference of Bayesian network based on the obtained graph structure of the agent’s behavior model. As a result of the simulation experiments using the behavior model, it was found that the effectiveness of word-of-mouth communication between consumers depended on the kind of the medical insurance and the life stage. In particular, the word-of-mouth communication has large impact on the insurances: the specified disease and the nursing special. Also it was found that the purchasing behavior of consumers who have ever purchased the medical insurance is different from that of consumers who have never done. From these findings, we suggested that the sales activity with the word-of communication was effective because consumers who purchased the products recommended the products to their friends. And we showed that the marketing strategies to strengthen the point of contact with existing contractors was effective and the marketing strategies to acquiring new customers who never purchased is to sell the main contact of the general medical insurance product and the special contract of the cancer insurance.

Appendix

See Fig. 14.

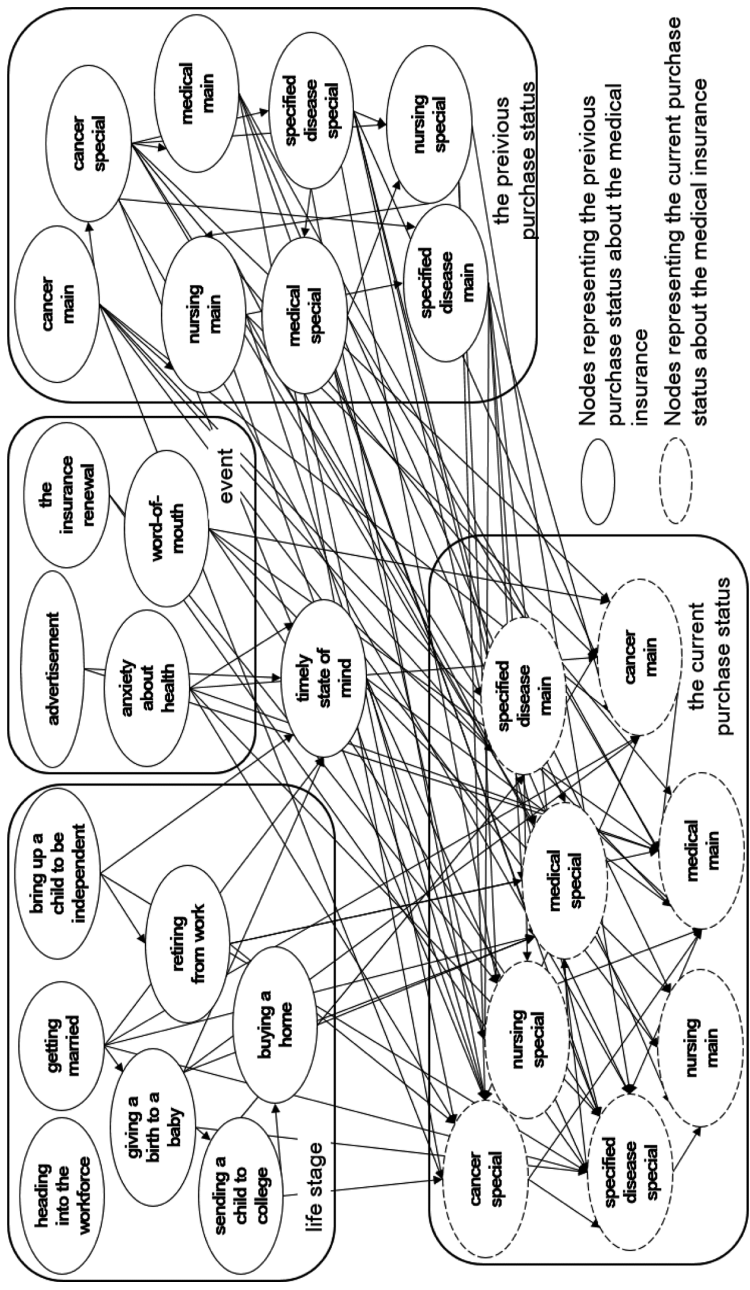


Fig. 14. Resultant Bayesian network

References

1. Tanaka, T.: A study on current life insurance sales and their action to consumers of life insurance industry. *J. Life Insur.* **180**, 47–75 (2012). Japan Institute of Life Insurance (in Japanese)
2. Ishino, Y.: Analysis and modeling of customer-perceived value of medical insurance products. In: Murata, T., Terano, T., Takahashi, S. (eds.) *Agent-Based Approaches in Economic and Social Complex Systems VII. Agent-Based Social Systems*, vol. 10, pp. 115–127. Springer, Tokyo (2013). https://doi.org/10.1007/978-4-431-54279-7_9
3. Miyazaki, M., Ishino, Y., Takahashi, S.: Effects of word-of-mouth communication on product diffusion: a case of medical insurance product. In: *Post-Proceedings of the AESCS International Workshop 2013, Agent-Based Approaches in Economic and Social Complex Systems VIII*. Springer (2013, in press)
4. Matsumoto, O., Miyazaki, M., Ishino, Y., Takahashi, S.: Method for getting parameters of agent-based modeling using bayesian network: a case of medical insurance market. In: Daud, A.R., Putro, U.S. (eds.) *Agent-Based Approaches in Economics and Social Complex Systems IX*, pp. 45–57. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-3662-0_4
5. Kocabas, V., Dragicevic, S.: Bayesian networks and agent-based modeling approach for urban land-use and population density change: a BNAS model. *J. Geogr. Syst.* **15**(4), 403–426 (2012)
6. Kurahashi, S., Minami, U., Terano, T.: Inverse simulation for analyzing emergent behaviors in artificial societies. *Trans. Soc. Instrum. Control Eng.* **35**(11), 1454–1461 (1999). (in Japanese)
7. Ohori, K., Iida, M., Takahashi, S.: Virtual grounding for facsimile model construction where real data is not available. *SICE J. Control Meas. Syst. Integr.* **6**(2), 108–116 (2013)
8. Ishino, Y.: Consumer survey data analysis to model internal state of an agent. In: *Proceedings of 5th Symposium of Technical Committee on Social Systems*, Okinawa, Japan, 5–7 March 2014, pp. 197–200 (2014). (in Japanese)
9. Kuribayashi, A.: Possibility of word-of-mouth in marketing of life insurance products, REPORT 2008.4, Report of Nissay Basic Research Center (2008). (in Japanese)
10. Aflac: Basics of Insurance You Want to Know. (in Japanese). <http://www.aflac.co.jp/soudan/guide/knowledge/worksheet/>. Accessed 6 Jan 2018
11. Ogai, Y., Matsumura, Y., Otani, T., Takatera, M., Hoshino, Y., Yasuda, T., Ohkura, K.: Complex network analyses on BtoB networks on Japanese textile and apparel industry. *J. Jpn. Res. Assoc. Text. End-Uses* **58**, 590–598 (2017). (in Japanese)
12. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**(1), 47–97 (2002)
13. Vázquez, A.: Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* **67**(5), 056104 (2003)
14. Tanaka, T.: Consideration concerning sales representatives of life insurance products. *J. Life Insur.* **169**, 1–28 (2009). Japan Institute of Life Insurance (in Japanese)



Do News Articles Have an Impact on Trading? - Korean Market Studies with High Frequency Data

Sungjae Yoon^(✉), Aiko Suge, and Hiroshi Takahashi

Graduate School of Business Administration, Keio University, Hiyoshi 4-1-1,
Kohoku-ku, Yokohama-shi, Kanagawa-Ken, Japan
{endeavourysj, aikosuge}@keio.jp, htaka@kbs.keio.ac.jp

Abstract. News is an important source of information for investment decision-making. Many studies analyzing listed companies in the US & Japan have been reported. However, the number of studies focusing on Korean stock markets is limited. This study analyzes the influence of news articles on Korean stock markets with high frequency trading data. Especially, we focus on analyses of the relationship between news articles and financial markets. Furthermore, we also analyze differences in market reactions according to language (English or Korean) of news articles and present three case studies.

Keywords: Finance · Asset pricing · High frequency trading data
Information economics · Big data analysis

1 Introduction

News articles are important sources of information for investment decision-making. Market efficiency is a central hypothesis in traditional financial theory, and information plays a significant role in efficient markets [3, 4, 7, 11, 15, 16]. Most empirical studies in finance use quantitative data such as asset prices, volumes, information release time, or other measurements that are easy to quantify. In the asset management business, investors make their investment decisions by utilizing various kinds of information, including textual information from media outlets such as newspapers, in addition to numerical data¹ (Fig. 1). Also, there are several analyses suggesting that textual information conveys different information to financial markets when compared to numerical information.

Many studies analyzing listed companies in the US & Japan have been reported. However, the number of studies focusing on Korean stock markets is limited. The features of Korean stock markets are as follows; (1) high trading turnover ratio (a measure of stock liquidity calculated by dividing the total number of shares traded over a period by the average number of shares outstanding for the period, 126% in 2016)²

¹ Recent progress in computer science contributes to financial studies. Various kinds of methods proposed in computer science -such as support vector machine, deep learning, agent-based modeling and network analysis- have been applied to financial research [2, 17, 18].

² Source: The World Bank.

and (2) high foreign stock ownership ratio, that is higher than Japan (45% vs 35%, 2012)³. Owing to those features, we assume that understanding the effect of news is more important for investors acting in Korean markets. With this in mind, we attempt to analyze the relationship between news articles and stock trading using high frequency trading data⁴. Furthermore, we also analyze differences in market reactions according to the language of news articles and present three case studies to develop a deeper understanding of the issues at hand. The remainder of the paper is organized as follows: First, we detail previous research and data employed in this analysis. Next, we show our method and results. Finally, we summarize this paper.



Fig. 1. Information and financial markets.

2 Related Works

Antweiler and Frank [1] analyze the messages posted on Yahoo! Finance and Raging Bull and find a significant relationship between stock messages and market volatility. Tetlock et al. [19, 20] quantify the language used in financial news stories from the Wall Street Journal and the Dow Jones News Service in an effort to predict firms' accounting earnings and stock returns. These studies suggest that textual media content, especially in terms of negative words, captures otherwise hard-to-quantify aspects of firms' fundamentals, which investors quickly incorporate into stock prices.

A great number of studies focusing on the relationship between the contents of news articles and investor behavior have been reported [14]. For example, Engelberg et al. [6] analyze the dataset of short sellers' trading patterns alongside a dataset of news releases. This study appears to confirm that short sellers' trading advantage comes largely from their ability to analyze publicly available information. Garcia and Parsons [5] identify a causal relationship between financial reporting and stock market performance.

Analyses focusing on the Japanese stock market and Korean stock market have been published [8–10, 12, 13]. Among those studies, Goshima and Takahashi [8–10] analyze news articles provided by Thomson Reuters through machine learning and deep learning to estimate news indices which evaluate the contents of news articles.

³ Calculated from Korea Exchange (KRX) database and Nikkei Economic Electronic Databank System.

⁴ We previously showed that news articles have an impact on stock trading volume using the high frequency data of Samsung Electronics [21].

This work finds a significant relationship between news indices and financial market data (such as stock prices and trading volumes).

Although the number of analyses focusing on the influence of news articles has been increasing, most studies analyze only daily stock data. In this study, we attempt to examine the relationship between news articles and financial markets using high frequency trading data in Korean markets (Fig. 2).

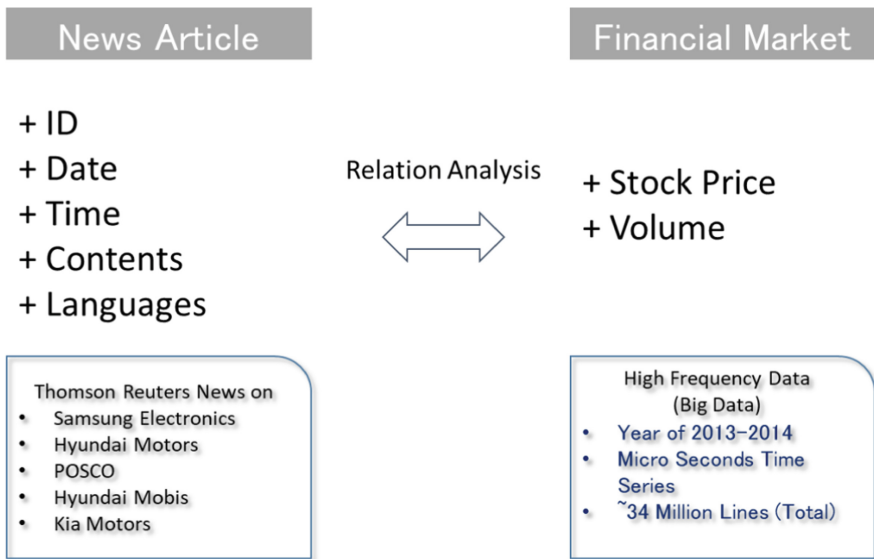


Fig. 2. Relation analysis of news articles on Korean companies and stock trading data.

3 Data

Stock Trading

We use high frequency trading data from Thomson Reuters. It contains micro seconds time series data of stock trading volume, stock price (bid/ask), and so on. This study focuses on the selected five Korean companies during the period of the financial years 2013 and 2014. As shown in Table 1, five companies are chosen by the rank of market capitalization value at the end of 2012. Their market capitalization value occupied around 30% of sum of all companies' market capitalization in the Korean stock market, so that a market share is sufficient to be representative of the features of the Korean stock market. Those companies stock trading data were extracted with ticker codes and the collected dataset includes more than 34 million lines so called "big data".

News Articles

The data of news articles is from Thomson Reuters, which is one of the most popular news sources in financial markets. News articles related to those companies above are extracted by the ticker codes and/or company names written in English or Korean.

Table 1. Details of selected five companies at the end of 2012.

Rank	Company	Market capitalization (billion KRW)	Occupation %	Ticker code
1	Samsung Electronics	224,189	19%	005930.KS
2	Hyundai Motors	48,130	4%	005380.KS
3	POSCO	30,428	3%	005490.KS
4	Hyundai Mobis	28,035	2%	012330.KS
5	Kia Motors	22,903	2%	000270.KS
Total		353,685	30%	

Those datasets contain various kinds of articles including those written in English, Spanish etc. In this study, we analyze articles written in 7 languages - English, Japanese, Chinese, Korean, German, Spanish or French -. Table 2 shows the numbers of news articles between 2013 and 2014. the second column (“24 h” column) means the whole news data in 24 h over a 2-year period. The total number of news articles is more than 12,500. The third column (“Trading time” column) represents the number of news articles released in daily trading time, from 9 a.m. to 3 p.m. in Korean time (UTC + 9). As shown in Table 2, English news is the major language in the dataset compared to the other 6 languages. Figures 3 and 4, respectively, display a time distribution of articles over 24 h (whole days) and in trading time (business days). Shown in those figures, the number of news articles increases just after the stock market closes and the number of news articles in trading time is about 28% of total news.

Table 2. Numbers of news articles (2013–2014).

Language	24 h (whole days)	Trading time (business days)
English	6,885	1,551
Japanese	1,647	781
Chinese	1,580	945
Korean	312	104
German	864	80
Spanish	755	40
French	464	21
Total	12,507	3,522

Amongst the five companies, the dataset of news articles relating to Samsung Electronics is the largest - more than half of all news articles in both 24 h data and trading time data. A similar tendency of distribution is observed in the news language comparison (data not shown).

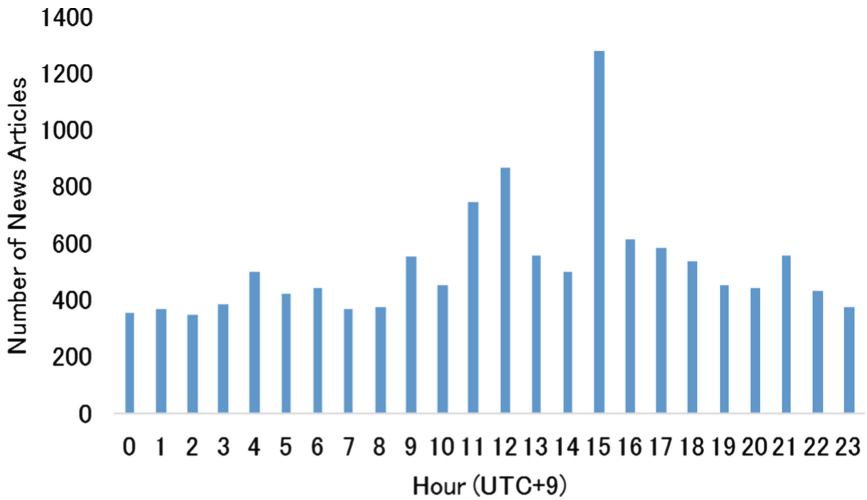


Fig. 3. Time distribution of news articles (24 h).

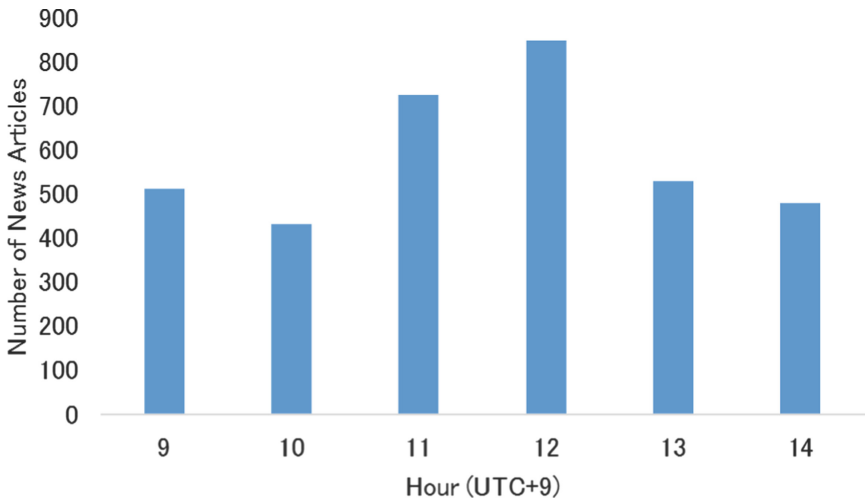


Fig. 4. Time distribution of news articles (trading time).

4 Methods

Firstly, we identify the time when each news article arrived (Fig. 5). Then we calculate the impact percentage that represents in relation to the whole trading volume of that company in one day's trading. In this analysis, we calculate the percentage every 15 min. At the same time, we also calculate the 1 min return volatility (standard deviation) for each period to observe price changes.

Next, we analyze three cases where news articles had large impacts on Samsung Electronics or Hyundai Motors in the period under analysis. By using these cases, we observe how a news order and/or news language has an impact on stock trading.

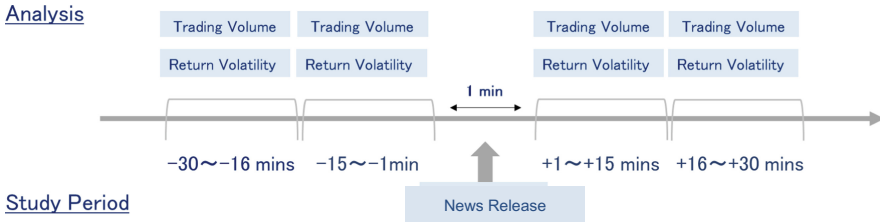


Fig. 5. News release and study periods.

5 Results

5.1 Trading Volume

Table 3 displays the trading volume percentage (%) of four periods; minus 30–16 min, minus 15–1 min, plus 1–15 min and plus 16–30 min. Figure 6 shows the result of three language cases (Total, English, and Korean cases).

Table 3. Trading volume (%).

Language	-30- -16 min	-15- -1 min	+1- + 15 min	+16- + 30 min
English	3.54	3.86	3.78	3.52
Japanese	3.64	3.47	3.45	3.41
Chinese	3.74	3.67	3.43	3.30
Korean	3.68	4.69	4.16	3.58
German	3.72	3.45	4.12	3.80
Spanish	4.49	4.71	4.52	4.83
French	3.99	4.02	5.67	2.50
Total	3.64	3.75	3.65	3.45

In most cases, the trading volume reaches its peak around the time that news is released. Those findings are consistent with previous studies which analyze Japanese stock markets [8–10].

The trading volume percentage and its peak time are different depending on the language. The differences might be due to the order -as well as the language- in which the news appears.

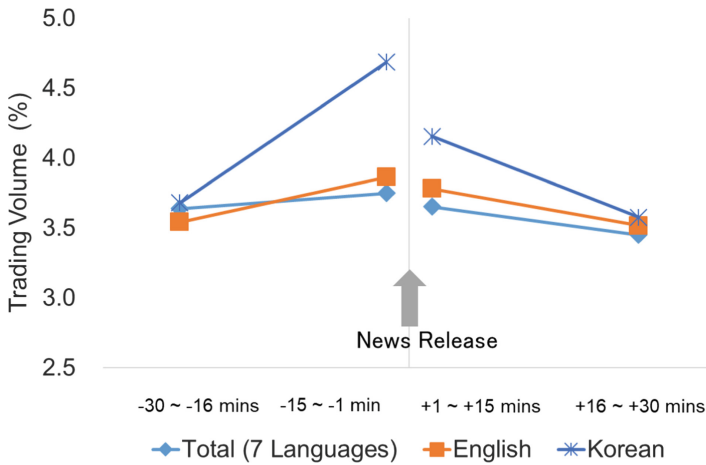


Fig. 6. Time course of trading volume percentage.

5.2 Return Volatility

In this section, we analyze how new articles affect the stock return volatility.

Table 4. Return volatility (0.1%).

Language	-30- -16 min	-15- -1 min	+1- + 15 min	+16- + 30 min
English	1.10	1.18	1.08	0.99
Japanese	1.27	1.10	1.04	1.02
Chinese	1.22	1.10	1.04	1.00
Korean	1.15	1.37	1.05	1.10
German	0.93	0.81	0.78	0.73
Spanish	0.97	0.99	0.87	0.74
French	0.78	0.88	0.81	0.58
Total	1.16	1.13	1.05	0.99

We analyze the 1-min stock return and then calculate the 15-min volatility (standard deviation). Results are shown in Table 4 and Fig. 7. From those results, news arrivals have almost no effect on the stock return volatility. Volatilities are limited to only about 0.1% and significant increases or decreases are not observed before and after news releases.

Periods of analysis are limited, so while it would appear clear that news articles have a direct impact on stock prices, it is less clear what effect they have on volatility over a more extended period. We are planning further analysis of stock price changes that news releases might cause.

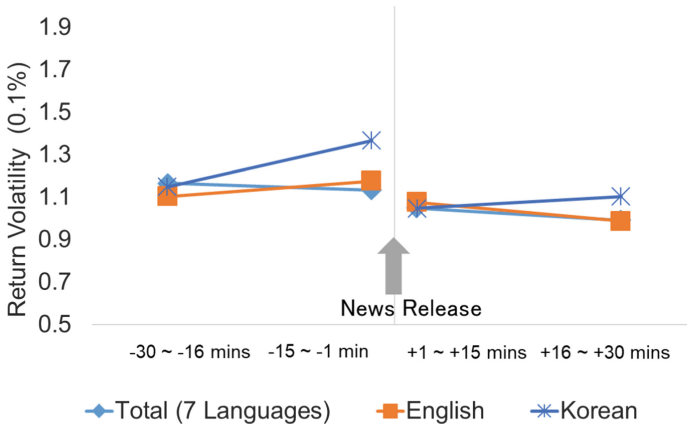


Fig. 7. Time course of return volatility.

In Sect. 5.3, we will touch on three cases in order to clarify our understanding of this observation.

5.3 Case Study

In this section, we will discuss changes of trading volumes and prices around news release time using three cases. The following cases were selected because we assume contents of news articles such as financial results, competitive information and investments, would have a more significant impact on stock trading.

Case 1: 8th January, 2013. The first case is a news article about 2012 fourth quarter financial results of Samsung Electronics (SE) released on 8th January, 2013. Figure 8 shows transitions of trading volume percentage of SE. In this case, an English news article was released before Korean news (09:08 and 09:43, respectively) and had a large impact on trading volume around 09:14.

Case 2: 6th August, 2014. The second case relates to the news of a withdrawal agreement regarding patent lawsuits between SE and Apple on 6th August, 2014. Figure 9 shows transitions of trading volume percentage of SE around that time. In this case, the Korean article of Thomson Reuters was released before the English news (11:11 and 11:21, respectively) and the latter seems to have had a larger impact on trading at around 11:23 than the Korean news article. However, further investigation reveals that most likely the earliest news at 11:03, relating to this withdrawal, was posted on NAVER (one of the most popular web portals in Korea). As a whole, the NAVER news article written in Korean produced the largest trading volume on this day.

Case 3: 18th September, 2014. The third case is about the huge investment of Hyundai Motor Group. Hyundai Motor Group was the highest bidder for the site of Korea Electric Power Corp <015760.KS>. Shown in Fig. 10, trading volume increases after the English news (10:44). While the earliest news had the greater impact in case 2,

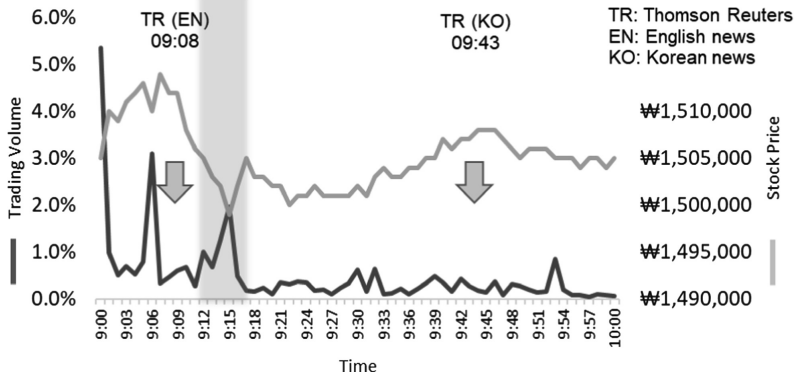


Fig. 8. Percentage of trading volume and stock price changes, 8th January, 2013. Announcement of 2012 fourth quarter financial results.

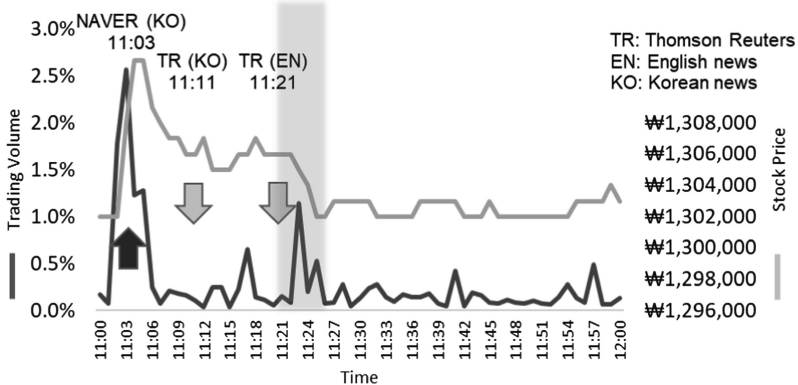


Fig. 9. Percentage of trading volume and stock price changes, 6th August, 2014. News of withdrawal agreement relating to patent lawsuits.

in this case no effect is observed after the news (10:29 from Yonhap; other major web portal in Korea) reported the same contents and was released at an earlier time. Those difference might reflect the news contents themselves and/or media power that releases news (ex. Naver vs Yonhap). We cannot draw a clear conclusion and further analysis is needed.

These three cases suggest that a significant relationship between news articles and financial markets exists. It would appear that English articles have a larger impact on trading volumes than those in Korean. That might be due to high foreign stock ownership ratio. Foreign investors presumably understand English more easily than Korean. This study also shows the possibility that the earliest news has a greater influence on trading volumes regardless of the language, but further analyses are planned.

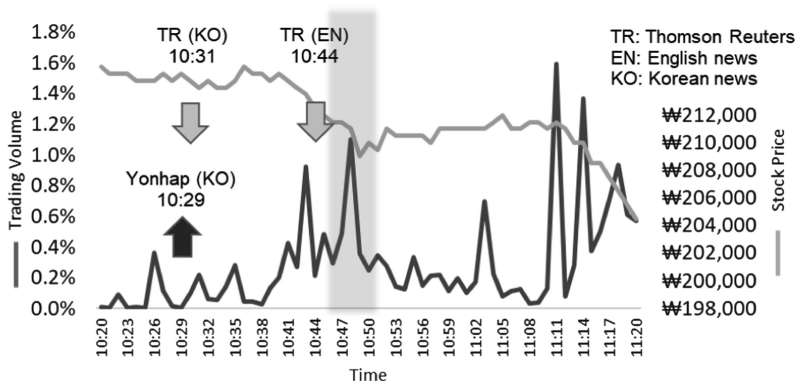


Fig. 10. Percentage of trading volume and stock price changes, 18th September, 2014. News of Hyundai Motors huge investment.

6 Conclusion

This article analyzes the influence of news articles on the stock trading of five Korean companies with high frequency trading data. In this study, we focus on the relationship between news articles and trading volume or return volatilities. As a result of intensive analyses, we find that the trading volume reaches its peak around the time when news articles are released, regardless of the language. These findings are consistent with previous studies [10]. We also observe that news articles seem to have almost no effect on return volatilities at least over the periods of time under analysis. Our study also suggests the possibility that English articles and more timely articles have a stronger influence on trading volume. However there remain complex questions over how exactly the news impacts on stock trading. Detailed analysis, such as text analysis using machine learning techniques, is planned for the future.

Acknowledgements. This research was supported by a grant-in-aid from the Kayamori Foundation of Informational Science Advancement.

References

1. Antweiler, W., Frank, M.Z.: Is all that talk just noise? The information content of internet stock message boards. *J. Financ.* **59**(3), 1259–1293 (2004)
2. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, Heidelberg (2006). <https://doi.org/10.1007/978-1-4615-7566-5>
3. Black, F., Scholes, M.: The pricing of options and corporate liabilities. *J. Polit. Econ.* **81**(3), 637–654 (1973)
4. Campbell, J.Y., Lo, A.W., MacKinlay, A.C.: *The Econometrics of Financial Markets*. Princeton University Press, Princeton (1997)
5. Dougal, C., Engelberg, J., Garcia, D., Parsons, C.A.: Journalists and the stock market. *Rev. Financ. Stud.* **25**(3), 639–679 (2012)

6. Engelberg, J., Reed, A.V., Ringgenberg, M.C.: How are shorts informed? Short sellers, news, and information processing. *J. Financ. Econ.* **105**(2), 260–278 (2012)
7. Fama, E.: Efficient capital markets: a review of theory and empirical work. *J. Financ.* **25**(2), 383–417 (1970)
8. Goshima, K., Takahashi, H.: Quantifying news tone to analyze Tokyo stock exchange with recursive neural networks. *Secur. Anal. J.* **54**(3), 76–86 (2016)
9. Goshima, K., Takahashi, H., Terano, T.: Estimating financial words' negative-positive from stock prices. In: *The 21st International Conference Computing in Economics and Finance* (2015)
10. Goshima, K., Takahashi, H.: Analyzing the relationship between news articles and high frequency trading data in Japanese stock markets. In: *The 24th Annual Meeting - Nippon Finance Association* (2016)
11. Ingersoll, J.E.: *Theory of Financial Decision Making*. Rowman & Littlefield, Lanham (1987)
12. Kim, Y.M., Willett, T.D.: News and the behavior of the Korean stock market during the global financial crisis. *Korea World Econ.* **15**(3), 395–419 (2014)
13. Lee, D.W., Cho, J.H.: Stock price reactions to news and the momentum effect in the Korean stock market. *Asia-Pac. J. Financ. Stud.* **43**, 556–588 (2014)
14. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* **66**(1), 35–65 (2011)
15. Luenberger, D.G.: *Investment Science*. Oxford University Press, Oxford (2000)
16. Sharpe, W.F.: Capital asset prices: a theory of market equilibrium under condition of risk. *J. Financ.* **19**(3), 425–442 (1964)
17. Takahashi, H., Terano, T.: Agent-based approach to investors' behavior and asset price fluctuation in financial markets. *J. Artif. Soc. Soc. Simul.* **6**(3), 1–3 (2003)
18. Takahashi, H.: An analysis of the influence of dispersion of valuations on financial markets through agent-based modeling. *Int. J. Inf. Technol. Decis. Mak.* **11**, 143–166 (2012)
19. Tetlock, P.C.: Giving content to investor sentiment: the role of media in the stock market. *J. Financ.* **62**(3), 1139–1168 (2007)
20. Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S.: More than words: quantifying language to measure firms fundamentals. *J. Financ.* **63**(3), 1437–1467 (2008)
21. Yoon, S.J., Suge A., Takahashi H.: *JSAI International Symposia on AI, Workshop 3: Artificial Intelligence of and for Business (AI-Biz 2017)*



Detecting Short-Term Mean Reverting Phenomenon in the Stock Market and OLMAR Method

Kazunori Umino¹(✉), Takamasa Kikuchi², Masaaki Kunigami¹,
Takashi Yamada³, and Takao Terano¹

¹ Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku,
Yokohama 226-8502, Japan

kl.umino@gmail.com, mkunigami@nifty.ne.jp,
terano@dis.titech.ac.jp

² Keio University, 4-1-1 Hiyoshi, Kohoku-ku, Yokohama 223-8526, Japan
tkikuchi63@gmail.com

³ Yamaguchi University, 1677-1, Yoshida, Yamaguchi 753-8541, Japan
tyamada@yamaguchi-u.ac.jp

Abstract. In this study, we examined the “short-term Mean Reverting Phenomenon” from two aspects. First, we clarified that excess return can be obtained by using the short-term Mean Reverting Phenomenon for the On-Line Moving Average Reversion (OLMAR) method, which is a portfolio selection algorithm and reportedly exhibits high performance. Then, we examined why the method was able to maintain superiority over the long term. In addition, we proposed an evaluation index of the short-term Mean Reverting Phenomenon present in the stock price dataset and analyzed it. The OLMAR method proved that excessive return was obtained by using the characteristic property showing the mean reverting tendency, which can be selected by the moving average divergence rate. Then, we confirmed that the advantage of the OLMAR method disappears by invalidating the above property existing in the stock price dataset. In addition, we proposed an evaluation index of the Mean Reverting and analyzed the stock price dataset using it.

Keywords: On-line portfolio rebalance algorithm · Anomaly · Mean reverting OLMAR · Evaluation index

1 Introduction

The purpose of this research is to detect the “short-term Mean Reverting Phenomenon” existing in the actual stock price dataset of about 20 years. We analyze cases of detection and use of Mean Reverting Phenomena in the actual stock price dataset, and propose the index showing the size of Mean Reverting existing in the stock dataset.

If stock price fluctuation is close to random, it is difficult to exceed the stock market index over the long term regardless of any trading method. Index is a simple average of returns in all stocks. Therefore, to outperform the index for a long time, it is necessary to find long-term and consistently detectable characteristics in the stock market and use

it to obtain excess returns. Therefore, the asset management method that outperforms the index for a long time needs to find long-term and consistently detectable characteristics in the stock market and use it to obtain excess returns.

Reports on the Portfolio selection algorithm (PSA) outperforming long-term index in simulation using an actual stock dataset are limited. In the existing research, the On-Line Moving Average Reversion (OLMAR) method has been reported as a PSA that provides the highest returns [1–4]. The OLMAR method is a PSA that selects stocks whereby stock prices have temporarily declined from the moving average, and from several days to several months, stocks that are likely to return to the moving average.

The OLMAR method is an approach with a reported ability to obtain excess return even if appropriate transaction costs are considered, and various improvements are added [5–8]. However, what kind of average regression phenomenon is detected and used has not been studied. In addition, no evaluation method has been developed to determine whether short-term mean reverting phenomenon exists in a target stock dataset and how large it is.

The subjects of this research are as follows: (1) Analysis of how the OLMAR method detects and uses “short-term Mean Reverting” that exists in the stock market to obtain excess return. (2) Proposal of a Mean Reverting Score (MRScore) and evaluate the trend and magnitude of the Mean Reverting in the stock price dataset.

Section 2 describes the portfolio reconstruction method and the OLMAR method. Section 3 shows evaluation indices, mean-reverting score, and definitions used in this research. Section 4 explains the analysis method of this paper. The results of verification of the newly prepared datasets are shown in Sect. 5. Finally, a summary is provided in Sect. 6.

2 Related Research

2.1 Portfolio Selection Algorithm

Algorithms that take advantage of the characteristics of price fluctuations more actively than usual have been developed from methods that control mean and variance [9], and have reportedly been classified into the following four types [3].

- (1) **Follow-the-Winner** is a method of positively selecting winning stocks (strong momentum stocks). A representative method is the Exponent Gradient (**EG**).
- (2) **Follow-the-Loser** is a method of expecting prices to shift from declining to rising and returning to average, by selecting a group of losers (prices of stocks that are falling). This is the **OLMAR** method [2, 3].
- (3) **Pattern-Matching Approaches** is a method that optimizes the portfolio selection method based on some pattern, with Correlation-driven Nonparametric Learning Approach for Portfolio Selection (**CORN**) [12] as a representative method.
- (4) **Meta Learning Algorithms** is a method of optimizing at the meta level; a typical method is the New Newton Step (**ONS**) [13] using Newton’s method.

Among the above four categories, **Follow-the-Loser** shows the highest performance, and the **OLMAR method** is an algorithm that shows the highest stable performance [3].

2.2 OLMAR Method

The OLMAR method preferentially a stock group with a large negative deviation rate of moving average deviation (MAD), and it is said to be one of the most return-oriented methods among all datasets [2, 3].

- (1) The maximum drawdown (MDD) is large and the temporary decline in asset price is large [2].
- (2) Frequent asset replacement occurs, and the return is reduced sharply with a slight increase in transaction cost [2, 6].

3 Evaluation Index, Score and Definition

3.1 Evaluation Index

The indicators used for evaluation in this research are explained. Let P_t be the price of the asset at time t ; the return R_t when holding this asset from time $t - 1$ to time t is as follows:

$$R_t = \frac{P_t}{P_{t-1}},$$

Indicators used in this paper are described below.

(MAD: Moving average divergence rate)

$$\begin{aligned} simpleMA_t^n &= \frac{1}{n} \sum_{k=1}^n P_{t-k+1}, \\ MAD_t^n &= 100(P_t - simpleMA_t^n)/P_t, \end{aligned}$$

(AR: Annualized return)

$$\begin{aligned} R^{day} &= \frac{1}{(n-1)} \sum_{k=2}^n \left(\frac{P_k}{P_{k-1}} \right), \\ AR &= (R^{day})^{252}, \end{aligned}$$

(MDD: Maximum Drawdown)

$$\begin{aligned} M_t &= \max_{u \in [0, t]} P_u, \\ D_t &= M_t - P_t, \\ MDD_t &= \max_{u \in [0, t]} D_u, \end{aligned}$$

3.2 MRScore (Index of Mean Reverting)

The MRScore is an index showing the magnitude of the average regression that occurs after the stock price falls in the stock price dataset. This method measures the momentum effect during the two periods of the previous term and the current term. The length in days of each term is expressed as “period”.

Equation 1 shows the momentum, which is the rate of change for a certain period of time. Equation 2 sets the product of the decline rate and the log return as the *Mean-Reverting-Index-Each* of the day for only cases where the momentum is minus. Equation 3 is a way of obtaining the Mean-Reverting-Index for a certain period of time and calculates the average value of Mean-Reverting-Index-Each of all stocks (N) in all periods (K). Equation 4 is the evaluation value obtained by correcting the Mean-Reverting-Index according to the period, which is called the MRScore.

If the MRScore is positive, there is a Mean Reverting trend, which is close to random if it is 0; if the MRScore is minus, it has a momentum effect instead of a Mean Reverting trend. When the momentum effect works, it shows a tendency to further decline after falling [14, 15].

$$R_{i,t}^{period} = \frac{P_{i,t}}{P_{i,t-period}}, \tag{1}$$

$$\begin{aligned} & \text{Mean - Reverting - Index - Each}(R_i, t, period) \\ &= \begin{cases} \left(\frac{1}{R_{i,t-period}^{period}} \right) \left(\log R_{i,t}^{period} \right) & \text{if } R_{i,t-period}^{period} < 1.0 \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{2}$$

$$\begin{aligned} & \text{Mean - Reverting - Index}^{period} \\ &= \frac{1}{N} \frac{1}{K - period} \\ & \times \left(\sum_{i=1}^N \sum_{t=period}^K \text{Mean - Reverting - Index - Each}(R_i, t, period) \right), \end{aligned} \tag{3}$$

$$MRScore = \sum_{m=1}^{25} \left(\sqrt{\frac{252}{10days}} \text{Mean - Reverting - Index}^{10m} \right), \tag{4}$$

3.3 Definition

Symbols and meanings used in this verification are defined below.

- N : Number of assets comprising the portfolio
- i : Stock i ($i = 1, \dots, N$).
- t : Time t ($t = 1, \dots, S$).
- $R_{i,t}$: Return of stock i at time t
- $MAD_{i,t}^{40}$: Moving average divergence rate of stock i at time t using a period of 40 days.
- $MDD_i^{40}(t1 : t2)$: Maximum drawdown of stock i from time $t1$ to $t2$

3.4 MAD-MRC and ReturnCurve

MAD-MRC (Moving average divergence rate - Mean reverting characteristics) is to obtain the neighborhood average value of the return on the next day below the lower checkPct% of the MAD value (Algorithm 1). In this paper, we deal with only the mean reverting phenomenon occurring after stock price declines. Therefore, the MAD value is “small” means that “negative deviation value is large”.

“ReturnCurve” is a diagram in which the x-axis is the MAD value and the “Daily return” of the neighborhood average value when all data are rearranged in the ascending order of MAD is taken as the y-axis.

The calculation algorithm is shown in Appendix 1.

4 Validation Contents

4.1 Analysis of Stock Price Fluctuation Characteristics

In this study, we analyze the stock price fluctuation characteristics that can obtain excess return by filtering stocks based on the moving average divergence rate used in the OLMAR method according to the following procedure.

- (1) Calculate the MAD of all stocks daily data.
- (2) After sorting in ascending order of MAD, an average of 2000 points in the vicinity is obtained, the value is taken as the average of return, and a return curve is obtained.
- (3) The average value of the lower $L = [0.25, 0.5, 1, 3]$ percent of the entire data is obtained from the return curve.

4.2 Data Correction Operation to Stock Price Fluctuation Characteristics

In this section, we verify that over profit cannot be obtained by the OLMAR method by corrected characteristics whereby excess return is obtained by MAD. In this research, when the return curve obtained by the moving average divergence rate has a certain tendency, the correction operation is performed on the data so that the return curve approaches a straight line ($y = \text{average value}$).

- (1) After obtaining $\text{MAD}_{i,t}^k$ in period k , find the return curve and rewrite the data value based on the formula below. Divide the unbalanced $L \leq 1.0$ (%) value into data dU of return equal to or more than 1.0 and other dD and obtain the standard deviation of each. Manipulate the data so that the standard deviations of dU and dD are equal, and replace modify_dU with dU (5, 6). When data are manipulated, if a difference occurs in the sum of returns, the add_mean is the difference that is evenly distributed to all data of $\text{MAD} < 0$ (7).

$$\text{ratio} = \frac{\text{stddev}(dD)}{\text{stddev}(dU)}, \quad (5)$$

$$\text{modify_dU} = 1.0 + \text{ratio} (\text{dU} - 1.0), \tag{6}$$

$$\text{add_mean} = \frac{(\text{sum}(\text{dU}) - \text{sum}(\text{modify_dU}))}{\text{number_of_data}(\text{dU and dD})}, \tag{7}$$

- (2) Repeat the above-described correction (operation of data) in the order of period $k = [20, 40, 80]$ to reduce the distortion (difference of standard deviation of rising/falling data) of the lower 1% with respect to the three moving averages.
- (3) It is also possible to modify the distortion according to the correction ratio a , if necessary. $A = 1.0$ in the case of total modification and $0 < a < 1.0$ in the case of partial modification (8).

$$\text{modify_dU} = 1.0 + (1 - a(1 - \text{ratio}))(\text{dU} - 1.0), \tag{8}$$

4.3 OLMAR_random Method

The average number of stocks constituting the portfolio and the average holding period of stocks is equal to the OLMAR method, and the number of transactions is the same. This method shows the performance when the detection and prediction ability of the mean reverting of the OLMAR method is invalidated. Details of the OLMAR_random method are shown in Appendix 2 (Algorithm 2).

4.4 Verification Procedure

In this research, we conducted the following verification:

- (1) Using the MAD, we analyzed the characteristics (return imbalance: distortion and trend of the return curve) by paying attention to the negative MAD. The OLMAR method is an algorithm that handles only falling stocks [3].
- (2) If there was an imbalance in return, a correction operation was made and the difference of the standard deviation of the rise/fall was corrected. This was done for a multi-moving average: MAD-MRC correct.
- (3) We simulated the original stock price data and the data on which the correction operation was performed by the OLMAR method and the OLMAR_random method, and analyzed the difference.

5 Validation Results

5.1 Dataset

The dataset used in this verification comprised a total of four types including the two original stock price datasets and the data correction operation shown in Sect. 4.2.

Characteristics of each dataset are shown (Table 1).

Table 1. Dataset and its information: Four types of datasets were used. Two real stock datasets (SP and NK) included “Anomalies”, while SP_mod and NL_mod had “Anomalies” removed.

Data set name	Information	Other features
SP: S&P500 N = 1688338	1999.1–2016.4 4359days 391stocks Hurst exponent (mean) = 0.46	Asset multiplier (mean) = 8.86 (median) = 4.60 range: 0.06–434.4 ave MDD = 71.6
NK: Nikkei225 N = 791244	1999.1–2016.4 4254days 183 stocks Hurst exponent (mean) = 0.48	Asset multiplier (mean) = 2.62 (median) = 1.62 range: 0.14–56.6 ave MDD = 78.7
SP_mod: SP corrected characteristics by correction operation	4359days 391stocks Hurst exponent (mean) = 0.54 <i>Ref to Sect. 4.2</i>	Asset multiplier (mean) = 44.4 (median) = 11.0 range: 0.024–6098.9 ave MDD = 70.7
NK_mod: NK corrected characteristics by correction operation	4254days 183 stocks Hurst exponent (mean) = 0.49 <i>Ref to Sect. 4.2</i>	Asset multiplier (mean) = 1.95 (median) = 1.21 range: 5.6×10^{-6}–27.9 ave MDD = 79.4

Looking at the difference of the average of Hurst’s exponent [10, 11], in the correction operation on the SP data, the mean regression changed to a “trend “ sustainability, but in the NK data, such a change did not occur. Additionally, in the dataset subjected to the correction operation, the asset multiplier range was large.

5.2 Parameter Setting

OLMAR and OLMAR_random method: transaction cost is 0.3% per transaction value (buy or sell).

MAD-MRC: eval_period = 40 days, From t 1 to t 2 the whole period is selected,
 check_pct = [0.25, 0.5, 1.0, 3.0], ngbhdChkN = 2000.
 ※ checkPct default is 1.0

5.3 Price Fluctuation Characteristics

The dataset used in this verification comprised a total of four types including the original stock price dataset 2, and the data correction operation is shown in Sect. 4.2.

Figure 1 shows SP stock data (N = 1,688,338) plotted by MAD and the next day’s return. The MAD and return of SP and NK are shown in Table 2. From the above, we found that for SP and NK data, a statistically significant high return is obtained if MAD is 1% or less.

In this research, we call this stock price fluctuation characteristic “**MAD-MRC**” (characteristic of mean regression phenomenon occurring below a one percent by moving average divergence rate). This characteristic is a short-term “Mean reverting phenomenon” existing in the stock market, which has not been clarified in previous research.

In this verification, a return curve was generated by averaging 2000 points in the neighborhood (Fig. 1) (Table 2). According to the **Brunner Munzel Test**, the null hypothesis that returns and is equal in all cases is rejected with risk factor $p < 0.001$. We used the Brunner Munzel test because we assumed that there was no normality and equal variance. In addition, in the test, NK and SP dataset of the MAD (40) compared the less than $N\% = [0.25, 0.5, 1, 3]\%$, and mean data of the MAD (40).

This verification revealed that MAD-MRC shows a statistically significant high return in the long-term datasets (SP and NK), indicating that it may be linked to the long-term superiority of the OLMAR method.

Table 2. Return on the next day (anomaly) when the lower N% of MAD is selected. The return tends to increase with **MAD-MRC**, and the mean return increases due to the expansion of the negative width of MAD. According to the **Brunner Munzel test**, the null hypothesis that returns and is equal in all cases is rejected with risk factor $p < 0.001$.

Select lower N%	MAD of data SP	Next day’s return *** $p < 0.001$
0.25	-33.38	1.0061***
0.5	-27.83	1.0031***
1.0	-22.32	1.0042***
3.0	-14.50	1.0017***
50.0	1.19	1.0012
Select lower N%	MAD of data NK	Next day’s return *** $p < 0.001$
0.25	-28.06	1.0072***
0.5	-24.39	1.0049***
1.0	-20.77	1.0041***
3.0	-15.17	1.0010***
50.0	0.36	0.9997

5.4 Data Correction Operation on Price Fluctuation Characteristics

Figure 2 shows the result of the decreasing effect of **MAD-MRC** (lower 1%: mean reverting characteristic) according to the proposed method in Sect. 4.2. After applying the correction operation to the SP data (Fig. 1 is original), there is no tendency for the return to rise according to **MAD**.

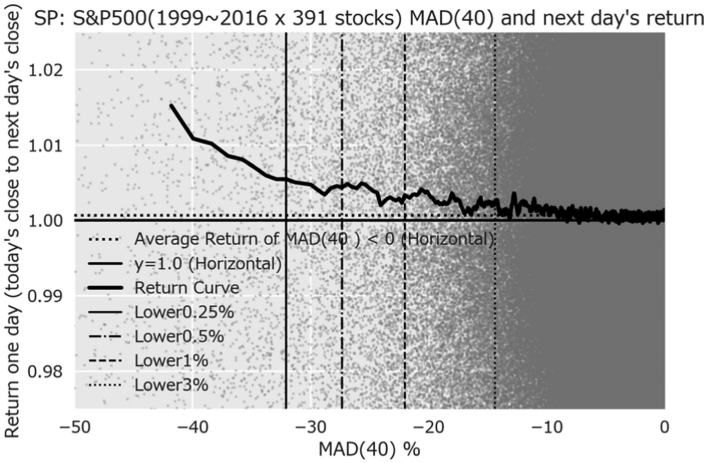


Fig. 1. Rising tendency (Anomaly) of the Return curve of MAD-MRC (40) using all stock data of SP: the x-axis shows the MAD (40) 40-day moving average divergence rate, and the gray point shows 1-day’s data (N = 1,688,338). The y-axis shows 1-day’s return from the closing price on the day to the closing price on the next day; sorting all the data in ascending order and plotting the average value of 2000 points in the vicinity is the **Return Curve. MAD-MRC (MAD-Mean reverting characteristics)** whereby returns rise occurs when the moving average divergence rate is less than or equal to a lower percentage; **the return becomes higher as the value of MAD gets smaller**). Each vertical line from the lower left shows 0.25%, 0.5%, 1% and 3% of the MAD.

The correction operation is performed in three representative periods $k = [20, 40, 80]$ in Sect. 4.2. Performing a typical period correction was also effective for the moving average, such as the period (e.g., 60, 100 days) (Fig. 3).

However, corrections in a short period (20 days) are inaccurate and need improvement in the future.

This result focused on the rare situation where MAD-MRC occurred and showed that excess return could be obtained by holding the relevant stock for a certain period.

5.5 Comparison of Simulation Results and “Index” by the OLMAR Method and OLMAR Random Method

From the results in Fig. 4, The index AR was 13.4% in 18 years of data, for the same period the OLMAR method asset AR was 35.4%. The maximum drawdown (MDD) increased from 49.5% to 98.28%, and the asset price is temporarily less than 1/50. This confirmed the high drawdown problem described in the paper [2].

Simulation results using SP (including anomalies) and SP_mod (anomalies removed) are shown in Figs. 4 and 5. Figure 4 shows the superiority of the OLMAR method, indicating that the OLMAR random method has no advantage. Figure 5 shows that there is no superiority of the OLMAR method in SP_mod data in which MAD-MRC is corrected.

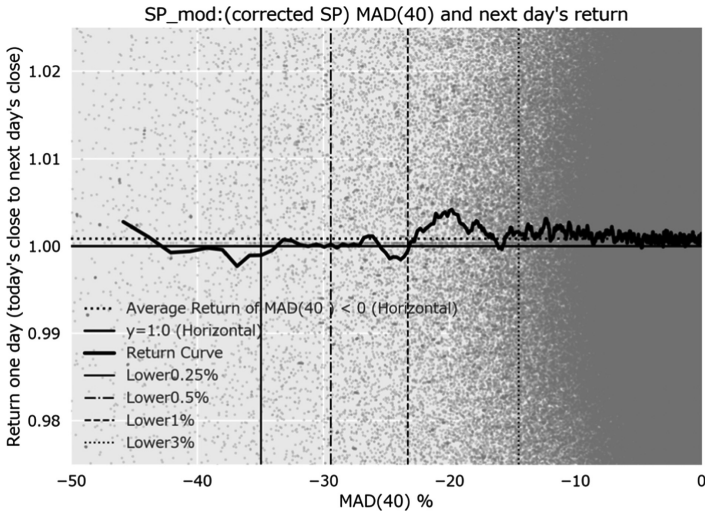


Fig. 2. The **Return curve** in the **SP_mod (anomalies removed)**: comprises data after correction operation to SP (N = 1,688,338). After applying the correction operation to the data SP (Fig. 1), there is **no tendency for the return to rise** according to the magnitude of the minus value of MAD.

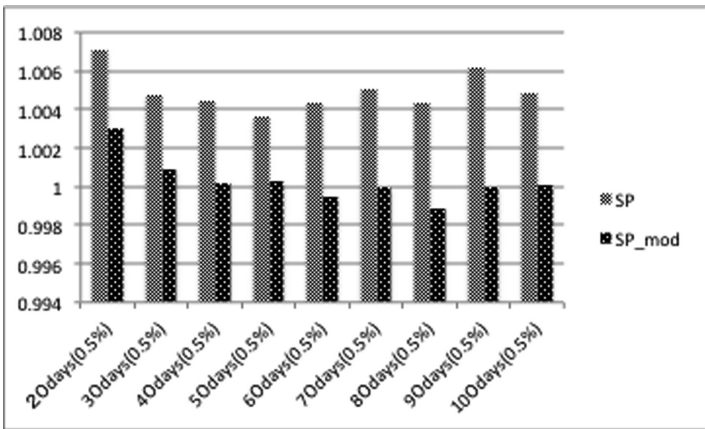


Fig. 3. Decrease of anomalies in “SP_mod” **MAD-MRC** after SP (including Anomalies)” **correction operation**. We verified the return of MAD (k, m parameters). During period $k = [20, 40, 80]$, the correction operation was performed and $m = [20, 30, \dots, 100]$ was a verification period. By performing the correction operation for the periods of set k , **MAD-MRC** in the set m period including other periods was also corrected.

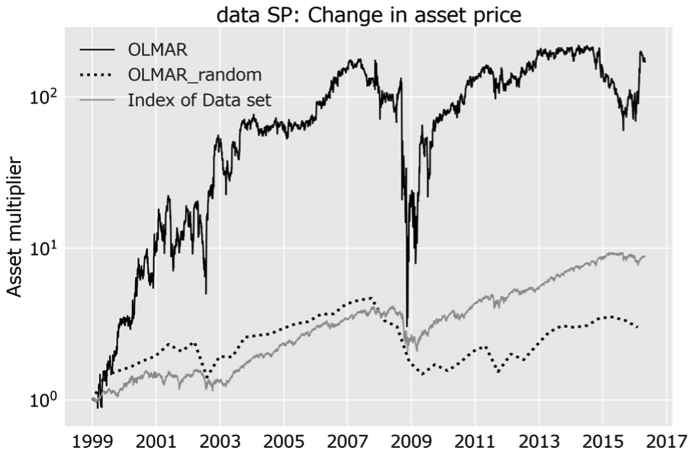


Fig. 4. Original Data, including anomalies, for **SP**: change in asset price of the **OLMAR** method and **OLMAR_random** (invalidation of deviation from the moving average) method and “**Index of dataset**”. For this result of using the moving average of 40 days at a transaction cost of 0.3% (one way of trading value), the OLMAR method outperforms the index and the asset multiplier is about 170 times. However, because OLMAR_random needs a transaction cost, it underperforms the index. The index final asset multiplier is 8.86 times.

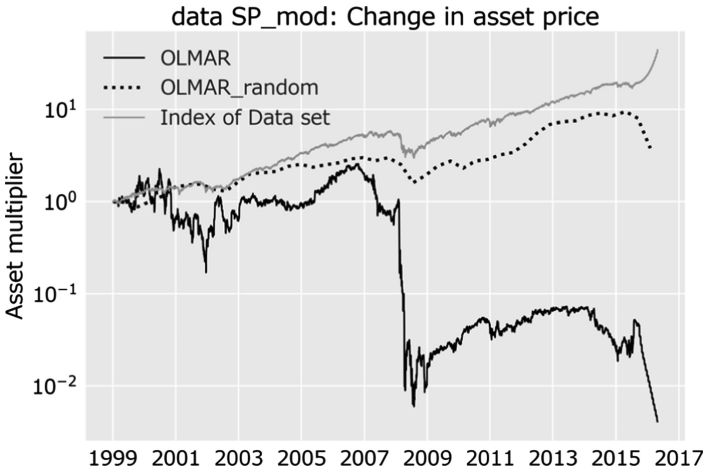


Fig. 5. Data in which **MAD-MRC** is corrected; **SP_mod** (anomalies removed): change in asset price of the **OLMAR** method and **OLMAR_random** (invalidation of deviation from the moving average) method and the “**Index of the dataset**”. Simulation is performed under the same parameters as shown in Fig. 4. Although the asset has risen to more than **40 times**, the “**Index of the dataset**”, the assets of the OLMAR method have declined sharply (the asset multiplier is 0.004: 1/250). The superiority of the **OLMAR** method has been lost due to the **corrected of MAD-MRC**.

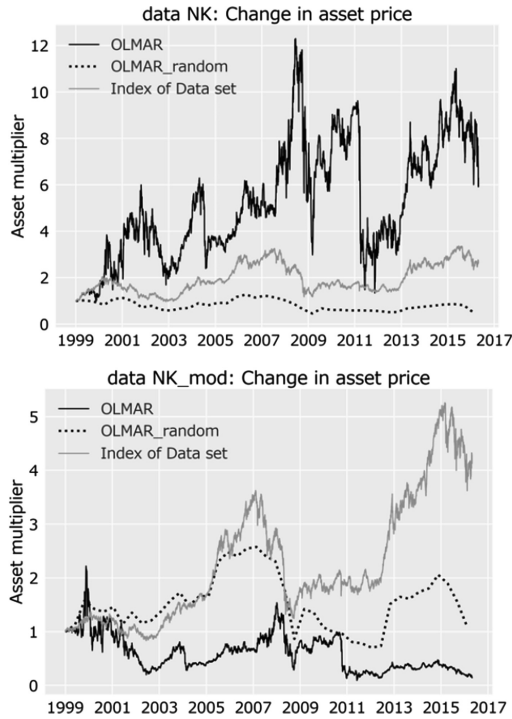


Fig. 6. Upper: using NK data (including anomalies). **Lower:** using NK_mod data (anomalies removed; corrected MAD-MRC). Both cases are simulation results with a moving average of 110 days and a transaction cost of 0.3%. In the NK data, the OLMAR method has an advantage, and in the NK_mod data, the superiority disappears. OLMAR method’s asset whereby NK and NK_mod are compared: AM is 6.29 to 0.15. MDD is 89.13% to 95.9%.

Similar results were obtained for NK (including anomalies) and NK_mod (anomalies removed) data (Fig. 6). The performance of the OLMAR method in the NK data is smaller than that of the SP data, but the asset price is over two times the index. In NK_mod, the OLMAR method fell the most, and underperformed the index and OLMAR_random method. The results are shown in Table 3.

From the results of this verification, we found the following:

- (1) With the new Japan and US dataset, the **OLMAR method could be confirmed to outperform the index** even when considering the transaction cost. Additionally, its advantage was lost due to correcting the **MAD-MRC**.
- (2) The OLMAR method uses **MAD-MRC** (small according to the moving average divergence rate being lower than 1%, with **return tending to enlarge**: Mean Reverting Characteristic). This is **short-term mean reverting** in the stock market, the phenomenon has not been clarified in past studies.

- (3) We found that the performance and long-term superiority of the **OLMAR method are dependent on the MAD-MRC**. Because the OLMAR method is a technique using **very limited characteristics**, manipulating its properties results in a completely different performance.
- (4) The OLMAR method has a high maximum drawdown risk even when it shows a high return. In the data SP simulation, **asset prices temporarily declined to 1/50**, and it was reconfirmed that OLMAR is a method with a high drawdown risk [2].

Table 3. Simulation result of the asset multiplier, annual average return and Max Drawdown. OLMAR has superiority in the original data for SP (including anomalies) and NK (including anomalies). However, there is no advantage in the data for SP_mod (anomalies removed) and NK_mod (anomalies removed). Additionally, the OLMAR method always has a high risk of Max Drawdown. The OLMAR_random method shows the average and standard deviation of 100 simulation results because it is a method of randomly selecting stocks.

Data and method	Asset multiplier <i>Higher is better</i>	Annual average return <i>Higher is better</i>	Max draw down <i>Lower is better</i>
SP: OLMAR	170.08	34.57%	98.28%
SP: Index	8.86	13.44%	49.56%
SP:	3.69(Ave)	5.65(Ave)	68.46(Ave)
OLMAR_random	3.70 (SD)	5.33 (SD)	9.98 (SD)
SP_mod: OLMAR	0.004	-27.22%	99.84%
SP_mod: Index	44.47	24.53%	49.80%
SP_mod:	5.33(Ave)	7.79(Ave)	70.04(Ave)
OLMAR_random	5.71 (SD)	5.54 (SD)	8.37 (SD)
NK: OLMAR	6.29	11.51%	89.13%
NK: Index	2.62	5.88%	63.97%
NK:	0.94(Ave)	-2.54(Ave)	81.9(Ave)
OLMAR_random	1.06 (SD)	5.06 (SD)	8.51 (SD)
NK_mod: OLMAR	0.15	-10.62%	95.87%
NK_mod: Index	4.12	8.76%	65.71%
NK_mod:	0.65(Ave)	-4.2(Ave)	81.1(Ave)
OLMAR_random	0.61 (SD)	4.53 (SD)	7.81 (SD)

5.6 Evaluation by MRScore

In this section, the Mean Reverting trend of SP and NK was evaluated by MRScore.

The results of the MRScore are shown in Table 4. The size of the MRScore has an influence on the asset multiplier of OLMAR method. The advantage of this indicator is that easily estimates the magnitude of the Mean Reverting trend without using the OLMAR method or any other PSA.

The MRScore of datasets SP and NK shows a Mean Reverting trend when the stock price declines. Furthermore, the MRScore shows that the Mean Reverting trend for SP is larger than that of NK. However, the relationship between the MRScore difference and asset multiplier difference are the future research topic.

Table 4. Evaluation of the stock price dataset according to MRScore: both the SP and NK datasets show the Mean Reverting trend after a price fall, because both MRScores are greater than 0. However, SP tends to be stronger than NK, which has an exponential influence on asset prices, so the difference is large.

Dataset name	Mean reverting score	Asset multiplier of OLMAR method
SP	3.90	170.08
NK	0.45	6.29

6 Summary

In this research, we have clarified anomalies that have not been previously clarified and examined their existence. Moreover, the phenomena were separated by a simple filter, the occurrence frequency was small, and what has been treated as noise or as an abnormal value until now was shown to be an anomaly used in the OLMAR method.

In this study, we showed that the OLMAR method gains excess return based on some robust price fluctuation characteristics of MAD-MRC. And, the long-term superiority of the OLMAR method is that it efficiently selects MAD-MRC. MAD-MRC is a phenomenon newly shown by this research, with a fluctuation occurring at a frequency of 1% or less on the stock market, and the return tending to increase according to the negative divergence width of MAD. MAD-MRC is a short-term mean reverting characteristic of the stock market that has not been clarified in previous studies.

Generally, construction of a portfolio based on using the main nature of stock price movements, but with OLMAR as a distinctive superiority, was established only by phenomena that very rarely occur (generally called an outlier). Therefore, we confirmed that the superiority of the OLMAR method is lost in stock data corrected by applying the MAD-MRC proposed in this paper.

However, actively using these phenomena is greatly damaged by phenomena change. It is necessary to think about how to balance these problems and refine the algorithm. Furthermore, the magnitude of the MDD risk of the OLMAR method represents the “Instability of the Anomalies”, and a more detailed analysis is needed. Instability means that “Anomalies do not always exist”.

In this study, we showed the existence of an average regression trend in SP and NK data using MRScore, which is the evaluation index of an average regression, and it was possible to grasp the size difference.

A future task is to increase the accuracy of the correction operation and to analyze the relationship between the moving average divergence and the return in more detail. Additionally, both SP and NK data used in this study have survivor bias. Because the excess return is obtained by the OLMAR method, there is a possibility that it may be influenced by survivor bias, so it is necessary to verify this in the future.

Appendix 1 Algorithm 1: MAD-MRC and Return Curve Algorithm

Algorithm 1: MAD-MRC and ReturnCurve

```

1: Input:
    $P_t^i$ : Daily stock price data of stock_id  $i$ , time(day_id)  $t$ ,  $1 \leq i \leq M$ ,  $1 \leq t \leq N$ ;
   eval_period: evaluation period (days);
   t1: evaluation start day_id;
   t2: evaluation end day_id;
   checkPct: Select daily return of stocks MAD(Moving Average Divergence rate)
           value less than or equal to the low order checkPct % in the selected period;
   ngbhdChkN: How many data (one side) of neighborhood data are calculated to average?;
2: Output: ReturnCurveX: vector of (All stock's daily MAD values sorted in ascending order).
           ReturnCurveY: vector of (Corresponds to ReturnCurveX:
           Next day's neighborhood average daily return).
           v: next day's return of the neighborhood average value
3: Procedure:
4: MAD[N][M]: allocate and initialize zero for MAD data;
5: for  $i \leftarrow 1$  to  $M$  do:
6:   for  $t \leftarrow t1 + \text{eval\_period}$  to  $t2$  do:
7:     MAD[t][i] ← calculate MAD( $p^i$ , eval_period,  $t$ ) # refer § 3.1 MAD definition
8:   end for
9: end for
10: MADV: List format data initialize;
11: MADV ← sort by Ascending order MAD value of MAD[:,i];
12: size ←  $N \times M$ 
13: ReturnCurveX[size]: allocate and initialize zero (ReturnCurve X data: MAD value);
14: ReturnCurveY[size]: allocate and initialize zero (ReturnCurve Y data: Next Day's return );
15: for  $k \leftarrow 1$  to size do:
16:   ReturnCurveX[k] ← MADV[k]
17: end for
18: for  $k \leftarrow 1$  to size do:
19:   if ngbhdChkN  $\leq k$  and  $k \leq \text{size} - \text{ngbhdChkN}$  then:
20:     ReturnCurveY[k] ← Calculate mean MADV[ $k - \text{ngbhdChkN} : k + \text{ngbhdChkN}$ ]
21:   elif  $k < \text{ngbhdChkN}$  then:
22:     ReturnCurveY[k] ← Calculate mean MADV[ $1 : 2 \times k$ ]
23:   elif  $k > \text{size} - \text{ngbhdChkN}$  then:
24:     npoints ← size - ngbhdChkN
25:     ReturnCurveY[k] ← Calculate mean MADV[ $\text{size} - 2 \times \text{npoints} : \text{size}$ ]
26:   endif
27: end for
28: idx ← convert_integer(size × checkPct × 0.01)
29: v ← ReturnCurveY[idx]
30: return ReturnCurveX, ReturnCurveY, v

```

Appendix 2 Algorithm 2: OLMAR_Random Method

The OLMAR_random method, once simulated by the OLMAR method, extracts the composition weight of the portfolio and randomly replaces the selected stocks. In addition, a sequence of weights greater than 0 is recognized as one chunk and is copied to new randomly selected stock_id. The OLMAR_random method generates a portfolio based on the new weight data generated by Algorithm 2.

Algorithm2 OLMAR_random generate weight of portfolio

```

1: Input:
   W[n_days][n_stocks]: n_days is the period of data and n_stocks is
   the number of stocks. From the calculation result of OLMAR method,
   extract weight from portfolio of each stock.
2: Output: W2[n_days][n_stocks]: OLMAR_random portfolio weight data.
3: Procedure:
4: W2[n_days][n_stocks] : allocate and initialize zero;
5: for  $k \leftarrow 1$  to n_stocks do:
6:   period_lst  $\leftarrow$  retrieve a period pair ( $p1, p2$ ) list of the when consecutive weights
   are greater than 0 data for each stock.
7:   for ( $p1, p2$ ) in period_lst do:
8:     flag  $\leftarrow$  True
9:     while flag==True do:
10:       $k2 \leftarrow$  randomly select stock id between 1 and n_stocks;
11:      if is_zero(W2[ $p1:p2$ ][ $k2$ ]) then
12:        W2[ $p1:p2$ ][ $k2$ ] = W[ $p1:p2$ ][ $k$ ]
        # Select stock randomly and write data in cases
        # where no data exists during that period.
13:      flag  $\leftarrow$  False
14:      end if
15:     end for
16: end for

```

References

1. Li, B., Hoi, S.C.H.: On-line portfolio selection with moving average reversion. arXiv preprint [arXiv:1206.4626](https://arxiv.org/abs/1206.4626) (2012)
2. Li, B., et al.: Moving average reversion strategy for on-line portfolio selection. *Artif. Intell.* **222**, 104–123 (2015)
3. Li, B., Hoi, S.C.: Online portfolio selection: a survey. *ACM Comput. Surv. (CSUR)* **46**(3), 35 (2014)
4. Li, B., Sahoo, D., Hoi, S.C.: OLPS: a toolbox for on-line portfolio selection. *J. Mach. Learn. Res.* **17**(35), 1–5 (2016)
5. Nyikosa, F.M., Osborne, M.A., Roberts, S.A.: Adaptive Bayesian optimisation for online portfolio selection. In: *Workshop on Bayesian Optimization at NIPS*, vol. 2015 (2015)
6. Ha, Y.: Online portfolio selection with transaction costs including market impact costs. *Browser Download This Paper* (2016)
7. Gao, L., Zhang, W.: Weighted moving average passive aggressive algorithm for online portfolio selection. In: *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1. IEEE (2013)
8. Lin, X., Zhang, M., Zhang, Y., Gu, Z., Liu, Y., Ma, S.: Boosting moving average reversion strategy for online portfolio selection: a meta-learning approach. In: *Candan, S., Chen, L., Pedersen, T.B., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS*, vol. 10178, pp. 494–510. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55699-4_30
9. Kroll, Y., Levy, H., Markowitz, H.M.: Mean-variance versus direct utility maximization. *J. Financ.* **39**(1), 47–61 (1984)
10. Rasheed, K., Qian, B.: Hurst exponent and financial market predictability. In: *IATED Conference on Financial Engineering and Applications (FEA 2004)* (2004)

11. Carbone, A., Castelli, G., Stanley, H.E.: Time-dependent Hurst exponent in financial time series. *Phys. A: Stat. Mech. Appl.* **344**(1), 267–271 (2004)
12. Li, B., Hoi, S.C., Gopalkrishnan, V.: CORN: correlation-driven nonparametric learning approach for portfolio selection. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 21 (2011)
13. Agarwal, A., et al.: Algorithms for portfolio management based on the newton method. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM (2006)
14. Fama, E.F.: Market efficiency, long-term returns, and behavioral finance. *J. Financ. Econ.* **49**(3), 283–306 (1998)
15. Jegadeesh, N., Titman, S.: Profitability of momentum strategies: an evaluation of alternative explanations. *J. Financ.* **56**(2), 699–720 (2001)



A Study on Technology Structure Clustering Through the Analyses of Patent Classification Codes with Link Mining

Masashi Shibata^(✉) and Masakazu Takahashi

Graduate School of Sciences and Technology for Innovation,
Yamaguchi University, Yamaguchi, Japan
{g501wc, masakazu}@yamaguchi-u.ac.jp

Abstract. This paper provides the technology structure analysis and the technology field clustering through the analyses of patent classification codes with link mining method. Knowledge extraction from patent information has been made thus far, but conventional patent analysis methods depend on personal heuristic knowledge. It makes it hard to extract the technology structure. We are focusing on classification codes in the patent. They are assigned to capture the technology fields of patent. With the proposed method, we are succeeding in the clustering of various technology fields.

Keywords: Patent · Patent analyses · Machine learning · Link mining Clustering · Technology structure analyses

1 Introduction

In corporate activities, it is important to find problems and solutions to perform new product development. Finding completely new solution in a field, often cause innovations. The technology fields which have a similar technology structure often have the same problems and the solutions. Thus, exploring the similar technology fields are an effective way to find the solutions. There are many methods to tackle them, such as patent analysis, paper survey, and so on. Among the analysis methods, the patent analysis is widely used in the business sector. This is because the information source is open to the public. Therefore, it is easy to obtain the information through the Internet. Furthermore, the patent information has the following characteristics; (a) Due to the patent submission is originally for industrialization basis, easy to understand the technological tastes, (b) It has rich information on not only the body but the classification codes.

Thus, the patent information is one of the useful data sources in order for analyzing the technological structure. However, the analytical method focusing on the patent document has some challenges. It is difficult to apply common text mining method for the unique wording and the long sentences. Since the chart has an important meaning, only the document information is not enough to extract the content sufficiently. Therefore, the patent analysis is performed manually, and its results are heuristic and depend on analyst's skills and experiences.

The Link mining method is one of the techniques to visualize the relationship structure of things. This method is mainly used for the relationship analysis for networks such as web pages, citation, gene networks.

In this paper, we propose the method of finding similar technology fields that does not depend on skills and experiences. For analyzing the technology structure, we focus on the classification codes in patents. The technology relations are visualized by the graph made by the classification codes. Then, the features of the graph are calculated by the link mining method. With the proposed method, we succeeded in visualizing the technological structure and clustering the technology fields that have similar structure.

The rest of the paper is organized as follows: Sect. 2 discusses the backgrounds of the research and related work; Sect. 3 briefly summarizes the gathered data on the target patent sector; Sect. 4 describes the analytics of the data and presents analytical results; and Sect. 5 gives some concluding remarks and future work.

2 Related Work

Technology analysis using patent information is utilized for multiple purpose, and it is performed by many methods. For finding business solutions, TRIZ (Teoriya Resheniya Izobretatelskikh Zadatch) aims at performing technical development based on the structure of problem solving which appears repeatedly in the patent [1, 2]. Kawakami et al. propose the inconvenience idea gains support system aid of inventive problem solving theory TRIZ [3]. For revealing the companies' activities, patent citation is often used. Narin et al. reveal the companies' strength by combining patent citation and other indicators [4]. Shide et al. performed finding the change of the positioning for customer of research and development activities of the company using the patents analyses [5]. Kimura proposed a technology evaluation method based on patent analysis for technology strategy planning [6]. Nagaoka et al. reveal the process of innovation in Japan by performing the survey about patent to inventors [7]. Many measures have been taken as having centered on the patent for a business solution.

Next, we look down about the patent analytical skills. One of the techniques of patent information analysis is the patent mapping. Kiriya made content analyses with this method [8]. As for the analysis of important information derived from patent analysis, Carpenter analyzed for important cited patents. Muguruma showed the validity of the patent citation analysis to propose the FCA (Forward Citation Applicant) map [9, 10]. Sato et al., proposed the importance calculation method of the patent document based on the citation information [11]. Ogawa et al., proposed a basic patent extraction based on the citation information [12]. For citations, Albert conducted a validation of citation for important patent among the industry [13].

In place of that which were conventionally performed by human power, as to the classification of the patent, such as category of invention and problem, Tanaka proposed method of extracting the feature automatically [14]. Yamashita proposed a method of surveillance technology and specific method of patent classification with text mining [15]. Yamamoto et al., proposed a method to enhance the compatibility of the search by applying the information of related patent documents in search of academic papers. Yamamoto, proposed a method to find the scientific papers with a variety of

further information [16, 17]. Kleinberg extracted the topic and description of the relationship with graph theory [18]. Eto proposed a measure of co-citation based on structural units of the paper [19]. Ueda proposed the technical analysis with an active mining method that focuses on the cognitive processes of the patent examiner, utilizing patent classification such as IPC (International Patent Classification), FI (File Index), and F-term (File Forming Term) [20].

Thus, with the application of the technology of intelligent informatics, knowledge extraction is performed to patent information. Then, we look down about the mining technology, which is the one technique of knowledge extraction [21–23]. The relationship of analytical methods and technologies, technology analysis method using patent information is described.

Above all, structured technique using graph theory has been applied in various fields [24]. For example, chemical formula, WWW, social network, statements with grammatical structure and dependency. For patent analysis, it is used for representing the relationship of citations. However, the research to express the technological structure by the graph of the classification codes has not been available so far. Thus, we tackle the analysis of the patent by using this method in this paper.

3 Experimental Configuration

In general, patent information consists of metadata, such as application date, applicant, classification codes, literal information and graphic information. In addition to the IPC and CPC, FI and F-term are used as classification codes in the Japanese patent classification system. Both IPC and CPC indicate the technological fields of the main topics of the patent's claims. FI is used to indicate patent's technological fields more detail than IPC by adding extension symbol and/or file discrimination symbol to the IPC. The F-term is used to indicate patent's technological fields more detail than FI by adding the viewpoint of the problem and the solution. The F-term is separated to a theme code and a viewpoint [25]. Figure 1 shows an example of the F-term's notation. The theme code represents the technological field. The view point analyses the theme, such as material, purpose, operation, and manufacturing. The figure subdivides the viewpoint. A theme code is composed of 1 digit number, 1 alphabetic character, and 3 digit number. The first digit number takes the value of 2, 3, 4, and 5, and they represent residual technology, mechanics, chemistry, and electricity respectively. The first 2 letters of a theme code form a theme group that groups similar technological fields.

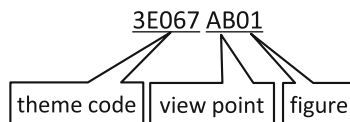


Fig. 1. F-term's notation

We have regarded 38 categories classified as theme groups as technological fields. We have selected one theme code for each theme group as the subject of the analysis. Table 1 shows the excerpted list of selected theme codes and their description. Since F-term is reviewed annually, the theme code of a certain technology field may change during the experiment period. Thus, some theme groups have more than one theme code. We have gathered the target data from the patent information database [26].

We have focused on patent publication. The reason is that it is ensured inventive step and inquiry by the examination. And we have gathered the data which are submitted within 2007–2009. This is because that we also analyze whether the global financial crisis, which occurred in 2008 and had a big negative social impact, influenced the technology structure or not. In above conditions, 26,705 patent data are extracted from the database.

In this study, the technological structure of a category is defined by the contained technological fields and their relations. It is represented by a graph structure. The theme codes given to a patent represent the technology fields and are used as nodes of the graph. First of all, for representing the technological structure of a patent, a complete graph is created for each patent. The graph is an unweighted undirected graph.

Then, the annual graph for each theme group is created by overlaying the patents' complete graphs in each year of application. The number of occurrences of nodes and the number of occurrences of links is counted one, thus the graphs are also unweighted undirected graphs. As a result, 117 graphs are created. They represent the annual technology structure for each technological field.

Then, the feature vector for each graph is created. As the elements of the feature vector, 3 types of features of the graph, such as (1) mean distance, (2) density, and (3) transitivity and the number of grants of the theme codes with the first digits 2, 3, 4, and 5 are calculated. Mean distance represents the mean shortest path between any two nodes in a graph. Density represents the ratio of the number of actual links to the number of the all links that can be existence in the graph. Transitivity indicates whether there is a link between node- i and node- k if there is a link between node- i and node- j and also between node- j and node- k . As a result, the feature vector of each graph is seven dimensional. Table 2 shows the excerpted list of the feature vector of the graphs.

x-means method [27, 28] is used as the classifier. x-means is the extended k-means method, and it can estimate the number of clusters. It hierarchically creates two clusters by k-means until the sum of the variances becomes minimal. For searching the minimal point, both the cluster members and the cluster center are varied. These formulae are given by (1) and (2) respectively [29]. Calculation (1) and (2) are repeated until convergence. Where $x^{(i)}$ is the sample- i , $C^{(i)}$ is the cluster of sample- i , μ_j is the center of the cluster- j , and m is the number of the samples. In this experiment, Python and its library, scikit-learn [30] are used to execute x-means.

$$C^{(i)} := \underset{j}{\operatorname{argmin}} \|x^{(i)} - \mu_j\|^2 \quad (1)$$

$$\mu_j := \frac{\sum_{i=1}^m \{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}} \quad (2)$$

Table 1. List of the theme codes (excerpted)

Theme group			Theme code	
Description			Description	
1	2B	Natural Resources	2B022	Cultivation of plants
2	2B	Natural Resources	2B104	Farming of fish and shellfish
3	2C	Office Machine	2C001	Electronic game machines
4	2D	Computerized Game, Study	2D001	Devices affording protection of roads or walls for sound insulation
5	2E	Living Environment	2E013	Bay windows, entrances, and structural adjustments related thereto
6	2F	Measurement	2F003, 2F103	Optical transform
7	2G	Applied Physics	2G001	Analyzing materials by the use of radiation
8	2H	Applied Optics	2H006	Eyeglasses
9	2K	Light-Device	2K103, 2K203	Projection apparatus
10	3B	Textile-Packing Machine	3B006	Electric suction cleaners
11	3C	Production Machinery	3C001	Automatic control of machine tools
12	3D	Transportation	3D001, 3D301	Chassis suspension devices
13	3E	Medical Equipment	3E001, 3E040	Handling of coins
14	3F	Conveyance	3F001, 3F343	Sheets, magazines, and separation thereof
15	3G	Power Machinery	3G001, 3G301	Electrical control of air or fuel supplied to internal-combustion engine
16	3H	Automated Control	3H001	Servomotors
17	3J	General Machinery	3J001	Connection of plates
18	3K	Living Equipment	3K001, 3K092	Resistance heating
19	3L	Heat Machinery	3L015, 3L018, 3L345	Cooling media circulation in refrigerators
20	4B	Biotechnology	4B004	Coffee makers
21	4C	Medical Care	4C001, 4C301, 4C601	Ultrasonic diagnostic devices
22	4D	Environmental Chemistry	4D001, 4D006	Separation using semi-permeable membranes
23	4E	Material Processing	4E001, 4E078, 4E079	Arc welding in general

4 Experimental Results

This chapter describes the result of the clustering of the technological structure of the target data using link mining.

Table 3 describes the result of the classification. The 117 groups are classified into 4 groups. In the theme groups assigned by multiple theme codes, only representative

theme code is described. Table 4 describes the coordinates of the center of the clusters. Most of the graphs of the same theme group are clustered into the same group even though the year of application is different. It is confirmed that the global financial crisis did not influence the technological structure.

By the result, Cluster0 is the cluster related to mechanics, cluster1 is the largest cluster, which relates every field, cluster2 is the cluster related to residual technology and electricity, and cluster3 is the cluster related to chemistry.

In cluster0, the theme groups which have only head value 3 are grouped. Many theme codes with head value 3 are allocated. Its mean distance is long, and density is low. Thus it is related to mechanics and made by a large amount of patents each covers a small area.

In cluster1, More than half of the theme groups are aggregated. The theme codes with the head value 2, 3, 4, and 5 are evenly assigned, but the number of them is small. Its mean distance is the shortest, and density and transitivity are the highest. Thus it covers every field and made by a large amount of patents each also covers a large area.

Table 2. The list of the feature vector (excerpted)

Year	Themecodes	2	3	4	5	Distance	Density	Transitivity
2007	2B022	18	7	34	0	1.8930	0.1070	0.2989
2007	2C001	15	4	2	65	1.9321	0.0679	0.1913
2007	2D001	19	2	7	4	1.8891	0.1109	0.1939
2007	2E013	2	0	0	0	1.0000	1.0000	1.0000
2007	2G001	33	6	19	17	1.9333	0.0667	0.1747
2007	2H006	21	8	48	3	1.9206	0.0794	0.2140
2007	2K103, 2K203	46	11	3	43	1.9098	0.1081	0.3543
2007	3B006	0	6	2	7	1.7143	0.2857	0.4233
2007	3C001	0	17	0	1	1.7320	0.2680	0.4176
2007	3F001, 3F343	15	28	3	7	1.8723	0.1277	0.3279
2007	3H001	0	5	0	0	1.3000	0.7000	0.8000
2007	3J001	16	44	10	14	1.9498	0.0502	0.0959
2007	3K001, 3K092	4	16	23	16	1.9234	0.0766	0.1594
2007	3L015, 3L018, 3L345	1	14	19	3	1.8889	0.1111	0.1880
2007	4B004	0	15	16	0	1.8430	0.1570	0.2545
2007	4C001, 4C301, 4C601	11	0	24	20	1.8774	0.1226	0.3027
2007	4D001, 4D006	17	22	129	13	1.9583	0.0417	0.1552
2007	4E001, 4E078, 4E079	2	8	20	0	1.8345	0.1655	0.3139
2007	4F040, 4F041, 4F042	58	42	60	40	2.1048	0.0440	0.1838
2007	4L031	12	69	67	5	1.9103	0.0897	0.2461
2007	4M118	44	67	18	59	1.9188	0.0812	0.3164
2007	5C001	10	18	4	13	1.7540	0.2460	0.4491
2007	5D002	0	4	0	4	1.0000	1.0000	1.0000
2007	5M024, 5B024	5	11	1	37	1.8224	0.1776	0.3811

Table 3. The result of the clustering

Cluster no.	Year	Theme	Year	Theme	Year	Theme	Year	Theme	Year	Theme	Year	Theme
0	2007	3D001	2007	3E001	2007	3F001	2007	3G001	2007	3J001		
	2008	3D001	2008	3E001	2008	3F001	2008	3G001	2008	3J001		
	2009	3D001	2009	3E001	2009	3F001	2009	3G001	2009	3J001		
1	2007	2B022	2007	2B104	2007	2D001	2007	2E013	2007	2F003	2007	2G001
	2008	2B022	2008	2B104	2008	2D001	2008	2E013	2008	2F003		
	2009	2B022	2009	2B104	2009	2D001	2009	2E013	2009	2F003		
	2007	3B006	2007	3C001	2007	3H001	2007	3K001	2007	3L015	2007	4B004
	2008	3B006	2008	3C001	2008	3H001	2008	3K001	2008	3L015	2008	4B004
	2009	3B006	2009	3C001	2009	3H001	2009	3K001	2009	3L015	2009	4B004
	2007	4C001	2007	4E001	2007	4G001	2007	4H001	2007	5C001	2007	5D002
	2008	4C001	2008	4E001	2008	4G001	2008	4H001	2008	5C001	2008	5D002
	2009	4C001	2009	4E001	2009	4G001	2009	4H001	2009	5C001	2009	5D002
	2007	5E001			2007	5G001	2007	5H001	2007	5J001		
	2008	5E001			2008	5G001	2008	5H001	2008	5J001		
	2009	5E001	2009	5F001	2009	5G001	2009	5H001	2009	5J001	2009	5M024
2	2007	2C001			2007	2K103	2007	4M118	2007	5F001	2007	5L096
	2008	2C001	2008	2G001	2008	2K103	2008	4M118	2008	5F001	2008	5L096
	2009	2C001	2009	2G001	2009	2K103	2009	4M118			2009	5L096
	2007	5M024										
	2008	5M024										
3	2007	2H006	2007	4D001	2007	4F040	2007	4J029	2007	4K001	2007	4L031
	2008	2H006	2008	4D001	2008	4F040	2008	4J029	2008	4K001	2008	4L031
	2009	2H006	2009	4D001	2009	4F040	2009	4J029	2009	4K001	2009	4L031

In cluster2, it consists of the patents whose theme codes' head value are 2, 4, and 5. There are relatively many theme codes which have head value 2 and 5 compared to the other elements. Its density is the lowest. Thus it is related to residual technology and electricity and made by a large amount of patents each covers small areas.

In cluster3, it consists of the patents whose theme codes' head value are 2 and 4. There are relatively many theme codes which have head value 4 compared to other elements, and its mean distance is the longest, and density and transitivity are the lowest. Thus it is related to chemistry, and made by a large amount of patents each covers a small area.

Table 4. Cluster center coordinates

Cluster no.	2	3	4	5	Distance	Density	Transitivity
0	9.667	39.733	5.400	14.133	1.912	0.109	0.274
1	6.879	6.121	12.303	9.697	1.746	0.252	0.373
2	27.167	6.667	9.556	47.722	1.894	0.107	0.284
3	22.222	18.778	73.056	15.278	1.964	0.064	0.203

The number of elements of cluster1 is the largest, and 66 theme groups are classified into the group. Different technological areas are classified in the same group. It is considered that they have a similar technological structure.

5 Concluding Remarks

This paper has presented the clustering of 38 technological fields which are represented as theme groups. As a result, it was found that some different technological fields have similar technological structures. In addition, it was found that the global financial crisis did not influence the technological structure.

Our future work is improving the clustering accuracy by optimizing the graphs' structures, extracting effective feature vector, and using appropriate classify method.

References

1. Altshuller, G.: *The Innovation Algorithm: TRIZ, Systematic Innovation, and Technical Creativity*. Technical Innovation Center, Worcester (1999)
2. Altshuller, G.: *40 Principles: Extended Edition*. Technical Innovation Center, Worcester (2005)
3. Kawakami, H., et al.: Idea generation support system for implementing benefit of inconvenience by employing the theory of inventive problem solving. *Trans. Soc. Instrum. Control Eng.* **49**(10), 911–917 (2013)
4. Narin, F., et al.: Patents as indicators of corporate technological strength. *Res. Policy* **16**(2–4), 143–155 (1987)
5. Shide, K., et al.: The shift of positioning of Japanese general contractors' R&D activities: a comparative value network analysis between condominium and semiconductor factory building markets. *J. Archit. Plan.* **76**(668), 1929–1935 (2011)
6. Kimura, H.: One approach of technology stocktaking and evaluation for corporate technology strategies: emphasizing future intentions and quantification through patent analysis. *J. Sci. Policy Res. Manag.* **26**(1/2), 52–61 (2012)
7. Nagaoka, S., et al.: The process of innovation in Japan seen by inventors, RIETI Discussion Paper Series (2007)
8. Kiriya, T.: IP information analysis (<special feature> patent information: analysis and effective utilization). *J. Inf. Sci. Technol. Assoc.* **60**(8), 306–312 (2010)
9. Carpenter, M.P.: Citation rated to technologically important patents. *World Patent Inf.* **4**, 160–163 (1981)
10. Muguruma, M.: The usefulness of patent forward citation analysis and its practical examples. *J. Inf. Sci. Technol. Assoc.* **56**(3), 114–119 (2006)
11. Sato, Y., et al.: A study of patent document score based on citation analysis. *IPSJ SIG Notes, Inf. Process. Soc. Jpn.* **59**, 9–16 (2006)
12. Ogawa, T., et al.: Finding basic patents using patent citations. *IPSJ SIG Notes, Inf. Process. Soc. Jpn.* **35**, 41–48 (2005)
13. Albert, M.B.: Direct validation of citation counts as indicators of industrially important patents. *Res. Policy* **20**, 251–259 (1991)
14. Tanaka, K.: Multi-viewpoint clustering of patent documents. *IPSJ SIG Notes* **4**, 9–14 (2008)

15. Yamashita, Y.: Text mining technology for patent analysis and patent search: patent search and patent analysis service patent integration. *J. Inf. Process. Manag.* **52**(10), 581–591 (2010)
16. Yamamoto, M., et al.: A journal paper filtering using the profile revised by patent document information. *IEEJ Trans. Electron. Inf. Syst.* **130**(2), 358–366 (2010)
17. Yamamoto, M., et al.: A journal paper filtering using the multiple information. *IEEJ Trans. Electron. Inf. Syst.* **131**(6), 1250–1259 (2013)
18. Kleinberg, J.: Authoritative sources in a hyperlinked environment. *J. ACM* **46**(5), 604–632 (1999)
19. Eto, M.: A new co-citation measure based on structures of citing papers. *Inf. Process. Soc. Jpn. Database* **49**, 1–15 (2008). (SIG 7 (TOD 37))
20. Ueda, I.: “Active mining utilizing the patent classification IPC, F1, F Term” on the basis of the cognitive processes of the examiner In: *Proceedings of SIG-FAI, Japanese Society for Artificial Intelligence*, vol. 46, pp. 13–21 (2001)
21. Karamon, J., et al.: Link mining from networks of academic papers, Technical report of IEICE (2007). *KBSE*, **106** (473), 73–78
22. Kashima, H.: Mining graphs and networks. *J. Inst. Electron. Inf. Commun. Eng.* **93**(9), 797–802 (2010)
23. Kajikawa, Y.: Utilization of citation information by link mining. *J. Inf. Sci. Technol. Assoc.* **60**(6), 224–229 (2010)
24. Gettor, L.: Link mining: a new data mining challenge. *ACM SIGKDD Explor. Newsl.* **5**(1), 84–89 (2003)
25. Outline of FI/F-term. https://www.jpo.go.jp/torikumi_e/searchportal_e/pdf/classification/fi_f-term.pdf. Accessed 23 Aug 2017
26. YUPASS. <http://www.yupass.jp>. Accessed 17 May 2017
27. Pelleg, D., Moore, A.W.: X-means: extending K-means with efficient estimation of the number of clusters. In: *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734. Morgan Kaufmann Publishers Inc., San Francisco (2000)
28. Ishioka, T.: Extended K-means with an efficient estimation of the number of clusters. *Jpn. J. Appl. Stat.* **29**(3), 141–149 (2000)
29. K Means. <http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>. Accessed 4 Mar 2018
30. scikit-learn. <http://scikit-learn.org/stable/>. Accessed 4 Mar 2018

LENLS 14

Logic and Engineering of Natural Language Semantics (LENLS) 14

Katsuhiko Sano

Department of Philosophy, Graduate School of Letters, Hokkaido University,
Nishi 7 Chome, Kita 10 Jo, Kita-ku, Sapporo, Hokkaido 060-0810, Japan
v-sano@let.hokudai.ac.jp

1 The Workshop

The international workshop Logic and Engineering of Natural Language Semantics (LENLS) was started in 2005. Its purpose is to provide a venue for researchers working on natural language semantics and pragmatics, (formal) philosophy, logic, artificial intelligence and computational linguistics together for discussion and interdisciplinary communication. Over the lifespan of the workshop, whose 14th iteration was held at JSAI-isAI 2017 from 13th to 15th November 2017, many researchers have presented their work, and the workshop has become recognized internationally in the semantics-pragmatics community.

LENLS 14 had 3 one-hour invited lectures and 27 thirty-minute submitted talks selected by the program committee (the total number of the submission was 36, which is the largest since LENLS 8, according to the EasyChair record). The number of participants is about fifty. The invited speakers were Craig Roberts (The Ohio State University), Ivano Ciardelli (Munich Center for Mathematical Philosophy, LMU München), and Shoichi Takahashi (Aoyama Gakuin University). Professor Roberts spoke about formal semantics and pragmatics of the *de se* interpretation of indexicals. Professor Ciardelli talked about his new analysis of counterfactuals that combines the framework of inquisitive semantics and the notion of causal reasoning, Professor Takahashi discussed his syntactic analysis of a restrictor NP (noun phrase) of an overtly displaced phrase. Topics discussed by the submitted papers raised issues from syntax-semantics-pragmatics interface, morpho-semantic interfaces, semantics of conditionals, semantics of emotions, type theory, semantics of expressives, categorial grammar, attitude verbs and evidentials, among many others. The papers in the present volume represent a selection of the papers presented at the workshop. As the reader can see the topics in this volume, the wide range of topics is characteristic of LENLS. All in all, the workshop was very successful and productive for both organizers and participants. We hope to keep this tradition in future to promote international researches in the semantics-pragmatics community.

2 Acknowledgements

Let me acknowledge some of those who helped with the workshop. The program committee and organisers, in addition to myself, were Daisuke Bekki, Elin McCready, Koji Mineshima, Alastair Butler, Richard Dietz, Naoya Fujikawa, Yoshiki Mori, Yasuo Nakayama, David Y. Oshima, Osamu Sawada, Wataru Uegaki, Katsuhiko Yabushita, Tomoyuki Yamada, Shunsuke Yatabe and Kei Yoshimoto. I also would like to acknowledge the external subreviewers for the workshop. Finally, the organisers would like to thank the JST CREST Programs “Advanced Core Technologies for Big Data Integration” for financial support and JSAI International Symposia on AI 2017 for giving us the opportunity to hold the workshop.



Relating Intensional Semantic Theories: Established Methods and Surprising Results

Kristina Liefke^(✉)

Institute for Linguistics, Goethe University Frankfurt,
60323 Frankfurt am Main, Germany
Liefke@lingua.uni-frankfurt.de
<http://liefke.wixsite.com/kristinaliefke>

Abstract. Formal semantics comprises a plethora of theories which interpret natural language through the use of different ontological primitives (e.g. individuals, possible worlds, situations, propositions, individual concepts). The ontological relations between these theories are, today, still largely unexplored. In particular, it remains an open question whether the primitives from some of these theories can be coded in terms of objects from other theories, or whether the ontologies of some theories can even be *reduced* to the ontologies of other, ontologically poorer, theories. This paper answers the above questions for a proper subset of formal semantic theories which are designed for the interpretation of doxastic attitude reports. The paper formalizes some ontological relations between these theories that are only suggested (but are not made explicit) in the literature, and identifies several new relations. The paper uses these relations to show that ‘the’ unifying theory for attitude reports is, in fact, a *class* of theories whose members are equivalent up to coding.

Keywords: Interpretation of doxastic attitude reports
(Hyper-)Intensional semantics · Ontological relations · Unification
Reduction

1 Introduction

The semantics of natural language presupposes a rich ontology. For example, to interpret the sentence *Every boy admires Mary*, we assume the existence of individuals (i.e. boys, Mary), propositions (i.e. *Every boy admires Mary*), properties

K. Liefke—I wish to thank two anonymous referees for LENLS 14 for their valuable comments and suggestions on an earlier version of this paper. The paper has profited from discussions with Chris Barker, Daisuke Bekki, Lucas Champollion, Ivano Ciardelli, Shalom Lappin, Roussanka Loukanova, Ed Zalta, and Ede Zimmermann. The research for this paper is supported by the German Research Foundation (via Kristina Liefke’s grant LI 2562/1-1 and Ede Zimmermann’s grant ZI 683/13-1).

- (5) a. Len thinks $[_{CP} \text{that } [_{DP} \text{Hesperus}] \text{ is a planet}]$. (T)
 b. ~~At all indices, Hesperus is Phosphorus.~~ (T)
 c. ~~Len thinks $[_{CP} \text{that } [_{DP} \text{Phosphorus}] \text{ is a planet}]$.~~ (F)
- (6) a. Len thinks $[_{CP} \text{that Phil is a } [_{CN} \text{groundhog}]]$. (T)
 b. ~~At all indices, all groundhogs are woodchucks.~~ (T)
 c. ~~Len thinks $[_{CP} \text{that Phil is a } [_{CN} \text{woodchuck}]]$.~~ (F)

To avoid predicting intuitively invalid inferences like (4) to (6), many revisions of Montague’s semantics introduce more fine-grained meanings (sometimes called *hyperintensions*; see [7,32]) that distinguish the semantic contributions of logically equivalent expressions. Hyperintensions are obtained by extending the sets of possible individuals, possible worlds, and truth-values by impossible individuals⁵, partial possible⁶/impossible or centered⁷ worlds, and truth-combinations⁸ (in ‘generalized’ theories; see [23,29,38]) or by replacing sets of possible worlds and functions from indices to individuals in attitude contexts by primitive, i.e. unanalyzable, propositions (type p) and/or by primitive individual concepts (type i) (in *property theories*; see [4,32,36]). For easy reference, we hereafter call the class of theories that are directed at the interpretation of doxastic attitude reports *intensional theories*. This class includes the above hyperintensional theories as well as the more classical possible worlds-theories (see [12,25,27]).

The class of intensional theories is typically also taken to include ‘structured intensions’-theories (see [8,21]) and *operational theories* (see [20,28,30]). ‘Structured intensions’-theories obtain hyperintensions by additionally considering the syntactic structure of the expressions that are semantically represented by the structured intensions. Operational theories obtain hyperintensions by assuming denotationally equivalent, but operationally distinct functions. However, since these two classes of theories are crucially different from the above-presented theories – and since they are less commonly used in contemporary formal semantic practice – we exclude them from our present considerations.

The combination of classical Montagovian objects (e.g. individuals, possible worlds, truth-values) with some of the new hyperintensional objects (e.g. partial/impossible/centered worlds, truth-combinations, primitive propositions, primitive individual concepts) has recently resulted in a plethora of competing formal-semantic theories (see Table 1). However, the interpretation of *the same* (or very similar) fragments of natural language through the use of *different* sets of primitives has resulted in a disunified semantics that leaves the relations between the different theories largely unexplored. In particular, it is unclear whether the basic objects from some of these theories can be coded in terms of objects from

⁵ These include physically impossible objects (e.g. perpetuum mobiles) and logically impossible or self-contradictory objects (e.g. the round square).

⁶ Partial worlds are more commonly known as *situations*.

⁷ These are ordered pairs $\langle i, x \rangle$ consisting of an index i and an individual x .

⁸ These additionally include the empty set of truth-values, $\{\} =: \mathbf{N}$ (i.e. *neither-true-nor-false*), and/or the set $\{\mathbf{T}, \mathbf{F}\} =: \mathbf{B}$ (i.e. *both-true-and-false*).

other theories – s.t. the explanatory or predictive power of one class of theories (with a certain set of primitives) can also be attributed to other classes of theories (with different primitives) –, or whether the ontologies of some of these theories can even be *reduced* to the ontologies of (ontologically poorer) theories (e.g. [5]) which do not contain special objects like possible worlds.

Our paper seeks to identify coding relations between the ontological primitives of different intensional semantic theories, and to use these relations to identify representation and reduction relations between the ontologies of these theories. These relations will then achieve a unification of the different theories. We expect that these relations will yield insights into the requirements on ontologically ‘minimal’ theories of doxastic attitude reports and that they will contribute to a better understanding of the linguistic-semantic type system. Apart from some isolated reductions (e.g. [15, 39]), such an effort has never been undertaken. However, only this effort enables us to transfer the interpretive success of one (class of) theory(s) (with a certain ontology) to other theories.

The paper is organized as follows: Sect. 2 surveys the ontological relations between different intensional semantic theories. These include ontological relations between different classical possible worlds-theories (see Sect. 2.1), between classical and generalized theories (see Sect. 2.2), and between generalized and property theories (see Sect. 2.3). On the basis of these relations, Sect. 2.4 identifies two classes of intensional theories with mutually representable ontologies. Section 3 uses the above findings for the construction of ontologically parsimonious, but explanatorily strong *models* of attitude reports.

2 Relations Between Intensional Theories

Table 1 captures the similarities and differences between the ontologies of the different intensional theories and identifies coding relations between the objects in these ontologies. In the table, the existence of a primitive in the ontology of the respective theory is marked by a cross, \times . Bracketed crosses, (\times) , indicate that the primitive is not a fully-fledged type, but is introduced through an additional type-forming rule. Crosses with a type subscript α , \times_α , indicate that objects of the relevant type are represented in the domain of type α .⁹ Crosses of the form $[\times_M]$ indicate that objects of this type are only present in the ontology of the metatheory. Crosses with a superscript plus, \times^+ , indicate that the relevant domain contains *more* objects than same-type domains in classical theories. In the domains of the types s and t , these ‘additional’ objects include partial/impossible/centered worlds and the truth-combinations \mathbf{N} and \mathbf{B} , respectively. In the domain of the type e , these objects include impossible individuals (see [38]) and nominalized properties (see [4]).

In the Table, we read ‘ $\alpha \longrightarrow \beta$ ’ as ‘objects of type α can be represented (or coded) by objects of type β ’. ‘ $\gamma \longleftarrow \alpha \longrightarrow \beta$ ’ is read as ‘objects of type α can be coded by constructions out of objects of the types β and γ ’. Reduction relations

⁹ For example, Zalta [38, pp. 644–646] represents situations and impossible worlds by (abstract) objects in the domain of individuals.

Table 1. Ontological relations between intensional theories.

	Models \backslash Primitive types	e	s	t	(st)	i	p
Class. theories	Montague 1973 (IL)	\times	\times	\times			
	Gallin 1975 (TY ₂)	\times	\times	\times			
	Montague 1970a (IL ⁻)	\times	\times	\times			
Gener'd theories	Zalta 1997 (OT)	\times^+	\times^+	\times^+			
	Muskens 1995 (TY ₂ ³ , TY ₂ ⁴)	\times^+	\times^+	\times^+			
	Liefke forthcoming (IS)	\times	\times^+	\times^+			
Property theories	Muskens 2005 (TY ₃)	\times	\times	\times			\times
	Chierchia & Turner (PT)	\times^+	\times^+	\times^+			\times
	Thomason 1980 (IntL)	\times	\times	\times			\times
	Pollard 2015 (AHS)	\times		\times		\times	\times

between objects in the ontologies of intensional theories which are identified in the literature are indicated by black solid arrows; embedding relations are indicated by grey solid arrows. Reduction (or embedding) relations which are suggested by these results, but which are – to the best of our knowledge – not explicitly addressed in the literature, are indicated by dashed black (grey) arrows.

Ontological relations from the literature include the possibility of embedding the ontology of Montague’s Intensional Logic (IL) from [26, 27] in the ontology of Gallin’s two-sorted variant of Church’s Simple Theory of Types, TY₂ (see [12]), of reducing the ontology of the linguistically relevant part of TY₂ to the ontology of IL (see [39]), and of embedding the ontology of TY₂ in the ontology of a type-theoretic variant of situation semantics, TY₂³ (see [29]). Ontological relations which are only suggested in the literature include the possibility of reducing the ontology of TY₂³ to the ontology of TY₂, of embedding the ontology of TY₂³ in a property-theoretic ontology (e.g. in the ontology of IntL from [36]), and of reducing property-theoretic ontologies that assume primitive possible worlds (e.g. the ontology of TY₃ from [30]) to property-theoretic ontologies without primitive worlds (e.g. to the ontologies of PT/AHS/IntL from [4, 32, 36]).¹⁰

The different ontological relations are discussed in some detail below, starting with relations between classical theories. In subsequent discussion, we use Gallin’s [12] convention of subscripting a logic’s name by the number of its basic types, not counting the type t . For every natural number n , we thus let ‘TY _{n} ’ denote a logic with $n + 1$ basic types (granted the existence of the type t). Correspondingly, the logic TY₁ (with basic types t and e) is Church’s Simple Theory of Types (see [5]). The logic TY₂ (with basic types t , e , and s) is Gallin’s two-sorted variant of this logic (see [12]). Following [29], we use a superscript number k to identify a logic as a k -valued logic. To capture the standardness of classical two-valued logics, we drop the superscript 2, s.t. we write TY _{n} ² simply as ‘TY _{n} ’.

¹⁰ For reasons of perspicuity, not all last-mentioned relations are visualized in Table 1.

2.1 Relations Between Classical Theories

Gallin's [12] embedding of IL in TY_2 is arguably the best-known relation between classical intensional theories. This embedding enables the interpretation of natural language in a simpler theory with nicer formal properties.¹¹ TY_2 differs from IL w.r.t. the ontological status of indices: while indices have the same status as individuals in TY_2 (s.t. they are in the domain of quantification and lambda abstraction), they are only derivative objects in IL, where they are introduced by a rule for the formation of intensional types. This rule states that, if α is an IL-type, then so is $s\alpha$ (see [27, p. 256]; cf. [26, pp. 227–228]). The type $s\alpha$ is instantiated, e.g., by the type for characteristic functions of sets of indices (i.e. st), by the type for individual concepts (i.e. se), and by the type for functions from indices to sets of individuals (i.e. $s(et)$). However, as a result of this 'rule-based' introduction of the type s , IL cannot quantify over indices. Consequently, it is often said¹² to be unable to model linguistic phenomena that require explicit quantification over worlds or times (see [9, 19, 33]).

To embed IL in TY_2 , Gallin analyzes extensional objects (of type α) as the extensions of intensional (type- $s\alpha$) objects at the actual-world index, and analyzes Montague's intensionalization and extensionalization operators, \wedge and \vee , as abstraction over and application to this index, respectively. The necessity and possibility operators, \square and \diamond , are then analyzed as universal and existential quantification over indices.

Gallin's embedding shows the possibility of interpreting natural language in ontologically richer theories that assume an equal (or 'symmetric' [33]) treatment of individuals and indices. Zimmermann [39] has shown that the ontology of the (large) part of TY_2 that is relevant for the interpretation of natural language can also be reduced to the ontology of IL, such that there is a *reduction* of TY_2 to IL. This reduction is made possible by the fact that explicit abstraction from and quantification over possible worlds do not add any objects to the original IL-ontology. Zimmermann obtains this result by eliminating all bound type- s variables which do not occur in IL.

A similarly interesting relation to the one above is the reduction of IL (or of TY_2) to ontologically even more parsimonious theories (e.g. [6, 25, 37]) which restrict intensional objects to type- st propositions. As a result of this restriction, the ontologies of these theories exclude (type- se) individual concepts, constructions involving individual concepts (e.g. objects of type $(se)t$ or $se(st)$), and functions from indices to non- t objects (e.g. objects of type $s(et)$). We hereafter call the logic that is associated with these theories IL^- .

¹¹ In contrast to IL, TY_2 validates full universal instantiation, beta conversion, and Leibniz' Law and has the diamond property (see [29, pp. 23–24]; cf. [11, p. 323]).

¹² For a refutation of this claim, see the discussion of [39] below.

The ontological reduction of IL (or of TY_2) to IL^- can be performed through a number of different coding strategies. In particular, Kaplan [15] has shown¹³ that the ontology of IL can be reduced to the ontology of IL^- by representing objects A of type $s\alpha$ ¹⁴ by type- $\alpha(st)$ functions, $\lambda a^\alpha \lambda i^s [A(i) = a]$, from type- α objects a to the set of indices at which the extension of A is a . This representation includes, as a special case, the representation of individual concepts as functions from individuals to the set of indices (type $e(st)$) at which these concepts have the individual as their extension. Extensional properties of individual concepts (type $(se)t$) and functions from indices to such properties (type $s((se)t)$) are then represented by objects of the types $(e(st))t$ and $(e(st))(st)$, respectively.

To improve upon the generality of Kaplan's result – and to allow for the full ontological reduction of TY_2 to IL^- (incl. the reduction of free-standing indices) –, Liefke [22] represents indices w by the characteristic function, $\lambda i^s [i = w]$, of their singleton sets. This representation enables the representation of individual concepts by functions from propositions to individuals (type $(st)e$) and of objects of the types $(se)t$ and $s((se)t)$ by objects of the type $((st)e)t$ resp. $((st)e)(st)$.¹⁵

The above-discussed relations all obtain between the ontologies of *intensional* theories. Liefke and Sanders [24] have shown that a proper part of the ontology of TY_2 (which restricts intensional objects to individual concepts and to constructions involving these concepts) can even be reduced to the ontology of Church's [5] *extensional* one-sorted type theory TY_1 . This reduction is achieved by representing individual concepts by coded finite sequences of natural numbers¹⁶ (type 0^* , or e) and by representing continuous functionals by lower-type objects (type $0^* \rightarrow 0$, or et) called the *associates* of these functionals. Associates of continuous functionals are countable representations of these functionals which uniquely determine the value of these functionals for every argument (see [16, 17]).

The ontological relations between classical intensional (and extensional) theories are summarized in Fig. 1. In the table, $X \longrightarrow Y$ is read as ‘the (objects in the) ontology of X can be represented (by objects) within the ontology of Y ’, where X and Y are type-logical theories. Subscripts under arrows attribute the identification of the relevant relation. Dashed arrows indicate partial relations.

¹³ More accurately, Kaplan shows “that Frege’s ontology [can be represented] within that part of it which constitutes Russell’s ontology” (see [15, p. 719]). Our description of Kaplan’s result is motivated by the identity of Frege’s ontology (when combined with a haecceitist position on trans-world identity; see [ibid., pp. 725–729]) with the ontology of IL, and by the identity of Russell’s ontology with the ontology of IL^- .

¹⁴ Kaplan’s strategy can be significantly simplified by requiring that the type $s\alpha$ be in \diamond -normal form (see [22, p. 11]). This form gathers occurrences of the index-type s immediately before t . Thus, the type $s(et)$ has the \diamond -normal form $e(st)$. Since $e(st)$ is already an IL^- -type, objects of type $s(et)$ will be represented in this type, rather than in the more complex type $et(st)$.

¹⁵ The latter coding is justified w.r.t. to the \diamond -normal form of $s((se)t)$, i.e. $(se)(st)$.

¹⁶ The relevant literature uses 0 and 0^* as the types for natural numbers, respectively for coded *sequences* of natural numbers. To ensure the transferability of results, we identify 0 with e .

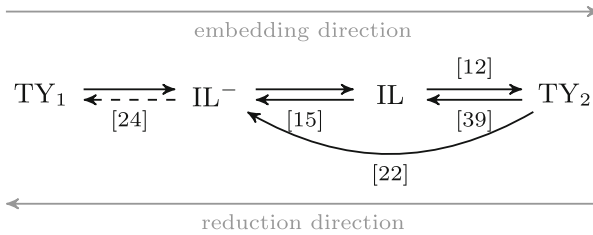


Fig. 1. Ontological relations between classical theories.

2.2 Relations Between Classical and Generalized Theories

We have suggested in Sect. 1 that generalized theories improve upon the predictive accuracy of classical possible worlds-theories by extending the ontology of classical theories with new kinds of individuals, worlds, and truth-values. Specifically, the generalized theories from Table 1 all extend the sets of possible worlds and truth-values from TY_2 by possible situations and by the truth-combination \mathbf{N} (*neither-true-nor-false*). The functional version¹⁷ of Zalta’s [38] *Object Theory* (hereafter, OT), the functional version of Muskens’ [29] relational type theory TT_2^4 (i.e. TY_2^4), and Liefke’s [23] *Integrated Semantics* (IS) further extend the sets of worlds and truth-values by impossible worlds (or by impossible situations; see [23,38]) and by the truth-combination \mathbf{B} (*both-true-and-false*). OT further adds, to the set of possible individuals from TY_2 , abstract individuals (incl. impossible individuals). To obtain more fine-grained linguistic meanings, IS replaces situations by centered situations (see [34]).

In virtue of the above extensions, the ontologies of TY_2^4 , of its partial variant TY_2^3 , of OT, and of IS all embed the ontologies of classical theories. However, results from universal algebra suggest that many of these embeddings can also be inverted. These results include the possibility of representing truth-combinations ξ by characteristic functions, $\lambda\vartheta^t [\vartheta \sqsubseteq \xi]$, of the set of truth-values that are included in ξ under the approximation-ordering on truth-combinations (see [1]; cf. [2]).¹⁸ They further include the possibility of representing situations and possible/impossible worlds σ by sets, $\lambda i^s [\sigma \leq i]$, of possible worlds whose information contains (a consistent part of) the information of σ (see [35]).¹⁹ In this

¹⁷ In contrast to the theory’s original relational formulation, this version has a primitive type for truth-combinations.

¹⁸ Since the truth-combination \mathbf{B} includes both \mathbf{T} and \mathbf{F} on this ordering, it will be represented by the set $\{\mathbf{T}, \mathbf{F}\}$. Since the truth-combination \mathbf{N} includes neither \mathbf{T} nor \mathbf{F} (or \mathbf{B}), it will be represented by the empty set. Since each of \mathbf{T} and \mathbf{F} only includes \mathbf{N} and itself, they will be represented by their singleton sets. The relation \sqsubseteq is discussed in detail in [2].

¹⁹ Possible worlds will then be represented by singleton sets containing these worlds. Situations/impossible worlds will be represented by sets of worlds whose members extend the information of the situation, resp. whose members capture a total consistent part of the world’s information. The relation \leq is introduced in [29, pp. 69–74].

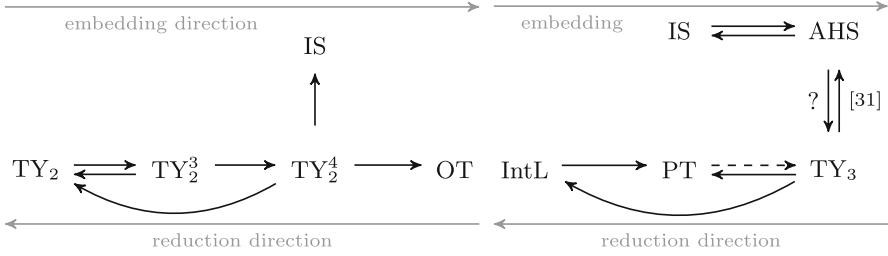


Fig. 2. Ontological relations between generalized theories.

Fig. 3. Ontological relations between property theories.

way, these results enable the representation of the ontologies of TY_2^3 and TY_2^4 in the ontology of TY_2 . The strategy of representing situations as sets of worlds also stands behind the representation of indices by their singletons from [22] (see Sect. 2.1).

Because of the large cardinality of OT’s type- e domain (see Table 1) – and the resulting impossibility of representing each individual from OT by an individual in the domain of TY_2^4 –, the above results only enable a partial reduction of the ontology of OT to the ontology of TY_2 . A related observation holds for the reduction of IS to TY_2 : since TY_2 does not have an operation for the formation of product types (s.t. its ontology does not contain representations of centered situations), the ontology of IS resists a full reduction to the ontology of TY_2 .

The ontological relations between generalized theories, and between generalized theories and TY_2 , are summarized in Fig. 2.

2.3 Relations Between Generalized and Property Theories

We have seen above that, by extending the familiar type-theoretic domains, generalized theories distinguish more intuitively different sentence-meanings than classical theories. However, the coarse grain of situations still limits the predictive accuracy of these theories. For example, since the nouns *groundhog* and *woodchuck* are co-intensional (s.t. every individual which is a groundhog in some situation is also a woodchuck in this situation, and *vice versa*), many²⁰ generalized theories still predict the validity of counterintuitive inferences like (6).

Property theories (e.g. [4, 30, 32, 36]) solve this problem²¹ by replacing sets of worlds/of situations by primitive propositions as the meanings of sentences. However, to interpret natural language modals and counterfactuals, to obtain truth at a world, and to predict the right entailments of intensional, *non*-hyperintensional

²⁰ This excludes OT and IS, which assign more fine-grained meanings to attitude complements (see Sect. 2.2).

²¹ The impossibility of representing primitive propositions by type- st (or type- $(st)(tt)$) objects motivates the distinction between generalized and property theories. This distinction will be discussed in more detail in Sect. 2.4.

constructions like (7),²² many property theories (e.g. [30]; see [36, Sect. 7]) still assume a dedicated domain of possible worlds, and of sets of possible worlds. Sets of worlds are then connected to primitive propositions via a homomorphism²³ from hyperintensions to intensions. To predict the right entailment-conditions for constructions like (7), the meanings of sentences in intensional contexts are identified with the values of this homomorphism for the relevant hyperintension.

- (7) a. The test shows $[_{CP}\text{that Phil is a } [_{CN}\text{groundhog}]]$. (T)
 b. At all indices, all groundhogs are woodchucks. (T)
 c. The test shows $[_{CP}\text{that Phil is a } [_{CN}\text{woodchuck}]]$. (T)!

Recently, Pollard [32] has shown that property theories can drop worlds as primitives without losing explanatory power (see [10, 31]). This is made possible by representing possible worlds w by ultrafilters in the boolean prealgebra of propositions (i.e. by characteristic functions of maximal consistent sets of propositions). The truth of a proposition at a world w is then defined via the proposition's membership in the ultrafilter representing w . Linguistic modal operators (e.g. necessarily) are interpreted as quantifiers over such ultrafilters. The representation of intensions of type $s\alpha$ by objects of the type $(pt)\alpha$ preserves the homomorphism from hyperintensions to intensions. This representation includes the representation of type-*se* individual concepts by objects of the type $(pt)e$, and the representation of extensional and intensional properties of individual concepts (types $(se)t$, $(se)st$) by objects of the type $((pt)e)t$, resp. $((pt)e)(pt)t$.

To obtain hyperfine-grained sentence-meanings in a compositional manner, property theories interpret proper names in the hyperintensional type for quantificational DPs, $(ep)p$ (see [29, 36]) or in the type for primitive individual concepts, i (see [10, 31, 32]). Theories of the latter sort associate different type- i interpretations of co-referential names with different ways of identifying the names' type- e referent at all indices.²⁴ Theories of the former sort associate different type- $(ep)p$ interpretations of co-referential names with the referent's different uniquely identifying properties. The analysis of names as rigid designators requires that these properties be essential properties (i.e. properties which are uniquely exemplified by the same individual at all indices). This requirement is implemented through semantic constraints on the property-theoretic models interpreting names (see [36, pp. 60–61]). To the best of our knowledge, the question whether primitive individual concepts can be represented by type- $(ep)p$ objects is still open.

The possibility of representing possible worlds by ultrafilters on primitive propositions suggests that the ontology of Muskens' [30, Sect. 4] 'propositional'

²² These contexts allow the truth-preserving substitution of equivalent expressions.

²³ To ensure the finer grain of hyperintensions (than intensions), it is commonly assumed that this homomorphism is not an isomorphism (see [30, 36]).

²⁴ This follows from the analysis of proper names as rigid designators (which have the same referent at all indices) (see [18]; cf. [31, p. 260]) and from the assumption that primitive individual concepts are more fine-grained than functions from indices to individuals (see [31, pp. 273, 277–278]).

enrichment of the logic TY_2 , i.e. TY_3 , can be represented in the ontology of Pollard's [31] *Agnostic Hyperintensional Semantics* (AHS). The latter is a propositional enrichment of the logic TY_1 that commands a.o. the types t , e , and p . Since the ontology of TY_3 contains the ontology of Chierchia and Turner's [4] *Property Theory* (hereafter, PT) and of Thomason's *Intentional Logic* (IntL) as its proper parts, Pollard's result also suggests the possibility of reducing the ontology of TY_3 to the ontologies of PT and IntL.²⁵ (Note however that, because of the large cardinality of PT's type- e domain, the ontological reduction of TY_3 to PT is only a partial reduction.)

Pollard's strategy for the elimination of primitive possible worlds can be straightforwardly applied to the elimination of impossible worlds and of (possible or impossible) situations. This elimination uses a four-valued²⁶ variant of the function, $\lambda i^s \lambda p^p [p@i]$, for the type- pt representation of possible worlds, where ' $p@i$ ' is read as ' p is true at i '. This function produces consistent sets of propositions on input possible situations (prime filters of propositions on input complete spatio-temporal parts of possible worlds), and produces inconsistent sets of propositions on input impossible worlds or situations. The possibility of giving type- $p(tt)$ representations of (possible or impossible) worlds and situations enables the representation of the ontologies of TY_2^3 and TY_2^4 in the ontology of TY_3 and, consequently, in the ontology of AHS.

We have noted in Sect. 2.2 that IS's functions from centered situations to truth-combinations are more fine-grained than their non-centered counterparts from TY_2^3 or TY_2^4 , such that the ontology of IS resists a full reduction to the ontology of TY_2 . In particular, IS-interpretations of attitude reports capture speakers' intuitions about the preservation of meaning under substitution, such that they correctly predict the substitution-resistance of (4) to (6) (see [23]). Since these intuitions also determine the level of granularity of primitive propositions in property theories (see [32, p. 553]), IS is ontologically equivalent to AHS.

The ontological relations between property theories are summarized in Fig. 3.

2.4 Classes of Intensional Theories

The above yields the network of ontological relations between intensional theories from Fig. 4.

Figure 4 identifies two classes of intensional semantic theories whose members allow the mutual representation of their ontologies. The difference between the two classes lies in the presence vs. absence of objects that are sufficiently

²⁵ In virtue of the possibility of representing possible worlds by ultrafilters of propositions, the ontologies of PT and IntL further enable the adequate interpretation of intensional constructions like (7).

²⁶ This variant reflects the possibility that *neither* a proposition nor its complement are true at a *situation* and that *both* a proposition and its complement are true at an *impossible* world or situation. To capture this possibility, we represent worlds and situations by (type- pt , or $-p(tt)$) functions from primitive propositions to truth-combinations.

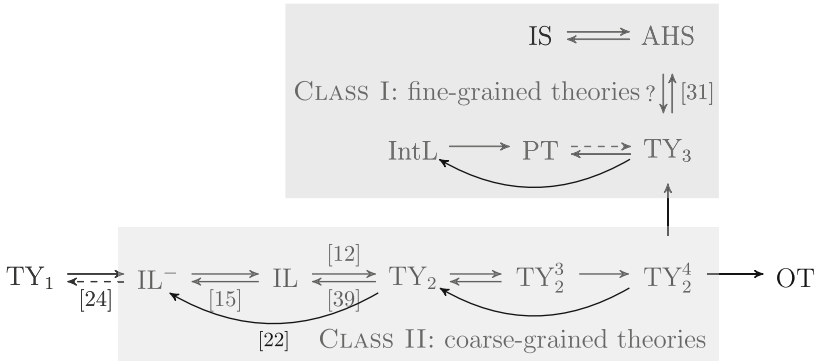


Fig. 4. Ontological relations b/w classical, generalized, and property theories.

fine-grained to serve the interpretation of doxastic attitude complements (e.g. the interpretation of the occurrences of (2a) and (3a) from (5a) and (6a)). Since primitive propositions (and objects representing these propositions) satisfy this granularity requirement, the ontologies of AHS, TY_3 , PT, IntL, OT, and IS all qualify as members of the class of *fine-grained* intensional theories. Since functions of the generalized type *st* (and objects representing such functions) do not satisfy this requirement, the ontologies of TY_2 , TY_2^3 , TY_2^4 , IL, and IL^- qualify as members of the class of *coarse-grained* intensional theories. The finer grain of primitive propositions in comparison to objects of the above type prevents the ontological reduction of fine-grained to coarse-grained theories.

3 Ontological Inter-model Reduction: A Sketch

Our previous considerations have been restricted to relations between the ontologies of different intensional semantic theories, i.e. to relations between the *frames* of models of these theories. However, so long as these models assume the familiar interpretation functions (which send terms of the theory’s language to the frame’s ‘default’ objects, rather than to more fine-grained objects that can be represented by these objects), they will not achieve the greater explanatory or predictive success of models which assume these fine-grained objects. For example, since standard models of TY_2 interpret (1b) as the *total* set of possible *worlds* in which 7 is a prime (rather than as the (coded) *partial* set of *situations* in which 7 is a prime), this model is still unable to explain the intuitive invalidity of (4).

To transfer the explanatory and predictive power of models of one theory (with a certain ontology) to models of another theory (with a different ontology), we thus need to supplement the coding relations from Sect. 2 with a number of further steps. For the indirect interpretation of natural language in type-logical models (see [26,27]), these steps include the following:

- (i) **type conversion:** the recursive specification of a function ξ – in line with the coding of objects from Sect. 2 – that converts the types of all objects of the reduced theory (above: TY_2^3) into the types of objects of the reducing theory (here: TY_2);
- (ii) **translation to converted-type terms:** the recursive specification of a function \bullet that translates all type- α terms A in the language²⁷ of the reduced theory into type- $\xi(\alpha)$ terms A^\bullet in the language of the reducing theory;
- (iii) **semantic restriction:** the specification of semantic constraints on the interpretation function, \mathcal{I}' , of the model of the reducing theory. These constraints ensure that \mathcal{I}' preserves the semantic distinctions that are induced by the interpretation function, \mathcal{I} , of the model of the reduced theory, s.t., for all terms A, B in the language of the reduced theory,

$$\mathcal{I}'(A^\bullet) = \mathcal{I}'(B^\bullet) \text{ iff } \mathcal{I}(A) = \mathcal{I}(B).$$

The constraints from step (iii) ensure that a sentence's interpretation in the reducing model preserves the predictions (*qua* sentential truth and entailment) of the reduced model.

For the reduction of a TY_2^3 -model to a TY_2 -model for the interpretation of natural language sentences (here: for the interpretation of (1a) and (1b)), a suitable candidate for the function ξ from step (i) is given below:

Definition 1 (TY_2^3 -to- TY_2 type-conversion). *The function ξ connects TY_2^3 types with TY_2 types via the following recursion:*

- I. (i) $\xi(e) = e$; (ii) $\xi(s) = st$; (iii) $\xi(t) = tt$;
- II. $\xi(\alpha \rightarrow \beta) = (\xi(\alpha) \rightarrow \xi(\beta))$

Clauses I.(ii) and I.(iii) capture the conversion of the TY_2^3 types for situations s and partial truth-combinations t to the TY_2 types for characteristic functions of sets of possible worlds, st , and for characteristic functions of sets of truth-values, tt (see Sect. 2.2). The remaining (sub-)clauses then enable the conversion of the TY_2^3 types for characteristic functions of partial sets of situations, st , and for functions from situations to partial sets of individuals, $s(et)$, to the TY_2 types $(st)(tt)$ and $(st)(e(tt))$, respectively, and enable the conversion of the TY_2^3 types of generalized quantifiers, i.e. $s((s(et))t)$, of determiners, i.e. $(s(et))((s(et))t)$, and of the copula **be**, i.e. $(s((s(et))t))(et)$, to the TY_2 types $(st)((st)(e(tt)))(tt)$, $((st)(e(tt)))((st)(e(tt)))(tt)$, and $((st)((st)(e(tt)))(tt))(e(tt))$, respectively.

A suitable candidate for the translation function from step (ii) will translate terms of the designated²⁸ TY_2^3 language \mathcal{L}^3 to terms of the designated reducing TY_2 language \mathcal{L}^2 as follows: (This translation uses the convention for TY_2^3 and TY_2 constants and variables from Tables 2 and 3).

²⁷ We assume that this is a *designated* language, whose constants are associated with the lexical elements of the target natural language (here: English).

²⁸ This is a language whose constants are associated with the lexical elements of (1a) and (1b).

Table 2. \mathcal{L}^3 constants and variables.

CONST.	TY ₂ ³ TYPE	VAR.	TY ₂ ³ TYPE
1, 2, 7	e		
<i>plus</i>	$e(ee)$	x	e
<i>prime</i>	$s(et)$	i	s

Table 3. \mathcal{L}^2 constants and variables.

CONST.	TY ₂ TYPE	VAR.	TY ₂ TYPE
1, 2, 7	e	ϑ	t
<i>plus</i>	$e(ee)$	x	e
<i>prime</i>	$(st)(e(tt))$	i	st

Definition 2 (TY₂-translation of TY₂³ terms). *The relation \bullet connects the designated terms of the logic TY₂³ with terms of the logic TY₂ as follows, where A , B , and C are suitably typed TY₂³ terms:*

- I. (i) $1^\bullet = 1$; $2^\bullet = 2$ $7^\bullet = 7$; $plus^\bullet = plus$; $prime^\bullet = \mathbf{prime}$;
(ii) $x^\bullet = x$; $i^\bullet = i$;
- II. (i) $(B(A))^\bullet = (B^\bullet(A^\bullet))$;
(ii) $(\lambda y_\alpha. A_\beta)^\bullet = \lambda y^\bullet \lambda \vartheta. A^\bullet(\vartheta)$ if $\xi(\beta) = tt$;
 $(\lambda y_\alpha. A_\beta)^\bullet = \lambda y^\bullet \lambda \vartheta. A^\bullet$ otherwise, if $\xi(\beta) = t$;
(iii) $(B = C)^\bullet = B^\bullet = C^\bullet$

The translations of TY₂³ terms containing connectives other than $=$ are obtained from the above via the definition of the connectives and quantifiers from [13].

The above enables the translation of the TY₂³ renderings of (1a) and (1b), i.e. $\lambda i [plus(1)(1) = 2]$ and $\lambda i [prime(i)(7)]$, into the TY₂ terms from (7) and (8). (Below, \rightsquigarrow is the translation function from natural language expressions to TY₂ terms.)

$$\begin{array}{ll}
 (8) \text{ One plus one equals two} & \text{Seven is a prime number} \quad (9) \\
 \rightsquigarrow (\lambda i [plus(1)(1) = 2])^\bullet & \rightsquigarrow (\lambda i [prime(i)(7)])^\bullet \\
 = \lambda i^\bullet \lambda \vartheta [plus(1)(1) = 2]^\bullet & = \lambda i^\bullet \lambda \vartheta [(prime(i)(7))^\bullet(\vartheta)] \text{ (by II.(ii))} \\
 = \lambda i^\bullet \lambda \vartheta [plus^\bullet(1^\bullet)(1^\bullet) = 2^\bullet] & = \lambda i^\bullet \lambda \vartheta [prime^\bullet(i^\bullet)(7^\bullet)(\vartheta)] \text{ (by II.(i))} \\
 = \lambda i \lambda \vartheta [plus(1)(1) = 2] & = \lambda i \lambda \vartheta [\mathbf{prime}(i)(7)(\vartheta)] \text{ (by I)}
 \end{array}$$

The semantic constraints on the interpretation function, \mathcal{I}' , of the reducing model of TY₂ (see step (iii)) are given in Definition 3:

Definition 3 (Semantic constraints on translating TY₂ terms). *The interpretation function, \mathcal{I}' , of the TY₂³-reducing model of TY₂ satisfies the following constraints:*

- I. $\mathcal{I}'(A_e^\bullet) = \mathcal{I}(A_e^\bullet)$; $\mathcal{I}'(A_s^\bullet) = \mathcal{I}(\lambda i [A \leq i])$; $\mathcal{I}'(A_t^\bullet) = \mathcal{I}(\lambda \vartheta^t [\vartheta \sqsubseteq A])$;
II. (i) $\mathcal{I}'((B(A))^\bullet) = \mathcal{I}'(B^\bullet)(\mathcal{I}'(A^\bullet))$;
(ii) $\mathcal{I}'((\lambda y_\alpha. A_\beta)^\bullet)$ = the function F with domain $D_{\xi(\alpha)} s.t.$
for all $d \in D_{\xi(\alpha)}$, $F(d) = \mathcal{I}'(A^\bullet)^g[d/y^\bullet]$;
(iii) $\mathcal{I}'((B = C)^\bullet) = \mathbf{T}$ iff $\mathcal{I}'(B^\bullet) = \mathcal{I}'(C^\bullet)$ iff $\mathcal{I}(B) = \mathcal{I}(C)$

The constraints from clause I, above, follow the description of the TY₂ coding of situations and truth-combinations from Sect. 2.2 (see [1, 35]). Together with the constraints from clause II, they ensure that the interpretations-under- \mathcal{I}' of (1a) and (1b) share the truth- and falsity-conditions of the interpretations of these sentences under the TY₂³ interpretation function \mathcal{I} . In particular, since (we assume that) there are some situations σ in the frame of

the designated TY_2^3 model at which the interpretation of $\lambda i [\textit{prime}(i)(7)]$ is undefined (s.t. $\mathcal{I}(\textit{prime}(\sigma)(7)) = \mathbf{N}$), there will also be sets of worlds σ in the frame of the reducing TY_2 model at which the interpretation-under- \mathcal{I}' of $\lambda i \lambda \vartheta [\textit{prime}(i)(7)(\vartheta)]$ is the empty set of truth-values, \emptyset . Since the interpretation of the TY_2 translation of the TY_2^3 rendering of (1a) yields a non-empty set of truth-values at each value of i ,²⁹ (1b) is not equivalent to (1a) in the reducing TY_2 model.

Notably, in the reducing TY_2 model, the notions of truth, entailment, and equivalence have a different definition than in ‘standard’ TY_2 models (e.g. [12]; see [29]). In standard TY_2 models, the notions of truth and (mutual) entailment are defined at the type for truth-values, t , or at the type for characteristic functions of sets of possible worlds, st . Since our reducing model interprets natural language sentences in the type $(st)(tt)$ (see (8), (9)), it demands that these notions instead be defined at the type $(st)(tt)$, or tt . The different-type interpretation of sentences in the reducing TY_2 model (w.r.t. the interpretation of sentences in standard TY_2 models) also demands that entailment be defined in terms of a different algebraic structure than entailment in standard TY_2 models, i.e. in terms of the logical ordering in the Kleene algebra on $\{\emptyset, \{\mathbf{T}\}, \{\mathbf{F}\}, \{\mathbf{T}, \mathbf{F}\}\}$ (see [1, 2]), rather than in terms of the logical ordering in the Boolean algebra on $\{\mathbf{T}, \mathbf{F}\}$. This definition strengthens the notion of entailment in the reducing TY_2 model, as is required to explain the difference between the ‘reducing’ TY_2 -interpretations of (1a) and (1b).

4 Conclusion

This paper has identified ontological relations between different formal semantic theories for the interpretation of doxastic attitude reports. The paper has used these relations to classify the theories according to their ontologies’ mutual representability, and has outlined a general strategy for the construction of ontologically parsimonious, but explanatorily strong models for attitude reports.

The ability to transfer the explanatory success of one theory (or model) to another, ontologically more parsimonious, theory (or model) is arguably a very desirable result. However, the identification of ontological intertheoretic (or inter-model) relations – and of classes of ontologically equivalent theories (or models) – also yields valuable insights about the requirements on ontologically minimal theories of attitude reports and about the semantic type system. In particular, the possibility of coding situations as sets of worlds – or of coding worlds as ultrafilters on primitive propositions – shows that an ontologically ‘minimal’ semantics for doxastic attitude reports need not assume primitive situations, or even worlds. The resulting robustness of semantic theories w.r.t. their exact choice of primitives – and the suitability of the more familiar classical possible worlds-theories for most purposes (e.g. for the explanation of the invalidity of (4)) – then explains the existence of the large number of competing intensional theories.

²⁹ This is due to the fact that $=$ is totally defined, s.t. the interpretation of $\lambda i \lambda \vartheta [\textit{plus}(1)(1) = 2]$ at a particular situation-argument is independent of this argument.

References

1. Belnap, N.D.: A useful four-valued logic. In: Dunn, J.M., Epstein, G. (eds.) *Modern Uses of Multiple-Valued Logics*, vol. 2. Springer, Dordrecht (1977). https://doi.org/10.1007/978-94-010-1161-7_2
2. Blamey, S.: Partial logic. In: Gabbay, D.M., Guenther, F. (eds.) *Handbook of Philosophical Logic*, vol. 5. Kluwer Academic Publishers, Dordrecht (2002)
3. Carnap, R.: *Meaning and Necessity: A Study in Semantics and Modal Logic*. University of Chicago Press, Chicago (1988)
4. Chierchia, G., Turner, R.: Semantics and property theory. *Linguist. Philos.* **11**(3), 261–302 (1988)
5. Church, A.: A formulation of the simple theory of types. *J. Symb. Log.* **5**(2), 56–68 (1940)
6. Cresswell, M.J.: *Logics and Languages*. Methuen Young Books, London (1973)
7. Cresswell, M.J.: Hyperintensional logic. *Stud. Logica* **34**(1), 25–38 (1975)
8. Cresswell, M.J.: *Structured Meanings*. MIT Press, Cambridge (1985)
9. Cresswell, M.J.: *Entities and Indices*. Kluwer Academic Publishers, Dordrecht (1990)
10. Fox, C., Lappin, S., Pollard, C.: A higher-order fine-grained logic for intensional semantics. In: *Proceedings of the 7th International Symposium on Logic and Language*, pp. 37–46 (2002)
11. Friedman, J., Warren, D.: Lambda normal forms in an intensional logic for English. *Stud. Logica* **39**, 311–324 (1980)
12. Gallin, D.: *Intensional and Higher-Order Modal Logic with Applications to Montague Semantics*. North Holland, Elsevier (1975)
13. Henkin, L.: Completeness in the theory of types. *J. Symb. Log.* **15**, 81–91 (1950)
14. Hintikka, J.: Impossible possible worlds vindicated. *J. Philos. Log.* **4**(4), 475–484 (1975)
15. Kaplan, D.: How to Russell a Frege-Church. *J. Philos.* **72**, 716–29 (1975)
16. Kleene, S.C.: Countable functionals. In: Heyting, A. (ed.) *Constructivity in Mathematics*, North-Holland, pp. 81–100 (1959)
17. Kreisel, G.: Interpretation of analysis by means of constructive functionals of finite types. In: *Constructivity in Mathematics*, pp. 101–128 (1959)
18. Kripke, S.A.: *Naming and Necessity*. Harvard University Press, Cambridge (1980)
19. Kusumoto, K.: On the quantification over times in natural language. *Nat. Lang. Semant.* **13**(4), 317–357 (2005)
20. Lappin, S.: Curry typing, polymorphism, and fine-grained intensionality. In: Lappin, S., Fox, C. (eds.) *Handbook of Contemporary Semantic Theory*, 2nd edn, pp. 408–428. Wiley-Blackwell (2015)
21. Lewis, D.: General semantics. *Synthese* **22**(1-2), 18–67 (1970)
22. Liefke, K.: Codability and robustness in formal natural language semantics. In: Murata, T., Mineshima, K., Bekki, D. (eds.) *JSAI-isAI 2014. LNCS (LNAI)*, vol. 9067, pp. 6–22. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48119-6_2
23. Liefke, K.: A compositional pluralist semantics for extensional and attitude verbs. In: Löbner, S., et al. (eds.) *Language, Cognition, and Mind. Selected Revised Papers from Cognitive Structures 16*. Springer, Heidelberg (forthcoming)
24. Liefke, K., Sanders, S.: A computable solution to Partee’s temperature puzzle. In: Amblard, M., de Groote, P., Pogodalla, S., Retoré, C. (eds.) *LACL 2016. LNCS*, vol. 10054, pp. 175–190. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53826-5_11

25. Montague, R.: English as a formal language. In: Thomason, R.H. (ed.) *Formal Philosophy. Selected Papers of Richard Montague*, Yale UP, pp. 188–221 (1976, 1970a)
26. Montague, R.: Universal grammar. In: Thomason, R.H. (ed.) *Formal Philosophy. Selected Papers of Richard Montague*, pp. 222–246 (1976, 1970b)
27. Montague, R.: The proper treatment of quantification in ordinary English. In: Thomason, R.H. (ed.) *Formal Philosophy. Selected Papers of Richard Montague*, pp. 247–270 (1976, 1973)
28. Moschovakis, Y.: A logical calculus of meaning and synonymy. *Linguist. Philos.* **29**(1), 27–89 (2006)
29. Muskens, R.: *Meaning and Partiality*. CSLI Publications, Stanford (1995)
30. Muskens, R.: Sense and the computation of reference. *Linguist. Philos.* **28**(4), 473–504 (2005)
31. Pollard, C.: Hyperintensions. *J. Log. Comput.* **18**(2), 257–282 (2008)
32. Pollard, C.: Agnostic hyperintensional semantics. *Synthese* **192**(3), 535–562 (2015)
33. Schlenker, P.: Ontological symmetry in language: a brief manifesto. *Mind Lang.* **21**(4), 504–539 (2006)
34. Stephenson, T.: Vivid attitudes: centered situations in the semantics of ‘remember’ and ‘imagine’. *Semant. Linguist. Theor.* **20**, 147–160 (2010)
35. Stone, M.H.: The theory of representation for Boolean algebras. *Trans. Am. Math. Soc.* **40**(1), 37–111 (1936)
36. Thomason, R.H.: A model theory for the propositional attitudes. *Linguist. Philos.* **4**(1), 47–70 (1980)
37. Turner, R.: Types. In: Benthem, J.V., et al. (eds.) *Handbook of Logic and Language*. Elsevier (1997)
38. Zalta, E.N.: A classically-based theory of impossible worlds. *Notre Dame J. Formal Log.* **38**(4), 640–660 (1997)
39. Zimmermann, T.E.: Intensional logic and two-sorted type theory. *J. Symb. Log.* **54**(1), 65–77 (1989)



Expressive Small Clauses in Japanese

Yu Izumi^(✉) and Shintaro Hayashi

Department of Anthropology and Philosophy, Nanzan University,
18 Yamazato-cho, Showa-ku, Nagoya, Aichi 466-8673, Japan
{yuizumi, sh0211}@nanzan-u.ac.jp

Abstract. This paper modifies and extends Potts and Roeper’s (2006) analysis of what they call Expressive Small Clauses, simple uses of epithets such as *You fool!*, to analogous phrases in Japanese. The original Potts and Roeper analysis is unable to account for two puzzling characteristics of Japanese Expressive Small Clauses that are not shared with those in English: first, the use of a second person pronoun is not permitted in the Japanese counterparts, whereas many other forms of pronouns and non-pronominal nouns are available; second, something like “you fool” in Japanese can indeed occur as an argument of a sentence. Drawing on the recent syntactic literature on the morphological variation of the analogous nominal epithets, the paper proposes an account that explains the differences between English and Japanese Expressive Small Clauses.

1 Introduction

In recent years, linguistic studies have been making contributions to the understanding of the socially problematic aspect of language use, including slurring, dog whistling, propaganda, and hate speech (Potts 2007; Croom 2011; Langton 2012; Stanley 2015 among others). Seemingly simple insults referred to as “Expressive Small Clauses” (ESC) by Potts and Roeper (2006, henceforth P&R), illustrated in (1), are also a part of “fighting words” that can provoke an audience and even incite violence (Greenawalt 1995).¹

- (1) a. Oh, You fool!
b. You idiot!

According to P&R, ESCs are a class of small clauses that are devoid of any functional structure and thereby “necessarily verbless” (p. 184). The expressive connotation of an ESC originates in the lexical property of the noun used. In an ESC, the denigrating noun does not carry ordinary descriptive content, such

This research was in part supported by Nanzan University Pache Research Subsidy 1-A-2 for the 2017 academic year.

¹ To be accurate, P&R’s focus is on ESCs that are used as self-disapprobation, not on those used to insult others, but their analysis applies to other-directed ESCs as well.

as one of type $\langle e, t \rangle$, but instead expressive content. For example, (2a) can be analyzed as (2b), where E is an expressive type that allows no further semantic computation. That is, P&R’s type theory excludes the possibility of types beginning with type E (e.g., $\langle E, e \rangle$, $\langle E, \langle e, t \rangle \rangle$).

- (2) a. You fool!
 b. fool (you) : E (P&R 2006, p. 196, ex. 40)
- \swarrow \searrow
 you : e fool : $\langle e, E \rangle$

It is important to note that P&R used the normal Function Application in (2b), not a multidimensional composition rule that is commonly appealed to when computing expressive content simultaneously with non-expressive, descriptive content (e.g., “CI Application” in Potts 2005, p. 64). If such a multidimensional rule were employed here, the ESC would have had a descriptive content of type e , in addition to its expressive content of type E . This consequence must be avoided because, as P&R observed, an ESC is never used in argument position and generally unembeddable, as shown in (3–4).

- (3) a. You fools should read more carefully.
 b. * You fool should read more carefully. (P&R 2006, p. 197, ex. 44)
- (4) * I consider you fool/nincompoop/screwball. (P&R 2006, p. 187, ex. 15a)

The contrast between (3a) and (3b) indicates that ESCs “do not behave in any sense like nominals” (p. 197), and P&R explained this fact by analyzing them as having “one-dimensional expressive meanings” (p. 197). Below, we will consider Japanese examples that appear to be in conflict with P&R’s observation here.

This paper modifies and extends P&R’s expressive analysis of ESCs to analogous insults in Japanese. First, Sect. 2 introduces Japanese ESCs and discusses how they differ from their English counterparts, showing that the original P&R analysis is unable to account for two puzzling characteristics of Japanese ESCs. Second, drawing on the recent syntactic literature on the morphological variation of ESCs, Sect. 3 offers solutions to the puzzles identified in Sect. 2.

2 Two Puzzles About Japanese ESCs

2.1 What Are ESCs in Japanese?

P&R presented a brief survey of cross-linguistic variation in the forms of ESCs, discussing languages such as German and Afrikaans, and they also presented (5a–b) as instances of ESCs in Japanese. Neither of them, however, is clearly similar to their basic example (1) above.

- (5) a. Ore tte baka da na.
 me TOPIC idiot COPULA PART
 “I am such a fool”
 b. Ore tte o-baka-san.
 me TOPIC HON-idiot-HON

(P&R 2006, p.192, ex. 31, translation added by the authors)

First, since (5a) includes the copula *da*, it fails to meet P&R’s own criterion for ESCs, according to which they must be verbless. Second, as opposed to ESCs, which are highly productive as noted by P&R, (5b) is not compatible with a different noun, as the contrast between the English example (6) and the Japanese purported counterpart (7) clearly indicates: (7) is simply incomprehensible even with enough contextual information. The contrast suggests that the phrase *o-baka-san* is a conventional, idiomatic phrase that does not have a productive usage.

- (6) a. You are trying to make allies again. You politician!
 b. You broke a glass again. You elephant! (Arsenijević 2007, p.89, ex. 4)
 (7) */?? Ore tte o-seijika/zou-san.
 me TOPIC HON-politician/elephant-HON

Third, the straightforward interpretation of (5a–b) is more aligned with the full-sentential counterparts of (1) (*I’m a fool*, etc.). (5a–b) do not seem to us to express emotive frustration about a present circumstance unlike (1); they are most naturally understood as a self-deprecating observation about one’s stable property, such as character or habit.

Based on these considerations, we doubt that (5a–b) are ESCs in P&R’s sense. In any case, whether P&R’s original examples turn out to be ESCs, there are obvious counterparts of (1) in Japanese, where two nouns are intervened by the morpheme *no*.² (8) contains a first-person pronoun and corresponds to an ESC used as self-disapprobation, while the examples in (9) contain a non-pronominal noun, and they are used as insults directed at others.

- (8) Ore/Watashi no baka/usotsuki!
 1st.sing no fool/liar
 (9) a. Oneechan/Okaasan no baka/usotsuki! (kinship terms)
 sister/mother no fool/liar
 b. Tanaka/Sacchan no baka/usotsuki! (proper names)

² The status of *no* in ESCs is not obvious at all. It might be a genitive case particle, just like *no* in a possessive NP such as *Taro no kuruma* (“Taro’s car”). It might be inserted for some other reason independent from case assignment (such as Kitagawa and Ross’s (1982) Mod-Insertion rule). See (Watanabe 2010) and citations therein. The following discussion does not depend on the nature of *no* in an ESC. We will simply gloss it as *no* below.

- c. Sensei/Shacho no baka/usotsuki! (“teacher”/“CEO”)
- d. Taifuu/Teisupe Pasokon no baka/usotsuki! (“typhoon”/“low-spec PC”) (from the Internet)
- e. Ko-no baka/usotsuki! (“this”)

First of all, the pattern emerging here—“Noun 1-*no*-Noun 2”—is highly productive, and one can be creative about the choice of Noun 2; for example, *Ko-no seijika!* (“You politician!”) would be totally fine with enough contextual information. A wide range of nominals can also appear as Noun 1. For example, (9a) has a kinship term such as “sister” or “mother” in the Noun 1 position, and (9d) shows that even a noun for an inanimate object can occur as Noun 1 (although, admittedly, the examples bear a hint of anthropomorphization).

As in (9e), the proximate demonstrative *ko* commonly appears in ESCs. The distant demonstrative *a* (“that”) is also permitted (e.g., *A-no baka/usotsuki!*). In the latter case, the addressee or the target of the insult does not have to be visibly present in the vicinity of the utterer. The intermediate demonstrative *so* (“the/that”) is, however, not permitted in an ESC.

- (10) *So-no baka/usotsuki!
 the/that-*no* fool/liar

In this paper, we will not venture into explaining these subtle differences between Japanese demonstratives.

Furthermore, we also set aside a range of variants of Japanese ESCs that end with the morphemes *me* and/or *ga*.

- (11) Ko-no baka/usotsuki me/(me)ga!
 this-*no* fool/liar *me/(me)ga*

As is the case for the morpheme *no*, *ga* here might be a case particle (nominative in this instance). If so, then it would be interesting to compare it with other case variations, such as the genitive marker in most Scandinavian languages, as discussed by Julien (2016) (e.g., *Din dust!* “your fool” in Norwegian).

In the remainder of the current paper, we assume that phrases like (8) and (9) are the counterparts of (1) in Japanese, and we will not discuss more elaborate cases such as (11).

2.2 The First Puzzle

The first puzzle regarding Japanese ESCs is that the use of a second person pronoun in the Noun 1 position is clearly degraded, as in (12), even though all other nouns that have a vocative use are acceptable in the same position, as we have seen in (9a–e) above.

- (12) ?? Anata/Omae/Kisama/Temee no baka/usotsuki!
 2nd.sing (progressively more impolite forms) *no* fool/liar

In vocatives, the second person pronouns are just as acceptable as the other nouns here.

- (13) Anata/Omae/Oneechan, kocchi kite!
 2nd.sing/2nd.sing(impolite)/Sister, here come
 “You/Sister, come here!”

What is so special, then, about the second person pronouns in ESCs? The P&R original analysis of ESCs on its own has no resource to address this puzzle. If a first person pronoun in Japanese were analyzed as an expression of type *e* to account for the acceptable case (8), then the analysis would have to predict (12) to be acceptable as well, unless one rejects the very plausible assumption that the first and second person pronouns are assigned the same basic semantic type.

One may suggest that some idiosyncratic syntactic property of the Japanese pronominal system deems examples like (12) illicit. Such a syntactic restriction is, however, not in the offing. Note, first, that (12) is not obviously ungrammatical—it merely sounds odd—and second, that there is no independently motivated restriction on the person of subjects in Japanese that would exclusively rule out second person pronouns. Some predicates of direct experience in Japanese are known to restrict the person of their subjects, but as (14) below indicates, they exclude everything but the first-person subjects (Kuroda 1973; Kuno 1973; Tenny 2006, among others).

- (14) Watashi/*Anata/*Tanaka wa kanashii yo.
 1st.sing/2nd.sing/Tanaka TOP sad part
 “I am/you are/Tanaka is sad.”

Thus, in what follows, instead of seeking a syntactic restriction on the second person pronouns in ESCs, we derive the oddity of (12) in the semantic composition of Japanese ESCs.

2.3 The Second Puzzle

The second puzzle regarding Japanese ESCs is that, unlike (3b) above, at least some of the Japanese ESCs mentioned above can be licitly used as an argument, as shown by (15). Crucially, however, such argumental ESCs lack a contextual requirement for prototypical ESCs: the addressees of ESCs need to be the very individuals picked up by the first nouns of the ESCs, as in (16a). Consider (16b), where an ESC is coupled with a vocative use of the name *Yamada*. Assuming *Yamada* and *Tanaka* are not coreferential, (16b) hardly makes sense. This is because the addressee of (16b) is fixed by the vocative *Yamada*, and that is incompatible with the conflicting addressee introduced by the ESC subject *Tanaka*. By contrast, (15) shows that the referent of an argumental ESC does not have to be the addressee of the utterance.

(15) (Addressing at Yamada)

Oi Yamada, **Tanaka no baka** ga mata shippai-shita yo.
 Hey Yamada, **Tanaka no fool** NOM again mistake-did PART

“Hey Yamada, that fool Tanaka made a mistake again.”

- (16) a. Oi Tanaka, **Tanaka no baka!**
 Hey Tanaka, **Tanaka no fool**
 b. # Oi Yamada, **Tanaka no baka!**
 Hey Yamada, **Tanaka no fool**

The P&R analysis of ESCs is incompatible with (15) because an argumental ESC is precisely what their semantic system is designed to exclude (recall that according to P&R, an ESC overall yields a meaning of type *E*, which cannot be an argument of a sentence).

One may suggest that *baka* (“fool”) is ambiguous between the expressive and descriptive interpretations ($\langle e, E \rangle$ vs $\langle e, t \rangle$), and *Tanaka no baka* in (15) is not an ESC after all (it is a noun of some sort). Although the underlying intuition here points to the right direction (as we will see below), merely appealing to ambiguity leads to more problems. If *baka* is ambiguous between the expressive and descriptive interpretations, and either denotation can be freely chosen, then why can’t we say the same for *fool* in English? For P&R, *fool* is ambiguous to begin with, but the phrase *you fool* can only have an expressive interpretation. What, then, makes *baka* in (15) so different from *fool* in (3b) that an argumental ESC is permitted only in the former? There must be a principled explanation for the differences between (15) and (3b).

To summarize, Japanese ESCs differ from their English counterparts in that, first, second person pronouns are not allowed in Japanese ESCs, and second, an argumental ESC is sometimes allowed in Japanese, but not in English. In what follows, we will propose to solve these two puzzles in one fell swoop by making modifications to the P&R analysis of ESCs.

3 (Dis)solving the Puzzles

3.1 Our Proposal

To account for the two puzzling phenomena presented in the previous section, we would like to propose a modified analysis of ESCs that retains P&R’s basic insight—that ESCs are expressives. Our analysis is designed to encompass the following two ideas.

- (17) What P&R have dubbed ESCs are not literally “small clauses” consisting minimally of the subject and the predicate, but in fact they have more structure than meets the eye.

- (18) The locus of the expressive property of an ESC is not an epithet noun, such as *fool*, in itself; rather, its expressive meaning is derived with the help of another syntactic head located above the epithet noun.

(17) is, we think, not particularly contentious. Syntacticians such as Julien (2016) and Corver (2008) cross-linguistically examine ESCs, and they both propose that there is an independent syntactic head that is responsible for the predication relation between the subject and the predicate of an ESC.³ To give a concrete example, according to Julien (2016), the Norwegian ESC *din lille fjott* (“you little dork”) is analyzed as having a head called “Pred,” which semantically connects “you” and “little dork” by means of predication; an ESC is not a mere concatenation of two nominals.

Other scholars also deny the small-clause status of ESCs (Arsenijević 2007; d’Avis and Meibauer 2013) and treat them as a class of vocative constructions. A strong piece of evidence comes from the Serbo-Croatian counterparts of (1).

- (19) a. E, budal-o!
oh, fool.VOC
- b. Idiot-e!
idiot.VOC
- c. E, Boban-e, budal-o!
oh, Boban.VOC fool.VOC
- d. E, Boban-e, idiot-e!
oh, Boban.VOC idiot.VOC

(Arsenijević 2007, pp. 91-2)

In (19), the nominal epithets occur in the vocative case. If one is willing to accept that vocative constructions involve some functional structure (e.g., Hill 2007; Portner 2007), then ESCs in Serbo-Croatian, at least, contain functional structure.

(18) is concerned with what we have pointed out in Sect. 2.3. Simply regarding epithet nouns as lexically ambiguous between descriptive and expressive interpretations would fail to capture the above-mentioned differences between Japanese and English ESCs. We thus must attribute the source of the expressive meaning to something else. We identify it with a higher syntactic head in the structure.

We also assume that Japanese is an NP language (Bošković 2012; Izumi 2011; Takahashi 2011, among many others), where NPs on their own can be an argument of a sentence without being headed by some functional element such as a D. This assumption, together with the view that an ESC includes a predicational structure, implies that English and Japanese ESCs are structurally different, as represented by (20).

³ Since they are not “small clauses” in the usual sense, Julien and Corver drop the term ESC altogether—ESCs are “possessive predicational vocatives” in Julien’s terminology and “evaluative vocatives” in Corver’s. In this paper, we do not presume ESCs to be a class of vocative constructions, and so we stick to P&R’s terminology.

(22) $\llbracket \text{anata} \rrbracket^c = \lambda x.x \text{ is } c_A$

(23) $\llbracket \text{Tanaka} \rrbracket^c = \lambda x.x \text{ is called "Tanaka"}$

Now, the derivation for *You fool!* based on the proposed structure (20a) is the same as the original P&R analysis (2b), except that *fool* is unambiguous and is first turned into an expressive predicate by E-Pred.

The derivation for the Japanese ESC (20b) proceeds in a parallel fashion:

(24) a. $\llbracket \text{Tanaka no baka} \rrbracket^c = \lambda x.x \text{ is called "Tanaka" and a fool (via Predicate Modification)}$

b. $\llbracket \text{E-Pred Tanaka no baka} \rrbracket^c = \lambda x.\{c' : x \text{ is } c'_A (= c_A) \text{ and } c'_S (= c_S) \text{ considers } c'_A \text{ to be unfavorably describable as being called "Tanaka" and a fool}\}$ (via Function Application)

c. $\llbracket \text{pro}_1 \text{ E-Pred Tanaka no baka} \rrbracket^c = \{c' : \text{the individual referred to by } \text{pro}_1 \text{ is } c'_A (= c_A) \text{ and } c'_S (= c_S) \text{ considers } c'_A \text{ to be unfavorably describable as being called "Tanaka" and a fool}\}$ (via Function Application and Traces and Pronouns for pro_1)

Thus, apart from the additional nominal predicate *Tanaka*, the English and Japanese ESCs (20a–b) are analyzed as having the same expressive meaning: the speaker expresses an insulting attitude toward the addressee.

One may worry that, by presenting the second step as (24b), we are claiming that the addressee of (20b) is insulted by being called “Tanaka.” This outcome would be incongruous with an intuitive reading of (20b). That is not what (21) implies, however, because it states that the addressee is insulted by being referred to as the whole NP, not just its parts (a fake diamond is not a diamond). Also, note that the first noun in an ESC could contribute a negative, derogatory meaning to the whole NP, as in the “low-spec PC” example in (9d). Admittedly, the assumption we make for (24a) is simplistic, ignoring the NP internal structure and the possible contribution of the morpheme *no* to the overall NP meaning. To be precise about the internal semantics of NPs, we would need to specify a theory of $\text{NP}_1\text{-no-NP}_2$ in Japanese, which is not an easy task, to say the least. The simplistic assumption is adequate for our present purposes.

3.2 A Solution to the First Puzzle

Applying the same derivation as (24) to the unacceptable (12), we obtain:

(25) $\llbracket \text{pro}_1 \text{ E-Pred Anata no baka} \rrbracket^c = \{c' : \text{the individual referred to by } \text{pro}_1 \text{ is } c'_A (= c_A) \text{ and } c'_S (= c_S) \text{ considers } c'_A (= c_A) \text{ to be unfavorably describable as being } c_A \text{ and a fool}\}$

This should sound odd, if not incoherent, because $c'_A (= c_A)$ is described **as** $c_A (= c'_A)$, namely, the addressee of the context is described **as the addressee of the context**—a redundant way of describing someone. We claim that this type of redundancy explains why a second person pronoun in Japanese ESCs is almost unacceptable but not obviously ungrammatical.

Put differently, an ESC always targets the addressee of the context (though the addressee is sometimes the speaker herself), and so a Japanese second person pronoun, which behaves like a predicate and thereby contributes the semantic information of “being the addressee” to the overall meaning, would be redundant in an ESC.

3.3 A Solution to (or a Dissolution of) the Second Puzzle

Since NPs on their own can occur as arguments in Japanese, the NP *Tanaka no baka* (“Tanaka *no* fool”) is licitly used as an argument; in other words, (15) in fact contains no genuine ESC, and the lack of E-Pred explains the lack of an expressive meaning in (15). On the other hand, in the case of English, neither *you fool* nor *fool* is permitted in argument position, because the former only has an expressive meaning of type *E* and the latter is a mere NP, which must be headed by a D to be an argument in DP languages.

By offering this solution, we can be seen as dissolving the second puzzle regarding ESCs, rather than solving it, because we deny that (15) has an ESC, understood as an expressive phrase, concurring with P&R that no ESC can be an argument of a sentence. Our account is an improvement over P&R’s insofar as it captures the differences between English and Japanese ESCs, while explaining why some phrases of the form NP₁ *no* NP₂ are not really ESCs.

One remaining issue is that personal pronouns do not permit a use analogous to (15), repeated here as (26).

- (26) **Tanaka no baka** ga mata shippai-shita yo.
Tanaka no fool NOM again mistake-did PART
 “That fool Tanaka made a mistake again.”
- (27) ?? **Omae no baka** ga mata shippai-shita yo.
You no fool NOM again mistake-did PART
- (28) ?? **Ore no baka** ga mata shippai-shita yo.
I no fool NOM again mistake-did PART

Our proposal above simply assumes that the NP₁-*no*-NP₂ series is a mere NP, which can be an argument of a sentence, and so it fails to explain why (27–28) only have possessive readings—for example, *omae no baka* in (27) could refer to “your idiocy,” but not the addressee herself.

We do not currently have a good solution to this problem. The place to begin would be to examine closely the internal structure of NP₁-*no*-NP₂ in (26). It may turn out that the two NPs are in the appositive relation, and also that personal pronouns cannot be modified by appositives. The following contrast may encourage research in this direction.

- (29) a. Taiyouou Louis-wa ...
 Sun.king Louis-TOP ...
 “Louis, the Sun King, ...”
- b. ??/* Taiyouou anata/watashi-wa ...
 Sun.king you/I-TOP ...
 “You/I, the Sun King, ...”

If (29a) is an instance of nominal apposition, (29b) indicates that personal pronouns are not always acceptable in the same environment. We leave further discussion of this issue for future research.

4 Concluding Remarks

We have presented two puzzling phenomena involving Japanese slurring ESCs and proposed solutions to the puzzles by modifying the P&R analysis of English ESCs. The proposal is mainly based on the NP/DP language distinction, so future studies must examine if there is a systematic difference between ESCs in NP and DP languages. Another research question is how to understand the internal structure of NPs in Japanese ESCs—What is the relation between the first noun and the second noun? What is the nature of the morpheme *no*? Whether it turns out to be empirically tenable, our proposal opens up many research questions concerning the structure and meaning of nominal expressions.

References

- Arsenijević, B.: Disapprobation expressions are vocative epithets. In: ACLC Working Papers, vol. 2, no. 2, pp. 87–98 (2007)
- Bošković, Z.: Phases in NPs/DPs. In: Gallego, Á.J. (ed.) *Phases: Developing the Framework*, pp. 343–383. Mouton de Gruyter, Berlin (2012)
- Bowers, J.: The syntax of predication. *Linguist. Inq.* **24**(4), 591–656 (1993)
- Burge, T.: Reference and proper names. *J. Philos.* **70**(14), 425–439 (1973)
- Corver, N.: Uniformity and diversity in the syntax of evaluative vocatives. *J. Comp. German. Linguist.* **11**, 43–93 (2008)
- Croom, A.M.: Slurs. *Lang. Sci.* **33**, 343–358 (2011)
- d’Avis, F., Meibauer, J.: *Du Idiot! Din idiot!* Pseudo-vocative constructions and insults in German (and Swedish). In: Patrizia, B.S., Hanna, N.A. (eds.) *Vocative!: Addressing Between System and Performance*, pp. 189–217. Mouton de Gruyter, Berlin (2013)
- Elbourne, P.D.: *Situations and Individuals*. The MIT Press, Cambridge (2005)
- Fara, D.G.: Names as predicates. *Philos. Rev.* **124**, 59–117 (2015)
- Greenawalt, K.: *Fighting Words: Individuals, Communities, and Liberties of Speech*. Princeton University Press, Princeton (1995)
- Gutzmann, D.: *Use-Conditional Meaning: Studies in Multidimensional Semantics*. Oxford University Press, Oxford (2015)

- Hill, V.: Vocatives and the pragmatics-syntax interface. *Lingua* **117**(12), 2077–2105 (2007)
- Izumi, Y.: Interpreting bare nouns: type-shifting vs. silent heads. In: *The Proceedings of the 21st Semantics and Linguistics Theory*, pp. 481–494 (2011)
- Izumi, Y.: *The semantics of proper names and other bare nominals*. Ph.D. thesis, University of Maryland, College Park (2012)
- Julien, M.: Possessive predicational vocatives in Scandinavian. *J. Comp. German. Linguist.* **19**, 75–108 (2016)
- Kitagawa, C., Ross, C.N.G.: Prenominal modification in Chinese and Japanese. *Linguist. Anal.* **9**, 19–53 (1982)
- Kuno, S.: *The Structure of the Japanese Language*. The MIT Press, Cambridge (1973)
- Kuroda, S.Y.: Where epistemology, style and grammar meet: a case study from Japanese. In: Kiparsky, P., Anderson, S.R. (eds.) *A Festschrift for Morris Halle*, pp. 377–391. Holt, Rinehart & Winston (1973)
- Langton, R.: Beyond belief: pragmatics in hate speech and pornography. In: Maitra, I., McGowan, M.K. (eds.) *Speech and Harm: Controversies over Free Speech*, pp. 72–93. Oxford University Press, Oxford (2012)
- Matushansky, O.: On the linguistic complexity of proper names. *Linguist. Philos.* **31**(5), 573–627 (2008)
- Noguchi, T.: Two types of pronouns and variable binding. *Language* **73**(4), 770–797 (1997)
- Portner, P.: Instructions for interpretation as separate performatives. In: Schwabe, K., Winkler, S. (eds.) *On Information Structure, Meaning and Form: Generalizations Across Languages*, pp. 407–425. John Benjamins Publishing Company, Amsterdam (2007)
- Potts, C.: *The Logic of Conventional Implicatures*. Oxford University Press, Oxford (2005)
- Potts, C.: The expressive dimension. *Theor. Linguist.* **33**, 165–198 (2007)
- Potts, C., Roeper, T.: The narrowing acquisition path: from declarative to expressive small clauses. In: Progovac, L., Paesani, K., Caselles-Suárez, E., Barton, E. (eds.) *The Syntax of Nonsententials: Multi-Disciplinary Perspectives*, pp. 183–201. John Benjamins, Amsterdam (2006)
- Predelli, S.: *Meaning Without Truth*. Oxford University Press, Oxford (2013)
- Stanley, J.: *How Propaganda Works*. Princeton University Press, New Jersey (2015)
- Takahashi, M.: *Some theoretical consequences of case-marking in Japanese*. Ph.D. thesis, University of Connecticut, Storrs (2011)
- Tenny, C.L.: Evidentiality, experiencers, and the syntax of sentience in Japanese. *J. East Asian Linguist.* **15**, 245–288 (2006)
- Watanabe, A.: Notes on nominal ellipsis and the nature of *no* and classifiers in Japanese. *J. East Asian Linguist.* **19**, 61–74 (2010)



Pictorial and Alphabet Writings in Asymmetric Signaling Games

Liping Tang^(✉)

Institute of Logic and Cognition, Department of Philosophy, Sun Yat-sen University,
Guangzhou, China
tanglp3@mail.sysu.edu.cn

Abstract. The goal of the paper is to differentiate the pictorial writings and alphabet writings (i.e. Chinese vs English) in signaling games. The idea is to emphasize the asymmetry of communicators' different cognitive statuses while they are writing and reading by using these two types of languages. For fulfilling the idea, we study two variations of the standard evolutionary signaling game in order to incorporate players' asymmetry on recognizing and learning signals. It is found that pictorial representation takes more advantages in the early stage of the development of a language, while alphabet representation works better in the long run.

Keywords: Pictorial and alphabet writings · Cognitive asymmetry
Signaling game · Reinforcement learning

1 Introduction

Language is the most important tool for communication. One of the significant differences among languages is the feature of pictorial form vs the alphabet form in the writing system. The typical example of the pictorial language is Chinese and the typical example of the alphabet language is English. Various discussions on the differences of these two types of language can be found in cognitive linguistics and anthropology literature [4, 6, 11, 15, 19].

In this paper, we investigate the differences of pictorial and alphabet language from a novel perspective. Treating written language as a tool of communication, there is an asymmetry between writing and reading with respect to the writer's and the reader's cognitive statuses. As Smith [17] stated, "There are radical differences between the skills and knowledge employed in reading and those employed in writing, just as there are considerable differences in the processes involved in learning to read and in learning to write. . . . A reasonable working hypothesis is that anything that tends to make writing easier will make reading more difficult, and vice versa."

Pictorial words, such as Chinese, normally take the form of pictures. Those pictures usually reflect fully or partially the meanings of the words, therefore, easy to read. Consequently, pictorial words, on the contrary, often have a complex

form. Therefore, according to Smith's hypothesis, pictorial words are easy for the reader to use but hard for the writer to use. Hence the cognitive burden of using the pictorial language goes more to the writer. A similar argument applies to the differences between a writer and a reader on learning a pictorial word, that is, the pictorial word is harder for the writer to learn but easier for the reader to learn.

In contrast, alphabet language, i.e. English, has the opposite feature. Since the meaning of an alphabet word is usually not embedded in the written form of the word, therefore, the meaning of the word is not apparent to the reader. But it is easy for the writer to use alphabet words because all the components of the alphabet words are from a small amount of letters. For example, all English words are formed from 26 letters. In conclusion, alphabet words are easier for the writer but more difficult for the reader to learn and use. The goal of the paper is to model the asymmetry between a writer and a reader on *using* and *learning* pictorial language and alphabet language through a game theoretical model, *Signaling Game*.

Signaling game captures a general communication scenario. There are two players in this game, a sender and a receiver. Sender sends a signal to the receiver conditioned on the information he observes. By receiving the signal, the receiver needs to take an act that decides both players' payoffs.

In this game, players have different roles in terms of the "use" of the signals. Sender's task is to send a signal¹ based on the information revealed to him. Receiver's job is to take an act by reading the signal. However, in classical signaling games, the sender and the receiver are treated equally on using the signals even though the sender plays the role of a writer and the receiver plays the role of a reader. It is because the focus of traditional signaling games is on the result of the communication. Communication is treated as a coordination interaction [8–10, 16]. It is more natural to treat players's efforts equally in terms of an efficient communication. Nevertheless, as we have discussed before, there are radical differences between a writer and a reader on using a language. Thus, for applying the signaling game to analyze the asymmetry between writing and reading in pictorial and alphabet languages, modifications of the standard signaling game are necessary.

Given a sender and a receiver's different cognitive burdens on using a language, pictorial representation and alphabet representation reflect distinct features. Two assumptions are made below based on previous analysis.

1. Cognitive perspective: pictorial words are easier for the receiver to read but harder for the sender to write; alphabet words are easier for the sender to write but more difficult for the receiver to read.
2. Learning perspective: pictorial words are faster for the receiver to learn than the sender; alphabet words are faster for the sender to learn than the receiver.

These two assumptions reflect the main distinctions of pictorial and alphabet words that we are interested in. An evolutionary game theoretical model is

¹ In this paper, we focus on writings.

provided to incorporate these two assumptions and show the differences of pictorial and alphabet representations. Through a couple of simulations, we show that the pictorial representation takes more initial advantages while the alphabet representation works better in the long run. The model we use here is the evolutionary signaling game with the reinforcement learning dynamics. Details about this formal model is discussed in the following sections.

The rest of the paper is organized in the following way. Section 2 is a brief introduction to Lewis's signaling game. In Sect. 3, we present the basic reinforcement learning rule and two modified versions of the original reinforcement learning model, in which the asymmetry between the sender and the receiver is characterized. In Sect. 4, we present the simulation results based on the two modified evolutionary signaling games. Paper ends with a short conclusion.

2 Lewis's Signaling Game

Lewis's signaling game describes a very general communication scenario where a sender observes the situations and sends a signal to a receiver. The receiver takes actions based on the signals he receives. Players have common interests that are decided by receiver's actions. According to Lewis, the simplest form of signaling game includes two states, two signals and two acts. Nature decides the state, the sender chooses a signal to send and the receiver takes an action according to the signal. The underlining payoff structure of this game takes the form in Table 1. It shows that when S_i is the true state and Act_i is chosen, both players get payoff 1. Otherwise, players get 0. According to this payoff structure, players have the same payoff function. Thus, players are willing to cooperate as much as possible and share all the available information.

Table 1. Payoff for players

	S_1	S_2
Act_1	1,1	0,0
Act_2	0,0	1,1

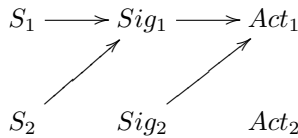
Meanwhile, being different from general game models, signaling game has more than one types of equilibrium with respect to various communication mechanisms. These equilibria are *Separating Equilibrium* and *Pooling Equilibrium*. Firstly, a signaling game can have a separating equilibrium under which signals carry complete information about the states. For example,

$$S_1 \longrightarrow Sig_1 \longrightarrow Act_1$$

$$S_2 \longrightarrow Sig_2 \longrightarrow Act_2$$

When S_1 occurs, the sender sends Sig_1 . By receiving Sig_1 , the receiver takes Act_1 . Similarly, When S_2 occurs, the sender sends Sig_2 . Players' this pair of strategies form a Nash equilibrium. Under this equilibrium, all the information about the states is communicated successfully between the sender and the receiver.

Moreover, the game can have a pooling equilibrium under which signals don't carry information about the states. The sender always sends the same signal for each state and receiver takes the same action all the time. For example, the following graph represents a pooling equilibrium.



In this example, the sender sends Sig_1 no matter what nature chooses. For the receiver, he ignores the signals and takes Act_1 all the time. As a result, no precise information about S_1 and S_2 is communicated.

In recent years, there is a growing number of literature on using signaling games to study various features of language. For example, Jäger [7] has applied a so called sim-max signaling game to study the convex property of the concept space. Crawford and Sobel [2] explored the communication property while players have conflicting interests through a cheap talk game. Santana [14] uses Lewis's signaling game to explore ambiguity in language.

In this paper, based on Lewis's signaling game, we want to investigate the differences between pictorial and alphabet representations. Before stating our model, we first introduce an important tool that is commonly used in evolutionary signaling game which is called reinforcement learning. It is used to capture players' learning process of using signals.

3 Reinforcement Learning in Signaling

3.1 Standard

Reinforcement learning is well known in behavioral psychology that describes a strengthen effect that an organism's future behavior is preceded by a specific antecedent stimulus. The strengthening effect can be measured as a higher frequency of behavior, a longer duration of time and so on. This learning model, due to its generality, has been widely applied in the fields of machine learning, game theory and information theory.

Many evolutionary signaling studies [8,12,13] have applied reinforcement learning rule as their basic learning method because it is not only psychologically natural but also assumes little cognitive abilities from the agents. Considering the structure of the reinforcement learning, it is equivalent to a simple urn model. Hoppe [5] introduced what he called "Polya-like urns" in 1984. In a

classic Polya urn process, we start with an urn containing various colored balls. Then we proceed in the following way. A ball is drawn at random from the urn. Then it is returned to the urn with another ball having the same color. All colors are treated in the same way. As the process continues, some color balls become more superior than other ones within the urn.

When applying reinforcement learning to the classical signaling game with two states, two signals, two acts and players get payoffs according to Table 1, the process can be illustrated as follows. The sender has two urns, one for S_1 and one for S_2 (see Fig. 1 for an illustration). Each urn contains balls of two colors. One color represents Sig_1 , the other represents Sig_2 . When the sender is informed of the state, she takes the appropriate urn and draws a ball. The color of the ball determines what signal is sent. If the receiver takes the appropriate action, then the sender returns her ball to her urn and adds another of the same color to the same urn. On the other hand, if the receiver does not take the appropriate action, the sender returns her ball, but does not add another.

The receiver's strategy is similar. He has two urns, one for Sig_1 and the other for Sig_2 . Each urn contains balls of two colors, one color represents Act_1 and the other Act_2 . If the receiver's act is appropriate for the state, he adds an extra ball to the appropriate urn, otherwise the urn is unchanged.

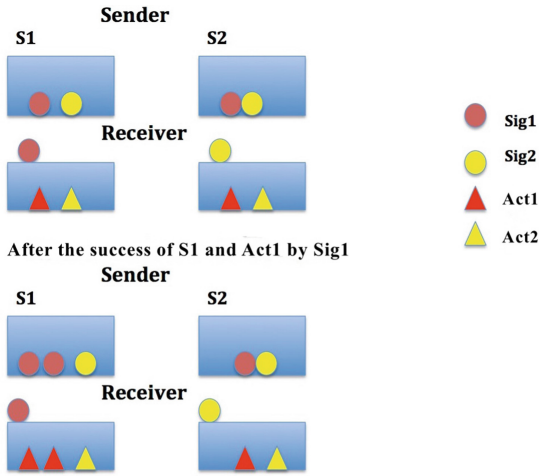


Fig. 1. An illustration of reinforcement learning in the two state, two signal, two act Lewis signaling game.

In Fig. 1, round balls indicate signals, triangle ones represent receiver's actions. At the initial stage, in state S_1 , Sig_1 and Sig_2 have equal probability to be chosen (the top picture in Fig. 1). The probability of Sig_1 in S_1 can be calculated by $1/(1 + 1) = 0.5$. Suppose nature chooses S_1 and the sender draws the red ball. As a result, Sig_1 is sent. Suppose the receiver draws the red triangle

from his Sig_1 urn. As a result, Act_1 is chosen by the receiver. Since the receiver took the correct action, one more red ball and red triangle should be added to S_1 urn and Sig_1 urn respectively. As a result, in the next stage, if S_1 occurs, the probability of Sig_1 to be chosen increases to $2/(1+2) = 2/3$ and the probability that act Act_1 is chosen when Sig_1 is sent increases to $2/(1+2) = 2/3$.

This learning process can be written in a formal way. Formally, let $S_i, i = 1, 2, \dots, n$ be all the states, $Sig_u, u = 1, 2, \dots, n$ be all the signals, $Act_j, j = 1, 2, \dots, n$, be all the acts. $w_{ui}(t), u = 1, 2, \dots, n$ are the weights for signal u for state i at time t . $p_{ui}(t)$ is the response probability of signal u for State i at time t for the sender. Suppose the initial weights $w_{ui}(0) = 1$ for all u and i . Then the updating rule is the following.

$$w_{ui}(t+1) = \begin{cases} w_{ui}(t) + 1 & \text{if } S_i \text{ occurs, } Sig_u \text{ is sent,} \\ & \text{and } Act_i \text{ is taken;} \\ w_{ui}(t) & \text{Otherwise.} \end{cases}$$

The response rule is

$$p_{ui}(t) = \frac{w_{ui}(t)}{\sum_v w_{vi}(t)} \text{ for } S_i \text{ at time } t$$

Similarly, the updating rule and the response rule for the receiver are the followings.

The updating rule is

$$w_{\alpha ui}(t+1) = \begin{cases} w_{\alpha ui}(t) + 1 & \text{if } S_i \text{ occurs, } Sig_u \text{ is sent,} \\ & \text{and } Act_i \text{ denoted as } \alpha \text{ is taken;} \\ w_{\alpha ui}(t) & \text{Otherwise.} \end{cases}$$

The response rule is

$$p_{\alpha ui}(t) = \frac{w_{\alpha ui}(t)}{\sum_{\beta} w_{\beta ui}(t)} \text{ for } S_i, Sig_u \text{ and } \alpha \text{ at time } t$$

The reinforcement learning signaling model has been widely applied in the studies of language, such as the works by Skyrms [16], Alexander et al. [8] and O'Connor [12, 13]. Variations of the standard models have been produced to accommodate special purposes. In Alexander et al. [8]'s paper on inventing new signals. A mutant signal is introduced to the original set of signals. Once the mutant signal is triggered, a completely new signal is invented in addition to the old set of signals. In O'Connor's paper [12] on vagueness, a different updating rule is adopted to capture a similarity relations among states. In Tang [18]'s paper on ambiguity, a crossing-state learning rule is produced to reflect the feature of ambiguity in language. For the purpose of this paper, two variations of the standard model are applied to fulfill the requirements withinin the two assumptions in Sect. 1.

3.2 Variations

Different Initial Weights. In probability theory, it is common to use uniform distribution to represent ignorance. This idea is exactly used in the standard reinforcement learning signaling game. Recalling the definition of the reinforcement dynamic, the initial weight $w_{ui}(0) = 1$ for all u and i . In other words, at the initial stage of the learning process, the probability of each signal being chosen by the sender is a uniform distribution (the prior probability). Similarly, for the receiver, the prior probability of each act being chosen is also a uniform distribution. Namely, players are indifferent among all the signals and acts and no player has an incentive to choose any particular signal or act. However, it is more natural to give players different prior probabilities in order to indicate the cognitive asymmetry between the sender and the receiver on using the same signal. As Smith [17] stated, reading and writing (learning to read and learning write) employ radical differences in knowledge and skills. Therefore, we reached the first criterion on how to model these differences:

Criterion 1: assign different players different prior probabilities on the signals (acts).

The next question is how to assign the initial weights such that the prior probability characterizes the feature in assumption 1. On the one hand, the same initial weight yields the uniform prior probability under which players have no knowledge about the signal(act). On the other hand, point distribution indicates the full knowledge. Besides the two extreme cases, there are all kinds of distributions sitting between these two and representing various degrees of players' cognitive statuses. According to Assumption 1, it is easier for the receiver to read the pictorial word than the sender to write the word. Therefore, the prior probability for the receiver should not be the uniform distribution if uniform distribution is assigned the sender. For achieving this goal, the easiest approach is to assign different initial weights to acts for the receiver such that receiver's prior probability carrying more information than the uniform distribution. In contrast, in the model for the alphabet word, it is easier for the sender than for the receiver in terms of the use of the alphabet word in a communication. Formally, the closer to 0 or 1 the probability is the more precise the probability is. Combing all the previous analysis, we can reach the second and third criterion:

Criterion 2: in the signaling model for pictorial representations, the prior probability for the receiver should be more precise than that of the sender; in the signaling model for alphabet representations, the prior probability for the sender should be more precise than that of the receiver.

Criterion 3: for achieving Criterion 2 of assigning different initial weights for different signals (acts), we should assign the initial weights by the following rule "the farther the weights apart from each other, the more precise the prior probability is".

For example, in a two-player Lewis signaling game. There are two states S_1, S_2 , two acts Act_1, Act_2 and two signals Sig_1, Sig_2 . In the reinforcement learning based on this simple game, the signal variables are conditioned on the states, and the act variables are conditioned on the states and the signals. Therefore, the number of the signal variable is 4 and the number of the act variables is 8. Suppose in the model for the pictorial word, we assign the initial weight for the sender to be $w_{ui}(0) = 1, u = 1, 2; i = 1, 2$. Thus, sender has a uniform prior distribution on the set of signals conditioned on the states. The initial weight for the receiver is $w_{1ui}(0) = 1$ for all $u = 1, 2; i = 1, 2$ and $w_{2ui}(0) = 2$ for all $u = 1, 2; i = 1, 2$. As a result, at each state and given each signal, the prior probability of choosing act 1 for the receiver is $1/3$ while that probability for act 2 is $2/3$. It means that the receiver incentives to choose act 2 much more often than act 1 at even the early stage of learning. The inherent reason for this intention comes from the receiver's easier "use" of the pictorial word.

Under the model of assigning different initial weights to signals or acts, we have developed simulations to test the differences between the pictorial writings and the alphabet writings. It is found that:

Under the modified reinforcement learning model where the initial weights are different, simulations show that the pictorial word model is better than the alphabet model in terms of the average utilities.

In other words, in the origin of language, pictorial word takes more advantages when only the initial advantages are emphasized. This result is consistent with the findings in Anthropology. In the history of language, not only the pictorial language takes the form of pictures, even the alphabet language takes the form of pictures in the early development of the language. The detailed simulation results are presented in Sect. 4.

Different Learning Speeds. The model with different initial weights only captures one aspect of the cognitive asymmetry between the sender and the receiver on perceiving different types of signals. As assumption 2 states, from the learning's perspective, pictorial word and alphabet word can be distinguished by players' different learning speeds. For characterizing different learning speeds within the dynamic signaling model, a logistic response rule is applied to the dynamic rule. The basic idea is to introduce a parameter λ into players' response rules. As the value of λ varies, different learning speeds can be described. Nevertheless, in the new model, the updating rules keep the same as in the standard model while only players' response rules are modified.

The updating rule for the sender keeps the same as in the standard model in the following form:

$$w_{ui}(t + 1) = \begin{cases} w_{ui}(t) + 1 & \text{if } S_i \text{ occurs, } Sig_u \text{ is sent} \\ & \text{and } Act_i \text{ is taken;} \\ w_{ui}(t) & \text{Otherwise.} \end{cases}$$

The updating rule for the receiver also keeps the same as in the standard model in the following form:

$$w_{\alpha}ui(t+1) = \begin{cases} w_{\alpha}ui(t) + 1 & \text{if } S_i \text{ occurs, } Sig_u \text{ is sent} \\ & \text{and } Act_i \text{ denoted as } \alpha \text{ is taken;} \\ w_{\alpha}ui(t) & \text{Otherwise.} \end{cases}$$

The response rules for the players reflect the key differences between our model and the classical reinforcement model. As we introduced briefly, the parameter λ is introduced into players' response rules to characterize players' different learning speeds. The way to realize this purpose is to apply a so called logistic response rule. Sender's new response rule takes the following form.

$$p_{ui}(t) = \frac{e^{\lambda w_u i(t)}}{\sum_v e^{\lambda w_v i(t)}} \text{ for } Sig_u \text{ at } S_i \text{ at time } t$$

Similarly, for the receiver, the logistic response rule is the following.

$$p_{\alpha}ui(t) = \frac{e^{\lambda w_{\alpha} ui(t)}}{\sum_{\beta} e^{\lambda w_{\beta} ui(t)}} \text{ for } Sig_u \text{ at } S_i \text{ and act } \alpha \text{ at time } t$$

In both formulas, the parameter λ represents the degree of "smoothness" of the function. The higher λ the more small difference affects the probability. As λ becomes larger, small difference in past payoffs correspond to greater differences in response probabilities². From the perspective of learning, more sensible to small changes in past payoffs means the learning is relatively fast. Therefore, the higher λ the higher learning speed can be captured.

Recalling Assumption 2 about learning, we get criterion 4.

Criterion 4: in the logistic learning model for pictorial word, higher λ should be assigned to the receiver; in the logistic learning model for alphabet word, higher λ should be assigned to the sender;

By criterion 4, simulations are operated to test the model. The general results are the following:

In the model where players have different learning speeds. The case where sender learns faster (alphabet model) is preferred than the case when the receiver learns faster (pictorial model). The intuition for this result is that from the language learning's perspective, alphabet word model gets more benefits. It explains the reason why alphabet language is easier for adopting and spreading than the pictorial language with respect to the number of and the complexity of the characters in a language [6, 15].

² This response rule is used in Zollman's work [1] to study forgetting in evolutionary games.

4 Simulation Results

4.1 Simulation 1 (Different Initial Weights)

Simulation 1 is operated based on the different initial weight model. Applying criterion 2 and criterion 3, different initial weights should be assigned to sender or receiver in different scenarios. In the pictorial model of two states, two signals and two acts, the sender has the same initial weight for all the signals, while the receiver has an initial weight assignment of 8 and 2 for his two acts. In contrast, in the alphabet model, the receiver has the same initial weight for all the acts, but the sender has an initial weight assignment of 8 and 2 for his two signals. The simulation is operated for 500, 10000, and 100000 iterations³. The probability assigned on the set of states is 0.2 and 0.8. Suppose the payoff is 1 for both players if the coordination is successful, then the utilities are collected in Table 2.

Table 2. Average utilities for the model with different initial weights I

Number of iterations	Pictorial model (Receiver’s weights: 8 vs 2 Sender’s weights: 5 vs 5)	Alphabet model (Sender’s weights: 8 vs 2 Receiver’s weights: 5 vs 5)
500	0.8829	0.8382
10000	0.9755	0.8905
100000	0.9836	0.8918

The following observations can be found from Table 2:

- Pictorial model is better than alphabet model in terms of players’ utilities all the time;
- The advantage of pictorial model is stable as the learning continues.

In comparison with the case of the initial weight assignment 8 vs 2. We tested a different weight assignment 6 vs 4 in the same game. Namely, in the pictorial model, receiver has the initial weights 6 and 4 for his two acts while the sender has the same initial weights. In contrast, in the alphabet model, sender has the initial weights 6 and 4 for his two signals while the receiver has the same initial weights on the acts. The simulation results for those two models are listed in Table 3.

Table 3 shows that the quality feature of the pictorial model and alphabet model stays the same as the results in Table 2, namely, the pictorial model behaves better than the alphabet model. The distinction is that the overall utilities in the 8 vs 2 case is better than the 6 vs 4 case. According to our previous

³ The assumption of a different probability on the state set is essential. With the same probability on the states, the behavior of the two models are almost the same.

Table 3. Average utilities for the model with different initial weights II

Number of iterations	Pictorial model (Receiver's weights: 6 vs 4 Sender's weights: 5 vs 5)	Alphabet model (Sender's weights: 6 vs 4 Receiver's weights: 5 vs 5)
500	0.8738	0.8247
10000	0.9418	0.8646
100000	0.9503	0.8823

analysis, the larger difference the initial weights is the more extreme the prior probability on the signal (act) is. As a result, the player who holds the extreme probability gets more advantages in the learning process as if this players has already learned how to use the signal or act in advance to some extent.

4.2 Simulation 2 (Different Learning Speeds)

In simulation 2, we apply the logistic response rule in the reinforcement learning model such that players' different learning speeds can be characterized by different λ s in the response rule. The higher λ means the faster leaning. The basic signaling game is the same as in simulation 1. A simulation is operated for 500, 1000 and 8000 iterations. In the pictorial model, λ is set to be 1 and 5 for the sender and the receiver respectively, that is, we assume the receiver learns faster in the pictorial model than the sender. On the other hand, in the alphabet model, the opposite is assumed. λ for the receiver is 1 and for the sender is 5 such that the sender learns faster in the alphabet model.

Table 4 reflects the simulation results.

Table 4. Average utilities for the model with different learning speeds I

Number of iterations	Pictorial model (Receiver $\lambda = 5$ Sender $\lambda = 1$)	Alphabet model (Sender $\lambda = 5$ Receiver $\lambda = 1$)
500 iterations	0.7997	0.7685
1000 iterations	0.8	0.8917
8000 iterations	0.8041	0.99

Observations from Table 4 are the followings:

- In the short length of learning (500 iterations), the two models are competing in terms of players' utilities. Moreover, the pictorial model is even slightly better than the alphabet model;
- Alphabet model is better than alphabet model in the long run in the different learning speed model;
- The advantage of alphabet model is stable as the learning continues.

These results are consistent with our intuitions as well as the findings in Anthropology. Since alphabet words are built from a relatively smaller set of syllable-based symbols, comparing to pictorial words, they are easier to adopt and spread [15]. However, even the alphabet language takes the form of pictures at the origin of its development [3]. It shows that pictorial language does have advantages at the early stage of a language development. Those results on the other hand verifies the validity of our learning models.

For comparison, different λ s are also tested in the same game. Similar results can be obtained from the new simulations.

Table 5. Average utilities for the model with different learning speeds II

Number of iterations	Pictorial model (Receiver $\lambda = 5$ Sender $\lambda = 10$)	Alphabet model (Sender $\lambda = 5$ Receiver $\lambda = 10$)
500	0.8717	0.9379
1000	0.9133	0.9918
8000	0.9782	0.998

The utilities are better for both pictorial and alphabet models in Table 5 than that in Table 4. This is because in Table 5's simulation, much bigger λ is assigned to the players. As a result, players have much higher speed of learning in this simulation. Nevertheless, the same result that alphabet model behaves better than the pictorial model is sustained.

4.3 Combing Different Weights and Different Learning Speeds

In the last collection of simulations, we try to combine the model of different initial weights and the model with different learning speeds. The intuition behind this simulation is that there are languages contain both pictorial elements and alphabet elements, such as Japanese and Korean. The way to combine the two learning models is to let player both have different initial weights and different learning speeds. For example, one situation can be that the sender has a higher learning speed and the receiver has the initial weight advantage. We investigated two situations of combining the two learning models. The simulation results are in Table 6.

The result is very similar to the results in the learning model in Table 4 where the sender has the advantage on learning (alphabet model). A reasonable conjecture is that sender's learning advantage plays a more essential role in this signaling learning model. In other words, the alphabet model with different learning speeds gains more advantages in the learning process. In practice, this conjecture matches our intuition of the linguistics features in Japanese. Even

Table 6. Average utilities for the model with different learning speeds and different initial weights

Number of iterations	Pictorial model (Weight) & Alphabet model (Learning) (Receiver's weight: 8 vs 2 Sender $\lambda = 5$; Receiver $\lambda = 1$)	Alphabet model (Weight) & Pictorial Model (Learning) (Sender's weight: 8 vs 2 Sender $\lambda = 1$; Receiver $\lambda = 5$)
500	0.75	0.79
1000	0.89	0.8
8000	0.99	0.8

though Japanese has the pictorial elements, Kanji, however, the main component of the language is still dominated by the phenolic parts of Hiragana and Katakana. Nevertheless, further investigation is still required to explain this result.

5 Conclusion

The goal of the paper is to develop an evolutionary signaling model to capture the differences between the pictorial representation and the alphabet representation, i.e. Chinese vs English, from the cognitive difference between the writer and the reader. We changed the standard reinforcement learning in two different ways. One approach is to assign different initial weights to the sender and the receiver. The other is to give players different learning speeds. Under simulations in both models, we got the following main results.

Pictorial model is better than alphabet model all the time if only the initial cognitive advantages are emphasized. However, Alphabet model is better than pictorial model in the long run when the learning advantages are emphasized. Meanwhile, the advantage of alphabet model is stable as the learning continues.

Those results are intuitive and consistent with the findings in Anthropology literature. Pictorial representation is preferred in the early development stage of a language. For example, in the origin of English, all the alphabets are in the form of pictures. However, as the language develops, the alphabet language spreads more vastly which explains why so many existing languages take the form of the alphabet representation. Of course, language is such a complex process and it can not be completely captured by the signaling game model. But our work at least provides some insights about the differences between those two kinds of representations.

In terms of methodology, traditional signaling games merit a symmetric structure between the sender and the receiver and pays more attention to the cooperative perspective of the game. We developed the idea of bringing asymmetry between the sender and the receiver's behavior into the classical signaling games.

Meanwhile, the application of this method in this paper also verifies the significance of the method of evolutionary signaling game in the studies of language.

Acknowledgements. The author would like to acknowledge Kevin Zollman and Igor Yanovich whose helps and suggestions aided in the preparation of the manuscript. The author wishes to thank the editor and two anonymous reviewers for their valuable comments on the manuscript. The research reported in this paper was supported by Humanity and Social Science Youth Foundation of Ministry of Education of China (17YJC72040004).

References

1. Barrett, J., Zollman, K.J.S.: The role of forgetting in the evolution and learning of language. *J. Exp. Theor. Artif. Intell.* **21**(4), 293–309 (2009)
2. Crawford, V., Sobel, J.: Strategic information transmission. *Econometrica* **50**, 1431–51 (1982)
3. Crawford, J.: On the origin and history of written language. *Trans. Ethnol. Soc. London* **5**, 96–104 (1867)
4. Downing, J.: Is literacy acquisition easier in some languages than in others? *Visible Lang.* **7**(2), 145–154 (1973)
5. Hoppe, F.M.: Polya-like urns and the Ewens sampling formula. *J. Math. Biol.* **20**, 91–94 (1984)
6. Hung, D.L., Tzeng, O.J.: Orthographic variations and visual information processing. *Psychol. Bull.* **90**(3), 377 (1981)
7. Jäger, G., Metzger, L.P., Riedel, F.: Voronoi languages: equilibria in cheap-talk games with high-dimensional types and few signals. *Games Econ. Behav.* **73**(2), 517–537 (2011)
8. Zabell, S.L., Alexander, J.M., Skyrms, B.: Inventing new signals. *Dyn. Games Appl.* **2**(1), 129–145 (2012)
9. Zollman, K.J.S.: Talking to neighbors: the evolution of regional meaning. *Philos. Sci.* **1**, 69–85 (2005)
10. Lewis, D.: *Convention: A Philosophical Study*. Harvard University Press, Cambridge (1969)
11. Muter, P., Johns, E.E.: Learning logographies and alphabetic codes. *Hum. Learn.* **4**, 105–125 (1985)
12. O'Connor, C.: Ambiguity is kinda good sometimes. *Philos. Sci.* **82**(1), 110–121 (2014)
13. O'Connor, C.: The evolution of vagueness. *Erkenntnis* **79**(4), 707–727 (2014)
14. Santana, C.: Ambiguity in cooperative signaling. *Philos. Sci.* **81**(3), 398–422 (2014)
15. Scott, J.A., Ehri, L.C.: Sight word reading in prereaders: use of logographic vs. alphabetic access routes. *J. Reading Behav.* **22**(2), 149–166 (1990)
16. Skyrms, B.: *Signals: Evolution, Learning and Information*. Oxford University Press, New York (2010)
17. Smith, F.: Phonology and orthography: reading and writing. *Elementary Engl.* **49**(7), 1075–1088 (1972)
18. Tang, L.: Ambiguity in compositional signaling. In: Working Paper (2016)
19. Tulving, E., Thomson, D.M.: Encoding specificity and retrieval processes in episodic memory. *Psychol. Rev.* **80**(5), 352 (1973)



Collecting Weighted Coercions from Crowd-Sourced Lexical Data for Compositional Semantic Analysis

Mathieu Lafourcade¹, Bruno Mery^{2,3}(✉), Mehdi Mirzapour¹, Richard Moot¹,
and Christian Retoré¹

¹ LIRMM – UMR 5506, CNRS & Université de Montpellier, Montpellier, France
{mathieu.lafourcade, richard.moot, christian.retores}@lirmm.fr,
mehdi.mirzapour@gmail.com

² LaBRI – UMR 5800, CNRS & Université de Bordeaux, Bordeaux, France
bruno.mery@u-bordeaux.fr

³ IUT de Bordeaux, Université de Bordeaux, Gradignan Cedex, France

Abstract. Type-theoretic frameworks for compositional semantics are aimed at producing structured meaning representations of natural language utterances.

Using elements of lexical semantics, these frameworks are able to model many complex phenomena related to the polysemy of words and their context-dependent meanings. However, they are just as powerful as the lexical resources they can access. This paper explores ways to create and enrich wide-coverage, weighted lexical resources from crowd-sourced data. Specifically, we investigate how existing rich lexical networks – created and validated by serious games – can be used to infer linguistic coercions along with ranking corresponding to preferences in their interpretations.

1 Type-Theoretic Semantic Frameworks with Rich Lexical Information

The **semantic analysis** of natural language is a process that should produce a complete and structural meaning representation of a given text (such as a logical formula or a Discourse Representation Structure) that makes explicit the entities referenced in the text as well as their relationships. This is used for word sense disambiguation, resolution of co-references, natural language inference and other complex tasks.

Rich lexical information is required to compute the meaning of utterances such as *I am going to the bank*, as *bank* is ambiguous between a geographic feature and a service building. Frameworks based on theories such as [7, 22], including [1, 2, 6, 13, 23] are able to obtain a rich logical representation using a Montague-like compositional process that correctly interprets sentences such as *I am going to the bank; they blocked my account*. They not only interpret the types of the lexemes involved as indicating that *bank* is a service building, but

also that *they* is a reference to a financial institution that is introduced in the first part of the sentence, and then coerced to a relevant human agent in the second.

Such frameworks require a rich corpus of lexical resources incorporating some degree of world knowledge encoded as complex types and lexical coercions; several of these frameworks benefit from software implementations, such as [5, 15].

In these approaches, the natural language utterance is analysed in its syntax and semantics layers as in classical compositional **Montague grammar**, producing a **logical form**; typing the terms with a **rich system of sorts** intended to capture **restrictions of selection**, using a **semantic lexicon**, produces some **typing mismatches** whenever polysemous terms are **linguistically coerced** to one of their facets.

In the framework we have proposed, the *Montagovian Generative Lexicon* (MGL), detailed in [23], this is characterised by a **mismatched application**, such as a functional predicate P requiring an argument of type B being applied to an argument a of type A : $(P^{B \rightarrow t} a^A)$. This is resolved using a **lexical transformation** that serves as the representation of the **linguistic coercion** taking place, and which should be provided by either P or a . There are two possibilities: *adapting the argument* with a transformation $f_1^{A \rightarrow B}$, the application becoming $(P (f_1 a))$, and *adapting the predicate* with a transformation $f_2^{(B \rightarrow t) \rightarrow (A \rightarrow t)}$, the application becoming $((f_2 P) a)$. Depending on the transformations that are provided by the functional terms P and a , one or the other adaptation occurs.

The phrase *the dinner was delicious but took a long time* (adapted from a canonical example, see e.g. [1]) can schematically be represented as

$$(\text{and } (\lambda x.\text{delicious } x) (\lambda x.\text{long } x)) \text{ dinner}$$

Within a many-sorted system, *take a long time* is restricted to events (entities of sort **evt**, or a subtype thereof), *delicious* is restricted to food (entities of sort F), and at least a transformation is needed in order to predicate on the two facets of *dinner*. MGL resolves this by having *dinner* as a term of type $\text{evt} \rightarrow t$, possessing a transformation $f_c^{(\text{evt} \rightarrow t) \rightarrow (F \rightarrow t)}$ which is a function mapping the dinner event to the food that was served at the dinner in question.

A crucial issue for these systems is then to have sufficient lexical resources (as a rich lexicon incorporating types and coercions) to function. In order to be successful, MGL and other type-theoretic logical frameworks for lexical semantics require:

For the Syntax-Semantics Analysis: A suitable syntax-semantics analyser which is able to provide a sufficiently structured output for a Montague-style analysis. MGL can make use of Grail (presented in [20]) easily, yielding λ -DRT-based outputs based on Type-Logical Grammars. Grail operates on extensive, corpora-driven grammars for French.

A System of Sorts: Each word (lexeme) should be associated to a typed term in a type theory (or typed λ -calculus) where types are functionally and inductively built from *base types*. Some theories use *common name* as base types, as discussed in [14]. In our approach, the functional base types correspond to

lexical **sorts** that capture the semantic notion of *restrictions of selection*. In this paper, we use the word *sort* when referring to the lexical notion, and *base type* for the technical, λ -calculus notion necessary for computation. While the exact definition and scope of these lexical sorts is debated, this can be decided arbitrarily, as long as they provide a starting point that can be enriched and refined if untreated restrictions of selection become apparent; in this paper, we use sorts defined by the semantic features of our resources, as discussed in Sect. 2.1.

A Set of Lexical Coercions: The core of MGL-based lexical semantics is the set of lexical transformations (corresponding to the *linguistic coercions* available for each *lexeme*), that allows co-composition to occur. Transformations might also be *constrained* in their use (some may be incompatible with others), and be dependent on (or made easier by) a specific *context*. The difficulties in building a wide-coverage lexical database for MGL lies in the acquisition of all such relevant transformations for all lexemes.

To sum up, MGL needs lexical entries in a format such as the following:

Lexeme	Logical term	Type	Comments
dinner	$(\lambda x. \text{dinner } x)$	$\mathbf{evt} \rightarrow \mathbf{t}$	As in Montague semantics, nouns are predicates; \mathbf{evt} is for <i>events</i> , \mathbf{t} for <i>propositions</i>
with dinner	f_c	$(\mathbf{evt} \rightarrow \mathbf{t}) \rightarrow (F \rightarrow \mathbf{t})$	Transformation to “food that was served at dinner” (type $F \rightarrow \mathbf{t}$, base type F for the <i>foodstuff</i> sort)
to take a long time	$(\lambda x. \text{long } x)$	$\mathbf{evt} \rightarrow \mathbf{t}$	Predicate of events
to be delicious	$(\lambda x. \text{delicious } x)$	$F \rightarrow \mathbf{t}$	Predicate of food

(MGL encompasses higher-order composition mechanisms that will allow operators such as *the* and *and* to compute the correct predications).

The main goal of this paper is to show how we can obtain the *lexical transformations* (such as f_c above) for our lexical entries.

2 Lexical Data Crowd-Sourced from Serious Games

While lexical networks have been developed for a long time (*WordNet*, defined in [19], being the reference for English), there have been several recent efforts to build collaborative, crowd-sourced resources that can reflect the current uses and relations of words by language speakers. One approach, given in [3], is to engage as many people as possible in a “serious” game (or, more accurately, a “game with a purpose”) in order to identify and co-validate lexical and relational information by having different competent speakers of the language competing to identify lexical meaning and relations between words.

These games include *JeuxDeMots*, described in [10], which provides a lexical network that comprises more than 200 million relations between words (as strings of characters). *JeuxDeMots* is actively developed and has proven remarkably robust. One advantage of this network over expert-produced and corpus-based resources is that a large amount of world knowledge has been added by the players: facts such as restrictions of selection (as in *cats can meow*) or ontological inclusion (as in *armchairs are chairs*) are explicitly produced by human players, while they are hard to get from other sources because of their “trivial” nature.

Another advantage of JDM is that it provides weights based on the frequency of words and phrases which is obtained from the available data played by users. Generally, there are two kinds of weights provided as negative and positive integers. The positive weights show the degree of confidence in a relation, while the negative weights, in contrast, indicate the degree of opposition between the nodes. For instance, *Autruche* (ostrich) and *déplacement aérien* (flight) have a relation called *r-agent-1* with a negative weight of -65 as one of their properties in the *JeuxDeMots* database.

This network can be used as a relevant source of lexical data for type-theoretic frameworks, as discussed in [4] for MTT. While that publication has demonstrated the capacity in which the *types* for each lexeme can be extracted and derived (an MGL-compatible lexicon with a different set of lexical sorts can easily be produced), we want to focus on the extraction of the *lexical transformations* from such lexical networks.

We will be using the lexical network created by *JeuxDeMots*, amended by several related open, contributive resources to make a complete resource known as *Rezo*, together with Grail as a syntax-semantics analyser, and MGL for lexical semantics. This forms a complete, coherent treatment process for French text.

2.1 Lexemes, Sorts and Sub-Types

JeuxDeMots and many other lexical networks operate upon character strings, while MGL differentiate between *contrastively ambiguous* homonyms. Words such as *bank* are considered as having (at least) two different entries in the Generative Lexicon tradition: one of the sort *Financial Institution* and the other of sort *Geographical Feature*, that happen to have the same string representation. In *JeuxDeMots* where the character string is the basic unit, there are two ways to detect contrastive ambiguity: *Semantic Features* and *Refinements*. A *Semantic Feature* is similar to a sort in the lexicon (and a base functional type), as it reflects a broad category of things that the character string can denote; there might be several of such features associated to each string. *Refinements* are single-meaning facets for this string that have been crowd-sourced for the express purpose of resolving the contrastive ambiguity.

For example, in French, *un bar* is ambiguous and can denote at least three different things:

- a place where drinks are served (as in the English “bar”, roughly synonymous with “public house” and “café”, polysemous via metonymy with the furniture item sharing this name and function);

- a kind of fish (in English, “sea bass”);
- and a pressure unit (1 bar = 100 kPa).

This is denoted in the data from *JeuxDeMots* as having different refinements, as well as having several semantic features, including some that are incompatible with each other (here, “bar” has “location”, “artefact” and “living being” as features). Extracting the initial information from *JeuxDeMots* is straightforward. *Common nouns* are logical predicates $P(_)$ of some type $\tau \rightarrow \mathbf{t}$, with τ a *sort*, subtype of entities **evt** for which satisfying of not the predicate makes sense. The initial sort is given as a salient semantic feature in *JeuxDeMots*.

These initial sorts can be refined later as needed for selection restrictions, which are also included in the lexical network. The network details possible patients and agents of predicates, for example. If the crowd-sourced data indicate that several incompatible semantic features are available for a single word, it simply means that we will have several distinct entries for several different lexemes that have the same string representation.

The typing for word taking nouns as arguments (such as adjectives, verbs. . .) are derived from a similar process and has been thoroughly explored in [4]. Specific sorts are added as has been proposed for MGL in [16]: specific sorts can be introduced for nouns denoting *groups of a given sort* (such as *committee* being a predicate that denotes a group of people, \mathbf{g}_P) and *massive entities* (such as *water* being a mass physical noun, of sort \mathbf{m}_φ). Several “operational” terms, such as the polymorphic conjunction *and* and determiners derived from Hilbert operators (discussed in [17]), are added by hand.

We can also derive a sub-typing mechanism. However, as discussed in [18], MGL restricts this mechanism (as well as the strict notion of “coercion among types”) to *ontological inclusions*. In the system of sorts, this ontological hierarchy can be detected in data given from *JeuxDeMots* as denoting *hyperonymy* quite easily (such relations are pervasive in lexical networks). No other type-driven coercions are included in the system, contrary to other approaches—we think that this is a sensible restriction.

2.2 Lexical Transformations

From the definitions of the MGL system that have been presented before, the way to determine the lexical transformations never has been explicitly mentioned. By playing *JeuxDeMots*, players effectively create a database of **coercions between words** that we need to process according to the lexical entries and their typings.

As explained before, there are two ways to resolve a **type mismatch** in an application such as $(P^{B \rightarrow \mathbf{t}} a^A)$: adapting either **the functional argument** or **the predicate**.

2.3 Adapting the Argument

The most common adaptation is done on the argument, using a relevant transformation $f^{A \rightarrow B}$, yielding the correctly-typed $(P(f a))$. There are two **origins**

possible for the transformation f : the transformation f is either provided by the lexical entry associated with the argument term a itself, or by lexical entry associated with the predicate term P .

Argument-Driven Transformations are extracted directly from entries revealing several meanings, that will be constrained by predication. For example, the expression *a book* is itself polysemous; an applied predicate such as *read*, *finish*, or *pack* will have strong typing constraints that will select one or several meanings (we will elaborate on this canonical GL example).

In MGL, the lexeme *book* (understood as the common noun associated to a literary concept versus the calendar-related verb) will have a *single sort*, R for readable object and be typed $R \rightarrow \mathbf{t}$. The same lexeme will be associated to several transformations:

$f_{object}^{(R \rightarrow \mathbf{t}) \rightarrow (\varphi \rightarrow \mathbf{t})}$ that will be used in *heavy book*, $f_{read}^{(R \rightarrow \mathbf{t}) \rightarrow (\mathbf{evt} \rightarrow \mathbf{t})}$ and $f_{write}^{(R \rightarrow \mathbf{t}) \rightarrow (\mathbf{evt} \rightarrow \mathbf{t})}$ used in *to finish a book*, $f_{author}^{(R \rightarrow \mathbf{t}) \rightarrow (P \rightarrow \mathbf{t})}$ used in *a monarchist book*.

In all of these cases, the transformation is *part of the lexical entry for the argument* and *constrained by the typing of the predicate*: *pro-monarchy* applies to people (and is associated to the *agentive quale* in GL tradition) and is of type $P \rightarrow \mathbf{t}$, *to read* is of type $R \rightarrow \mathbf{t}$, *to finish* of type $\mathbf{evt} \rightarrow \mathbf{t}$ (both are associated to the *telic quale*), and *to pack* of type $\varphi \rightarrow \mathbf{t}$ (associated to the *physical facet* of the complex object *book*).

In order to infer these transformations from lexical networks such as *JeuxDeMots*, we list the various possible *target types* for the predicates that are listed as *common patients* for this word. Then, for each target type, we select the compatible *associations* listed for the word, and generate a lexical transformation associated to this word, of a given typing, labeled with the association that is given in the lexical network. We retain the weight (frequency) of the association in order to filter out the more dubious transformation at a latter stage (this is also used to rank the preferred interpretations, if several are available).

For this example, listing all strings associated with *livre* (book; there are more than 3000 associations in the network, weighted and labelled), we will scan for *nouns denoting humans* and obtain *auteur* (author) and *écrivain* (writer), both with very strong relative weights; for *action verbs denoting non-instantaneous events* (that can said to *begin* or *end*) and obtain *lire* (to read) and *écrire* (to write), with a much stronger relative weight for the former. This process is detailed in Sect. 4. Even if some associations provided by the players in the network are dubious, filtering by type and syntactic properties yields exactly what is needed for the lexicon.

Predicate-Driven Transformations are not (all) to be found in lexical networks or crowd-sourced data, as they characterize a predication made “in the spur of the moment”. For example, the predicate *to read* is associated to a transformation $f_{written}^{(\varphi \rightarrow \mathbf{t}) \rightarrow (R \rightarrow \mathbf{t})}$ that allows sentence such as *I read the wall* to be

felicitous (and simply supposes that there is something written on said wall). Of predicate-driven transformations, some will be derived from lexical data (*grinding* is a common example, whenever *food* is associated to a word denoting a *living being*); others will need to be generated whenever the predication occurs, and be invalidated or not. The dynamic aspect of the crowd-sourced lexical network will integrate additional transformations as they are deemed pertinent, or fashionable, by its community; this is a strong argument in favor of such resources.

2.4 Adapting the Predicate

Transformations that change the type of a predicate are less common in MGL mechanisms. They are either provided by adverbs or other modifiers to a predicate, or intrinsic to the predicate. The former will be changing the target typing of a predicate. Examples of the latter include the polysemy of readings between collective and distributive among plural predicates detailed in [16]; these should be added manually for the lexicon.

2.5 Constraints and Relaxation

Transformations are associated to compatibility constraints that can be defined either as logical operators, or as arbitrary functions that filter the possible combinations of transformations; these are intended to suppress or signal hazardous co-predications. Our interpretation of such phenomena is the following:

- *predicate-modifying* transformations are cumulative, as in *we all lifted the pianos, working in pairs* (our account will provide a predicate coercion, further constrained to a *covering* reading by the complement);
- *argument-driven* transformations are compatible with each other and not constrained, and give rise to *felicitous co-predications* such as *the dinner was delicious but took a long time* and *heavy yet interesting book*;
- *predicate-driven* transformations are exclusive: only one of them can be used on a given entity, exclusive of any other transformations and of the original term; this can be seen when making *infelicitous predications* such as **fast and delicious salmon* or **Liverpool won the match and voted Remain*.

As suggested by several studies including [23], the latter constraint can be relaxed. Lexical predicate-driven transformations such as *grinding* (that we can get from crowd-sourced data) are compatible with other meanings, as long as there is a syntactic break between the two predications (as in *that salmon was fast; it is delicious*). There are ways to relax the constraints of application of non-lexicalised predicate-driven transformations as well, but these are syntax-, discourse- or pragmatic- dependent.

Detecting and validating constraints on co-predication, beyond the simple claim above, can also be a crowd-sourced task. *JeuxDeMots* itself is not suitable for this; however, a recent effort, *Ambiguss* (available at <https://ambiguss.calyxe.fr/> and discussed in [11]) is a good blueprint for crowd-sourced co-predication validation. *Ambiguss* is a database of sentences containing ambiguous (polysemous) terms that asks players to compete in determining the possible readings for those terms in context, and thus will be able to detect whenever co-predications are accepted or rejected when enough data will have been crowd-sourced. Such resources would also be useful for the evaluation of the performance of MGL on word-sense disambiguation tasks.

3 Integrating and Ranking Transformations

3.1 Adding Collected Transformations to the Lexicon

The semantic lexicon in MGL associates a set of *lexical transformations* (as optional λ -terms) to each lexeme. When the pertinent data has been collected from *JeuxDeMots*, the coercion is transcribed as a transformation with a functional type $A \rightarrow \tau$ where A is the type of the source lexeme and τ is the expected typing, predicted during the collection process. The target concept serves as the name of the transformation. This data can be directly input and used in our prototype implementation of MGL given in [15] (as the only necessary data are a source type, a target type and a name), and is added to the list of transformations of the current lexeme.

3.2 Scoring Interpretations

Compositional Lexical Semantics can produce the precise meaning of a polysemous term in context. However, there are cases where the immediate context is not sufficient to totally determine the sense used for a word, and a composed phrase can still be ambiguous: in the example above, *finir un livre*, the entity *livre* (book) is coerced to an event with a duration, but there are many suitable lexicalised events that can be used, and thus many lexical transformations that have a correct typing that can be used to resolve the type mismatch. MGL usually resolves this by producing **several** interpretations: one interpretation (a well-typed logical formula with a suitable transformation placed whenever necessary) per available coercion. The production of MGL is thus not a single logical representation, but a **collection of representations**.

Associating a preference score to different possible interpretations given by a compositional formalism is not so common (although this practice is pervasive in statistical or machine learning approaches to word-sense disambiguation); the more convincing works on this matter are mostly derived from Preference Semantics described, e. g., in [8]. An interesting extension is [21], giving an implementation of the constraint-based preference scoring.

In order to provide a preference ranking, we exploit a strong advantage of extracting coercions from crowd-sourced data: the most common preferred coercions associated to a lexeme will be clearly represented in the lexical network (because *JeuxDeMots* counts the number of times each fact has been contributed or validated). Using the relative weights of different coercions, MGL will then be able to associate a *relative score* (similar to a probability measure) to each logical representation produced, allowing to rank the interpretations (while still being able to generate all possible meanings).

Our proposal for producing this **score** consists in a simple procedure. In order to analyse and rank interpretations for *finir un livre* (finish a book):

1. Starting from textual data syntactically analysed using Categorical Grammars, we obtain a full logical representation, the terms of which are typed using a **many-sorted semantic lexicon** (in which event sorts *DurableEvent* and *AtomicEvent* are differentiated), we first identify the **adaptation** taking place in applications with type mismatches such as (finish^{*DurableEvent*} a_book^{*Readable*}).
2. The lexicon then uses data available from *JeuxDeMots* in order to build the **set of possible coercions**: taking all **relations to the argument first livre** (book), we **select the ones with compatible types** with the typing of the predicate: verbs denoting a non-instant action. In this example, there are only *lire* (680) and *écrire* (210) in the first two hundred relations.
3. Finally, normalising to a probability-like measure, this yields a **score between 1 and 0** for several interpretations ranked by order of probability. In this example, this means a score of 0.764 for *lire* and 0.236 for *écrire*.
4. This is validated by the crowd-sourced glosses for the complete phrasal expression that appear in that same order above.

In the case where **no suitable coercion is given by the argument**, a predicate-driven transformation can be used, as discussed in Sect. 2.3.

Lexical transformations provided by the (functional) argument are always more specific and preferred to generic, predicate-driven transformations; thus the latter only occur when there is no other choice, and there is only one interpretation (hence this case does not need any preference score).

Specific contexts can provide coercions that are not scored by this procedure, or that have different preferences. For instance, different contexts of enunciation will change the order of preferences in *finish a book*, for instance placing *writing* before *reading* in the context of the Paris Book Fair, or adding new possible coercions in contexts such as a bookbinding workshop.

3.3 Correcting the Lexicon Using Different Sources

JeuxDeMots is a strong and rich resource for French as it is, encompassing at the moment more than 200 million relations between more than 2.7 million terms (words, names, and phrases). However, one of its more interesting characteristics for our purpose is that it is a living resource, constantly updated by

crowd-sourcing (there are several players logged in and actively contributing at all times) and experts (selected members of the team constantly correct, update and add linguistic data and world knowledge). This means that, after acquiring the initial lexicon used for a MGL system by converting relevant data from *JeuxDeMots* as outlined above, the same source can be used in order to continuously update and correct our lexicon. The presence of *phrases* in data from *JeuxDeMots* can also be used to evaluate and validate our compositional system. We can, for instance, derive the type and transformations for the lexical entry of *finir* (*to finish*), do the same thing for *livre* (*book*), and conclude from a MGL composition the meaning of *finir un livre* (*finish a book*: what is *finished* is the event of *reading* or *writing* a book); this is validated by the associated meanings of *finir un livre* that appears as a phrasal expression in *JeuxDeMots*.

This process can be systematised and automated, as *JeuxDeMots* can be enriched by asking the players the meaning of an expression that has not yet been analysed (within reasonable limits). The meaning of words in context, as predicted by our compositional system, can also be checked by the players of the serious game *Ambiguss*. Thus, the evaluation of our system, as well as its correction and continuous improvement can be crowd-sourced.

4 A Short Case Study

We will concentrate on some examples in order to demonstrate how the ideas that are described in previous sections can actually be achieved using *JeuxDeMots* by illustrating the translation procedure from *JeuxDeMots* to proper meaning representations in Montagovian Generative Lexicon (MGL) framework. The case study originates from the example in the influential book *Generative Lexicon*. Several publications, starting with [12], have remarked that *to finish a book* can be predicted to have different meanings depending on its *subject* (agent).

An account of this phenomena proposed in [9] is that meaning is constructed in three phases:

1. lexical meaning is extracted from the words involved;
2. compositional meaning is created by the combination of predicate and object, with different interpretations occurring and the most ranked being *reading*;
3. contextual (world-knowledge) meaning is combined by meaning associated to the subject and the interpretations can be re-ordered or filtered out.

We would attempt to provide an adequate prediction using the resources at our disposal, with the idea that the necessary pragmatic information or world knowledge is encoded in the lexicon (and is present in *JeuxDeMots* as well).

4.1 The Straightforward Case

Let us consider the phrase *To finish a book* in a situation where we have Claire, a researcher, reading a book by Villani:

- (1) Claire a fini le livre de Villani.
 ‘Claire finished Villani’s book.’

In order to analyse such a sentence, we search for a more generic phrase in *JeuxDeMots*: (a) *la femme a fini un livre* (the woman finished a book) for (1). Lexical coercions to actions such as *reading* or *writing* are excepted from *JeuxDeMots*. Moreover, *JeuxDeMots* also gives us the built-in weights relation that results in the expected ranking on the obtained lexical coercions.

4.2 Lexicon Organization in MGL and Meaning Representation

A detailed explanation on lexicon organisation in MGL is available in [23, Sect. 2.4]. In summary, the lexicon in MGL associates to each word w a principal λ -term which is Montague term with a much richer typed system and optional λ -terms known as *modifiers* or *transformations* modelling *lexical coercions*. The following sample lexicon is designed for the example (a):

Lexeme	Main λ -term	Optional λ -terms
livre (obj)	book $^{R \rightarrow t}$	$f_{read}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)}$ $f_{write}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)}$
femme	woman $^{P \rightarrow t}$	
fini	finish $^{\alpha \rightarrow (evt \rightarrow t)}$	

The sorts used here are as follows: P for *Person*, R for *Readable*, and **evt** for *Event*. Skipping unnecessary details on quantifiers and syntax, we can represent two possible meanings for (a) as

$((finish^{\alpha \rightarrow (evt \rightarrow t)} (the\ woman)^P) (a (f_{read}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)} book^{R \rightarrow t})) evt)$ and
 $((finish^{\alpha \rightarrow (evt \rightarrow t)} (the\ woman)^P) (a (f_{write}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)} book^{R \rightarrow t})) evt)$.

4.3 Collecting Coercions

The relations gained from the game players in *JeuxDeMots* can technically be represented in different structures. The two kinds of data representation that are implemented and available are **relational** and **graph-based databases**. A simple query on an SQL system or on *Cypher*, the graph query language, on the *JeuxDeMots* graph-based database can fulfill our demand. We can also use *Datalog* which is a declarative database query language (basically Prolog with only constants and variables, i.e without terms, but with a proper negation). Although all of the options are technically available, we illustrate the process using a simple PHP-like query syntax which has actually been experimented.

As for example (a), what we want to do is basically to find the coercions *lire* (read) and *écrire* (write) for a sentence with the object *livre* (book) and the subject of sort *Person*. As illustrated in Fig. 1, we can find all the nodes with

relation *r_associated* to the node *livre*. Two filters are then applied. The first one rules out the candidates that do not have the syntactic property of being a verb; to do so in *JeuxDeMots* we use *r_pos* relation that targets the node *Ver:Inf*. The second rules out the verbs that cannot have an agent of sort *Human*; to do so we use *r_agent* relation that targets the node *femme* (woman). We sort in descending order the final table with the built-in weight property of *r_associated* relation that exists in *JeuxDeMots*.

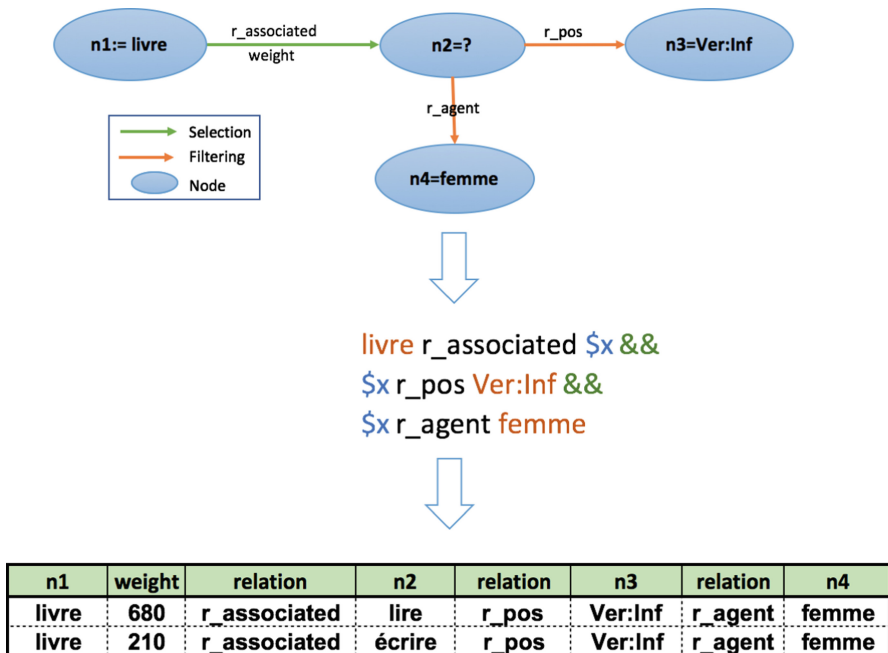


Fig. 1. Obtaining Coercions: General Scheme, Query Code and Outcome Table for Example (a)

4.4 The Non-Human Case

Regarding the extraction of the coercions from *JeuxDeMots*, we can see that the input of a query involves more than two words. For instance, considering the meaning of “dévorer un livre” (idiomatic French, *to devour a book* that can be used to denote binge-reading); in *JeuxDeMots* a relation between one of the meanings of “dévorer” and one of the meanings of “livre” depend on a third ingredient, namely the agent of “dévorer”. Assuming that Blanche is Claire’s pet goat, contrast the following:

- (2) a. Claire a dévoré le livre de Villani.
 ‘Claire devoured Villani’s book.’
 b. Blanche a dévoré le livre de Villani.
 ‘Blanche devoured Villani’s book.’
 c. Claire a permis à Blanche de finir le livre de Villani.
 ‘Claire allowed Blanche to finish Villani’s book.’
 d. Claire a promis à Blanche de finir le livre de Villani.
 ‘Claire promised Blanche to finish Villani’s book.’

Observe that the last two examples with control verbs require a syntactic computation to determine the agent/subject of the action that is performed on the book.

A further limit is that there are no sorts, types or sets in JeuxDeMots. If one is asked what a goat can eat, it is unlikely that a player answers “a book”. The answer: bedsides, grass, bushes, flowers, branches, leaves etc. a more generic answer would be “any object that is small and not too hard”, but there is no single word corresponding to this class, while players answer words.

For this particular case, the fact that a goat may eat a book **is**, actually, included in JeuxDeMots. Some players included and validated the fact that a goat may eat “paper” and books being made of paper (this is indicated in the entry for “book” as a constitutive coercion), they can be eaten by goats. In this case, there even is also a direct fact that goats can eat books, but with a much weaker confidence.

4.5 Limits of MGL

In MGL as it stands now, coercions are attached to one word – except ontological inclusions which are encoded via sub-typing – but the actual coercion used to fix a type mismatch could be the combination of several coercions provided by all the words in the expressions. As observed above a coercion may be triggered by several words, and may be the result of a sequence of relations. How can this be stored in MGL, in such a way that the general compositional mechanism of MGL produces the wanted readings and discards the unwanted ones? How can this be performed without trying all possible combinations of coercions, i. e., without going beyond reasonable time complexity limits?

A solution is to split the coercion in two (or more), each part being associated to each lexeme, the composition giving the complete result.

In that case, a type mismatch $P^{A \rightarrow \tau}(u^B)$ may be solved by first using a coercion $f^{B \rightarrow X}$ attached to u and a coercion $g^{X \rightarrow A}$ attached to P .

In all the above examples, there is a coercion from *books* to their physical facet. Having a *goat* as an agent will provide the verb *to eat* (which normally has an agent of sort *Animal* and a patient of sort *Food*) with a coercion from *Food* to physical objects, representing the *ingestion* of these objects and the world knowledge “fact” that “goats will eat (mostly) anything”. That way, the goat Blanche may well eat Villani’s book.

4.6 A Direct Solution

We would expand our sample lexicon to this:

Lexeme	Syntax constraint	Main λ -term	Optional λ -terms
livre (obj)	$subj: P$	$book^{R \rightarrow t}$	$f_{read}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)}$ $f_{write}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)}$
livre (obj)	$subj: A$	$book^{R \rightarrow t}$	$f_{eat}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)}$
chèvre		$goat^{A \rightarrow t}$	
femme		$woman^{P \rightarrow t}$	
fini		$finish^{\alpha \rightarrow (evt \rightarrow t)}$	

The syntactic constraint column extends the standard MGL lexicon in order to capture meanings that depend on the object and subject in a given sentence. In our case, having a subject of sort either *Animal* or *Person* can significantly change the meaning and it obviously needs different kind of coercions in the meaning representation layer.

Considering the example:

- (3) Blanche a fini le livre de Villani.
‘Blanche finished Villani’s book.’

We adapt this to a more generic sentence, **(b)** *la chèvre a fini un livre* (the goat finished a book) for (3), and the single reading for (b) should be $((finish^{\alpha \rightarrow (evt \rightarrow t)} (the\ goat)^A) (a (f_{eat}^{(R \rightarrow t) \rightarrow (evt \rightarrow t)} book^{R \rightarrow t})) evt)$.

We want to find the coercion *manger* (eat) for a sentence with the object *livre* (book), the subject being the word *chèvre* (goat). As illustrated in Fig. 2, we can find all the nodes with relation *r_patient* to the node *livre*. As before, two filters are applied, selecting for verbs that can have *chèvre* as object; we could then sort by weight if there were more than a single result.

This is direct relation extracted from *JeuxDeMots*; its confidence degree is light, and this is to be expected. In future work, we would like to acquire the coercion using “paper” as an intermediate step by searching for possible sequences (of limited length).

Thus, we can use *JeuxDeMots* to derive the coercions that we need incorporate into the MGL lexicon.

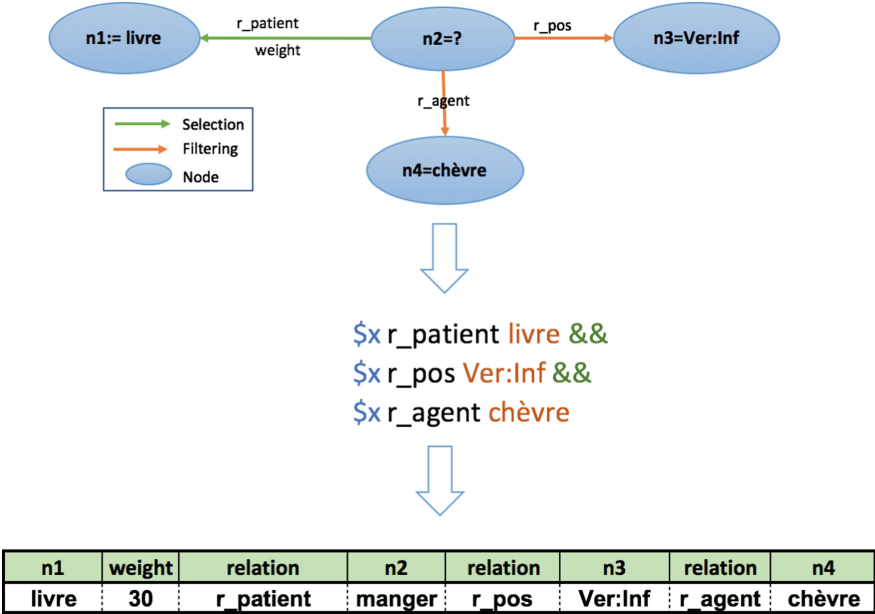


Fig. 2. Obtaining Coercions: General Scheme, Query Code and Outcome Table for Example (b)

5 Conclusion

The Grail syntax-semantics analyser, together with the type-theoretic account of lexical polysemy provided by MGL and based on compositional semantics and the ATY_n many-sorted logic, forms a computational system that is well-suited to lexical and semantic data crowd-sourced using *JeuxDeMots*. They can provide a complete chain of analysis that can process different complex linguistic phenomena for the French language. Grail has long been used in different versions, and has access to a large-covering French corpora-driven Type-Logical Grammar; *JeuxDeMots* provides access to a mature lexical network of words, phrases and relations that is continuously updated, with publicly available and queryable data; we also have previously demonstrated the pertinence and computational applications of MGL. We have presented a process and experimental data that shows that this treatment chain works, and can be automated. Moreover, this system can be evaluated, corrected and updated in a semi-automated fashion using similar crowd-sourced data.

What remains to be done is mostly a work of integration of these various components. The possibility of modifying the existing MGL framework for allowing multi-part coercions to be added, as a composition of transformations licensed from different lexemes, should be examined in detail, as well as the implications of this modification to the time complexity of the computation, and the expressive power of the resulting formalism.

References

1. Asher, N.: *Lexical Meaning in Context: A Web of Words*. Cambridge University Press, Cambridge (2011)
2. Bekki, D.: Dependent type semantics: an introduction. In: Christoff, Z., Galeazzi, P., Gierasimczuk, N., Marcoci, A., Smet, S. (eds.) *Logic and Interactive RAtionality (LIRa) Yearbook 2012*. vol. I, pp. 277–300. University of Amsterdam (2014)
3. Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., Poesio, M.: Using games to create language resources: successes and limitations of the approach. In: Gurevych, I., Kim, J. (eds.) *The People’s Web Meets NLP. Theory and Applications of Natural Language Processing*, pp. 3–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35085-6_1
4. Chatzikiyriakidis, S., Lafourcade, M., Ramadier, L., Zarrouk, M.: Modern type theories and lexical networks: using serious games as the basis for multi-sorted typed systems. *J. Lang. Model.* **5**, 229–272 (2017)
5. Chatzikiyriakidis, S., Luo, Z.: Natural language inference in coq. *J. Logic Lang. Inf.* **23**(4), 441–480 (2014). <https://doi.org/10.1007/s10849-014-9208-x>
6. Cooper, R.: Copredication, dynamic generalized quantification and lexical innovation by coercion. In: *Fourth International Workshop on Generative Approaches to the Lexicon* (2007)
7. Cruse, D.A.: *Lexical Semantics*. Cambridge, New York (1986)
8. Fass, D., Wilks, Y.: Preference semantics, ill-formedness, and metaphor. *Comput. Linguist.* **9**(3–4), 178–187 (1983). <http://dl.acm.org/citation.cfm?id=1334.980082>
9. Im, S., Lee, C.: A developed analysis of type coercion using ashers’ TCL and conventionality. In: Cooper, R., Retoré, C. (eds.) *Extended Abstracts of the ESSLLI 2015 Workshop TYTTLES: Types Theory and Lexical Semantics*, pp. 91–99, August 2015. <https://hal.archives-ouvertes.fr/hal-01584832>
10. Lafourcade, M.: Making people play for lexical acquisition with the JeuxDeMots prototype. In: *SNLP 2007: 7th International Symposium on Natural Language Processing*, Pattaya, Chonburi, Thailand, p. 7, December 2007. <https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883>
11. Lafourcade, M., Brun, N.L.: Ambiguss, a game for building a sense annotated corpus for French. In: *IWCS* (2017)
12. Lascarides, A.: The pragmatics of word meaning. In: *Proceedings of the AAAI Spring Symposium Series: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, pp. 75–80 (1995)
13. Luo, Z.: Contextual analysis of word meanings in type-theoretical semantics. In: Pogodalla, S., Prost, J.-P. (eds.) *LACL 2011. LNCS (LNAI)*, vol. 6736, pp. 159–174. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22221-4_11
14. Luo, Z.: Common nouns as types. In: Béchet, D., Dikovsky, A. (eds.) *LACL 2012. LNCS*, vol. 7351, pp. 173–185. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31262-5_12
15. Mery, B.: Challenges in the computational implementation of montagovian lexical semantics. In: Kurahashi, S., Ohta, Y., Arai, S., Satoh, K., Bekki, D. (eds.) *JSAI-isAI 2016. LNCS (LNAI)*, vol. 10247, pp. 90–107. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61572-1_7
16. Mery, B., Moot, R., Retoré, C.: Computing the semantics of plurals and massive entities using many-sorted types. In: Murata, T., Mineshima, K., Bekki, D. (eds.) *JSAI-isAI 2014. LNCS (LNAI)*, vol. 9067, pp. 144–159. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48119-6_11. <https://hal.inria.fr/hal-01214435>

17. Mery, B., Moot, R., Retoré, C.: Typed Hilbert operators for the lexical semantics of singular and plural determiner phrases. In: Epsilon 2015 - Hilbert's Epsilon and Tau in Logic, Informatics and Linguistics. Montpellier, France, June 2015
18. Mery, B., Retoré, C.: Are books events? ontological inclusions as coercive subtyping, lexical transfers as entailment. In: LENLS 2012, in JSAI 2015. Kanagawa, Japan, November 2015
19. Miller, G.A.: Wordnet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995). <https://doi.org/10.1145/219717.219748>
20. Moot, R.: The grail theorem prover: type theory for syntax and semantics. In: Chatzikyriakidis, S., Luo, Z. (eds.) *Modern Perspectives in Type-Theoretical Semantics*. SLP, vol. 98, pp. 247–277. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-50422-3_10
21. Nagao, K.: A preferential constraint satisfaction technique for natural language analysis. *IEICE Trans. Inf. Syst.* **77**(2), 161–170 (1994)
22. Pustejovsky, J.: *The Generative Lexicon*. MIT Press, Cambridge (1995)
23. Retoré, C.: The Montagovian Generative Lexicon Lambda Ty_n : a Type Theoretical Framework for Natural Language Semantics. In: 19th International Conference on Types for Proofs and Programs (TYPES 2013). Leibniz International Proceedings in Informatics (LIPIcs), vol. 26, pp. 202–229. Schloss Dagstuhl, Germany (2014)



How Dogwhistles Work

R. Henderson¹(✉) and Elin McCready²

¹ Department of Linguistics, University of Arizona, Tucson, USA
rhenderson@email.arizona.edu

² Department of English, Aoyama Gakuin University, Shibuya, Japan
mccready@c1.aoyama.ac.jp

Abstract. The paper focuses on the semantics and pragmatics of dogwhistles, namely expressions that send one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup. There are three questions that need to be resolved to understand the semantics and pragmatics of the phenomenon at hand: (i) What kind of meaning is dogwhistle content—implicature, conventional implicature, etc.; (ii) how do (some but not all) hearers recover the dogwhistle content, and (iii) how do expressions become endowed with dogwhistle content? These three questions are interrelated, but previous analyses have emphasized answers to a subset of these questions in ways that provide unsatisfactory answers to the others. The goal for this paper is to take stock of existing accounts, while showing a way forward that reconciles their differences.

1 Introduction

Dogwhistles can be defined as terms that send one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup. They are commonly deployed in political contexts to express opinions that politicians calculate will be unpalatable to some segment of their audience but which still yield some advantage when communicated to some other segment. Consider a recent example. On a 2014 radio program Representative Paul Ryan made the following statement, which was criticized shortly after by fellow Representative Barbara Lee as a “thinly veiled racial attack”.

- (1) We have got this tailspin of culture, in our inner cities in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

The authors would like to acknowledge the support of JSPS Kiban C Grant #16K02640, which partially supported this research, and to thank Heather Burnett and Judith Degen for extremely helpful discussion during the construction of the analysis presented here, as well as the participants at the 2017 Szklarska Poreba Workshop on the Roots of Pragmasemantics, the 2017 DGfS workshop on secondary content, LENLS 14 and the audience at Nagoya Gakuin University, especially Yu Izumi.

The “thin veil” refers to the phrase *inner-city*, which is code or euphemism for African American neighborhoods (especially stereotypically racialized views of such neighborhoods). Those aware of the dogwhistle heard Ryan promulgate the pernicious racial stereotype that African Americans are lazy. At the same time, Ryan maintained an amount of plausible deniability because this content was not explicit but only referenced a vague geographical location.

We see three questions that need to be resolved to understand the semantics and pragmatics of the phenomenon at hand: (i) What kind of meaning is dogwhistle content—implicature, conventional implicature, etc.; (ii) how do (some but not all) hearers recover the dogwhistle content, and (iii) how do expressions become endowed with dogwhistle content? These three questions are interrelated, but previous analyses have emphasized answers to a subset of these questions in ways that provide unsatisfactory answers to the others. The goal for this paper is to take stock of existing accounts, while showing a way forward that reconciles their differences.

We will start by examining these views with an eye to their strengths and weaknesses before pushing forward with our own analysis, which builds on [5]. We take the canonical example above as our test case in evaluating previous proposals and developing our own. As we will show, previous analyses all have their own strengths, but each either fails to address key aspects of the phenomenon or mischaracterizes the content of dogwhistles in problematic ways. The final proposal is that dogwhistles are neither Pottsian CIs as in the work of [16], nor hearer-centered, norm-violating inferences as in the account of [6], but are instead expressions that activate relevant normality conditionals in ingroup speakers but not in outgroup speakers; this can be viewed as a kind of audience-specific invited inference dependent on background knowledge in a way specific to dogwhistles. This analysis improves on the work of Potts and Khoo and fills a lacuna in the game-theoretic analysis of Henderson and McCready, ultimately yielding a substantial advance in our theoretical understanding of dogwhistled speech.

2 Views

There are three views on dogwhistles in the semantics and pragmatics literature. We consider them one by one and compare their perspectives on how dogwhistles work. What we will see is that while it is impossible to maintain a CI view of dogwhistles given the empirical facts, neither the inferentialist view of [6] and the game-theoretic view of [5] manage to account for certain aspects of how dogwhistles work. What we will propose in the light of these observations is a theory that blends aspects of the latter two theories, improving on both.

The CI View

Stanley (2015) is the first published view of dogwhistle speech in semantics and pragmatics. According to this view, dogwhistle terms introduce conventional implicatures: for instance, a term like *welfare*, which is associated with negative views of poor black people in the US, in addition to its literal meaning, carries

the conventional implicature that the speaker dislikes poor blacks. This makes it a mixed content bearer in the sense of [10], though Stanley himself doesn't use this term. However, the dogwhistled content has few to no properties of conventional implicatures, as pointed out by [5,6].

Most prominently, if a bit of content is conventional, as the not-at-issue content we see with mixed-content bearers is, it's not deniable any longer. This can be seen with pejoratives, which clearly carry conventional not-at-issue content, roughly that the speaker dislikes the group which is the target of the pejorative expression. It is very odd for a speaker to use a pejorative and then deny the expression of a negative attitude, as in the following dialogue: this is expected given that expressive items and conventional implicatures commit the speaker to the expressed not-at-issue content, as claimed by [12,13], and given that the not-at-issue content of pejoratives falls into one of these two classes.

- (2) A: Angela Merkel is a kraut.
 B: What do you have against Germans?
 A: #I don't have anything against Germans. Why do you think I might?

Such dialogues are fine with dogwhistles; in the following, there seems to be no entailment that A has the relevant attitude. This is unexpected if dogwhistled content is indeed the result of dogwhistles being mixed content bearers.

- (3) A: Donald is on welfare.
 B: What do you have against social programs?
 A: I don't have anything against social programs. Why do you think I might?

By this test, dogwhistles of all types can be concluded to not be conventional, and thus, a fortiori, not mixed content bearers.

The Inferentialist View

Another proposal on the market is the inferentialist view of [6], which we are quite sympathetic with in many respects and which addresses aspects of dogwhistles that [5] do not. Khoo's idea is that dogwhistles induce certain kinds of inferences: namely, those which the existing beliefs of interpreters coupled with the information provided by the dogwhistle combine to yield. Schematically, if the speaker claims that x is C and the interpreter believes that C 's are R 's, then the interpreter will conclude that x is R ; it's this kind of inference that Khoo thinks that dogwhistles license. The key point is that if the interpreter lacks the belief that C s are R s, the relevant inference won't arise: this is the way in which Khoo explains the difference in interpretation between ingroup and outgroup speakers. In the example we have focused on, suppose that the interpreter believes that *inner-city* neighborhoods are African American neighborhoods. Then the speaker saying that people who live in inner-city neighborhoods lack a culture of work licenses the inference that people who live in African American neighborhoods lack a culture of work. This is a kind of invited inference account which relies on the (at-issue) content of the dogwhistle itself and the background

beliefs interpreters have which license a constellation of inferences about things related to that content.

This kind of account gets around the problems of treating dogwhistles as CIs. Most importantly, the dogwhistle effect is not conventionalized, but is instead entirely listener-based, which preserves the speaker's deniability. This is the critical fact that CI accounts miss. At the same time, an account that is based entirely on the extensional content of the dogwhistle and the listener's background beliefs is too weak. As Khoo himself notes, the account predicts that any two coextensive terms should induce the same 'dogwhistle inferences,' but they don't: only certain terms do, namely those which can independently be identified as signaling certain aspects of speaker identity in the dogwhistley manner. We thus seem to require a theory in which the dogwhistle inferences are tied to specific linguistic expressions, but not a part of their conventionalized semantic meaning, as with a CI. This is a tricky middle way to find.

Khoo's solution is to appeal to work of [4] on belief fragmentation for a fix: the idea is that thinking of Xs in one way may not deliver the same inferences as thinking of them in a different way, so even coextensive terms may not give the same inferences. Indeed, he indicates (in his footnote 19) that a metalinguistic theory of these words is probably needed (though he makes no attempt to spell one out); 'beliefs about the code words themselves may be relevant.' We agree with this suggestion, though no detail is provided; we think it's precisely the use of the dogwhistle *qua* particularized expression that has to be taken into account when trying to compute what meaning is transmitted and what the likely intentions of the speaker are. However, we want to go further; we suggest that the need to look more closely at the linguistic expressions invalidates the inferentialist theory with its focus on *content*, and requires us to move to a view which induces the inferences arising from dogwhistles on the basis of the forms of the messages themselves, as [5] also emphasize.

The Game-Theoretic View

[5] provide an account of dogwhistles in a game-theoretic setting. In particular, we build off of pathbreaking work by [1, 2] on what she calls *Sociolinguistic Signalling Games* (SSGs). The core idea behind SSGs is that, in communication, speakers attempt to construct a sociolinguistic persona that listeners try to recover (as in Third Wave sociolinguistics, e.g. [3]). This process is mediated by linguistic expressions, which not only have a semantic meaning, but also a social meaning—namely those personae that the expression is consistent with. Finally, we assume that both speakers and listeners assign values to personae. In the speaker's case, the value is based on rankings on personae: the higher the ranking, the more the speaker wishes to be perceived as having that persona; in the listener's case, it is based on a ranking of personae that they (dis)approve of.

Against this backdrop, dogwhistle language arises under three conditions. First, a linguistic expression becomes associated with a particular persona. This sort of association is not so surprising. Certain groups of people speak a certain way, and any variation, including lexical choices can signal group membership,

a familiar point in sociolinguistic theory (e.g., [7]). Second, there is differential awareness in the population about how strongly that expression signals a particular persona. This is also not surprising. People not aware of a group's culture will not be aware of how they use language (e.g., [15]). Finally, there is a difference in the population in how individuals value that persona. If some individuals value a particular persona highly, but others strongly disapprove of it, there will be an incentive for speakers to signal their adoption of that persona only to certain groups. Under these conditions, it may become possible to use a linguistic expression to signal your persona to a subaudience, while hiding your persona from a large subaudience that would disapprove of that persona. In the game-theoretic perspective, this becomes a utility-maximizing strategy.

Note that in this discussion of persona, we have not talked about communication of 'genuine' linguistic content (i.e. plain vanilla semantic content), such as the kind of inference observed in the move from "inner city" to "African American neighborhoods" in the example we have been considering here. [5] show that there are actually two kinds of dogwhistle. Only the first involves exclusively the transmission of speaker personae. With the second kind (called Type 2), the content sends one message to all audience members, while the whistle enables the placement of an addendum on that message for a sub-audience which has a truth-conditional impact, something in the manner of the pragmatic enrichment of [14]. We argue that recovering this message is based on a listener recovering a particular persona for the speaker. The Ryan case above best fits this category. His use of "inner city" may convey to a subaudience that he has a particular persona, and so in virtue of that persona, when he says "inner city" he is referring to African American neighborhoods in those cities.

This kind of account is able to avoid the problems of the CI account. The reason is that speakers, in general, are able to deny that they have a particular persona in virtue of that fact that expressions often only loosely signal particular personas. In the case of Type 2 dogwhistles, denying the persona amounts to denying enriched content sent by way of that persona, e.g., the enriched meaning "African American neighborhood" from the expression "inner city".

What the account in [5] cannot do is explain how "inner city" is related to "African American neighborhood". Assuming that the connection between these two pieces of content is available, our game-theoretic account can provide an account of when the inference from one to the other will arise and that it is deniable, but unlike the account in [6], the relation between the dogwhistle expression and the dogwhistled content is opaque. This is an unfortunate feature of the game-theoretic account, which makes it incomplete as a full account of the pragmatics of dogwhistled expressions. We aim to rectify this situation in the remainder of the paper.

3 A Mixed View: Defaults, Backgrounds, and Form

From the three views described we can extract a number of considerations relevant to a full account. From the CI account and its problems it becomes clear that

dogwhistled content is pragmatic, but not fully conventionalized as a ‘proper’ nondefeasible part of meaning. From the inferentialist account we find a connection between prior beliefs and the content speakers can recover from dogwhistled communication, and also learn that the relevant inferences have to be conditioned on the form of the dogwhistle rather than just its extensional content. From the signaling game view, we see that considerations of utility maximization in particular communicative settings can explain the use of dogwhistles, but not how they arise initially. Our aim now is to bring these insights together into a single unified view.

Our ideal, then, is a theory of dogwhistles which (i) has a metalinguistic character, (ii) makes use of background, default information about how speakers use language, and (iii) can be reconciled with a game-theoretic account of the deployment and recovery of dogwhistled content. Fortunately we have the foundation of such a theory available off-the-shelf in [11], who uses just these components to analyze the way emotive underspecification is resolved in emotive adjectives. We will make use of this theory to underpin our refitting of the inferentialist view to solve its problems with extensionally equivalent expressions and integrate it with our game-theoretic model.

[11] considers cases of underspecified emotive content, which include adjectives like *damn* or *fucking*, particle exclamations like *Man!* and ordinary exclamatives like *What a hotel!*. All these expressions can be interpreted either as positive or negative in the right context.

- (4) What a hotel!
- a. We enter the hotel room: ocean view, palatial space, spotless white coverlet, bottle of champagne, etc. \rightsquigarrow **positive interpretation**
 - b. The hotel is a complete dump, roaches, springs coming out of the bed, plus the window doesn’t open and the AC is broken \rightsquigarrow **negative interpretation**

How can a hearer settle on a positive or negative interpretation? And how can a speaker navigate the potential of hearer misunderstanding of her intention? As these questions make clear, this case is a kind of toy version of the general problem of interpretation recovery in language. The strategy proposed by [11] for this is to condition the interpretation of underspecified emotives on hearer guesses (on the basis of her existing beliefs, modeled using a standard probability function) about speaker emotional states and background knowledge about how people use language with respect to their emotional states. Thus, given that a speaker is using an emotive expression with an underspecified interpretation, that (e.g.) in context (4a) she is in a positive emotional state, and that speakers, when they use underspecified emotives, ordinarily use them in a way that matches their intended interpretation with the speaker’s current emotional state, the hearer can conclude that the speaker likely means to communicate a positive meaning with the exclamation.

The basic ingredients of the analysis are the use of normality conditionals formulated in a default logic to formalize the way in which particular states of the

world associate with emotional states on the one hand and with how language is normally used and interpreted on the other. The first kind of conditional is less relevant to the current setting; it's the metalinguistic aspect of the analysis that makes it useful for overcoming the extensional identity problem. For the case above, we might have conditionals like the following (here, '>' is a normality conditional and $\lambda x.Emot(x)$ is a function yielding the emotional state of x , drawn from the set $\{pos, neg\}$):

$$(5) \quad reach_hotel(x) \wedge \exists y[roach(y) \wedge in_room(y)] \wedge \exists z[ac(z) \wedge in_room(z) \wedge broken(z)] > Emot(x) = neg$$

$$(6) \quad reach_hotel(x) \wedge \exists y[champagne(y) \wedge in_room(y)] \wedge \exists z[coverlet(z) \wedge in_room(z) \wedge spotless(z)] > Emot(x) = pos$$

These sorts of world knowledge axioms directly yield the speaker's emotional state, though defeasibly. The use of axioms about metalinguistic content is more general; instead of axioms relating to clearly specified situations like the above, we instead have schema indicating how speaker emotional state ordinarily relates to language use. For example, consider the following axiom from [11], which states that in the absence of defeaters, we can assume that any underspecified emotive expression included in a speaker's utterance should be interpreted in a way conforming with her overall emotional state:

$$(7) \quad (Emot(s^c) = E \wedge Use(s^c, S) \wedge EC \sqsubseteq S) > EC = E$$

Taking the above three axioms together, we are able to arrive at the proper (default) interpretation of the exclamative by chaining the applicable world knowledge axiom with the metalinguistic axiom about interpretation.

Our method for unifying the theories of dogwhistles we have discussed is to import the machinery of defaults about how language is used into our game-theoretic analysis. The result is a theory that blends aspects of [5, 6], improving on both. The core idea is that the default inferences Khoo uses to arrive at the dogwhistle message should be conditioned on the persona the speaker is aiming to present.

[5] provide a game-theoretic account of how speakers can recover the persona a speaker aims to present (to a subset of the audience) given the choice between distinct, though semantically equivalent expressions. When combined with Khoo's insight about the form of dogwhistle inference, we get an account of how speakers and hearers coordinate on the meaning of dogwhistles. This account is given by the logical formula in (8). According to this view, the speaker's use of a dogwhistle with content C and persona p licenses the inference that the speaker wants the hearer to believe R , given that hearer believes that C entails R in a Khoo-style inference. In (8), $[DW]$ indicates the dogwhistle itself *qua* linguistic form; this means that, if the speaker doesn't use the actual dogwhistle, the inference won't follow, so extensionally equivalent expressions won't do the job. In fact, the inference depends on the speaker's persona (because of the clause $Use(s, p, [DW])$), which is connected to the words used, so the extensionally

equivalent case is doubly out. Moreover, this account explains why dogwhistles have the enriched meanings they do, which was the aspect of dogwhistle pragmatics left out of [5]. In particular, they get their meanings from the speaker inviting the hearer, based on the speaker's persona, to make an inference they are prone to make based on the content of the dogwhistle.

$$(8) \quad Use(s, p, [DW]) \wedge Bel(h, \forall x[C(x) \rightarrow R(x)]) > Intend(s, Bel(h, R(x)))$$

for persona p and DW with dogwhistled content C .

Let's consider a specific example, that of *inner city*, which has been our test case throughout. This is a Type 2 dogwhistle which, given the recognition of the speaker's persona (cryptoracist), allows enrichment of the content *inner city* to *urban African American neighborhood*. The following axiom states that, given that a speaker s with persona p uses the dogwhistle *inner city*, and given that the hearer believes that inner city neighborhoods are all African American, then normally the speaker intends the inference from his phrasing to this enriched meaning to be made. This captures the proper meaning of the expression, and does so in a way that won't allow substitution of an extensionally equivalent expression such as *neighborhood in an urban area* to induce the relevant inference, for, given the form of the conditional on which the use of the precise expression [*inner-city*] is required, the inference won't be triggered in the absence of either recognition of the speaker's persona or the use of the dogwhistle itself.

$$(9) \quad Use(s, p, [inner_city]) \wedge Bel(h, \forall x[inner_city(x) \rightarrow urban_AA_neighborhood(x)]) > Intend(s, Bel(h, urban_AA_neighborhood(x)))$$

4 Conclusion

In this paper, we have presented three views of dogwhistle language in formal semantics and pragmatics, each flawed in its own way. The first takes them to be CIs; this theory fails due to the lack of conventional quality in dogwhistles. The second is an inferentialist view that aims to explain the kinds of semantic consequences that arise from using dogwhistles, by conditioning them on the background knowledge of those who hear them. This view is a major advance, but fails to explain the fact that dogwhistles don't have the same effects as other, extensionally equivalent, expressions; this is unexpected if only standard kinds of inference are in play, because such inferences are strictly content-based. The third theory conditions dogwhistled interpretations (that affect or enrich truth-conditional content) on the recovery of speaker personae in the sense of Third Wave sociolinguistics; this theory is successful, but makes no attempt to explain the precise kinds of enrichment that arise with dogwhistles, or why that precise content arises.

The present paper unified the second and third views by making use of metalinguistic default conditionals which condition on the perceived persona of the speaker: since personae are recovered on the basis of the particular expressions

used, the choice of expression is important, and extensionally equivalent items won't induce the same effects. This unification is a step forward in the analysis of dogwhistles, and shows further the importance and usefulness of personae in formal pragmatic work.

There are many future directions for this project. Here are two immediate ones. To our knowledge, there is no formal work on dogwhistles outside of English. We are interested to examine dogwhistles in political speech couched in other languages, and, importantly, in non-Western cultures; the results of such investigation should go further to show the (non)uniformity of Gricean considerations of rationality in speech and communication (cf. [8,9]). An initial test case in this regard might be the Japanese expression *junsui* 'pure', which has a common use in right-wing discourse and dogwhistles views about ethnic purity and Japanese racial homogeneity. Second, we want to look at phenomena which might be viewed as the inverse of dogwhistles in a certain sense: dogwhistles are centered on the recovery of speaker personae, but what about the hearer? Are there expressions which have particular meanings depending on whether the hearer has a particular persona, and lacks them otherwise? We believe so: one such is the phenomenon of *subtweeting*, where a general statement is made which is meant to apply to some specific hearer or set of hearers, possibly just those with a certain property. Thus, hearer self-identification as a target of the subtweet is required for its efficacy. In some cases (perhaps only when the target involves ascription of a property, such as *being a Nazi*), persona identification will be involved; we mean to tease apart some such cases as the next stage in the present project.

References

1. Burnett, H.: Signalling games, sociolinguistic variation and the construction of style. In: The 40th Penn Linguistics Colloquium, University of Pennsylvania (2016)
2. Burnett, H.: Sociolinguistic interaction and identity construction: the view from game-theoretic pragmatics. *Linguistics and Philosophy* (2017, to appear)
3. Eckert, P.: *Jocks and Burnouts: Social Identity in the High School*. Teachers College Press, New York (1989)
4. Elga, A., Rayo, A.: Fragmentation and information access. Manuscript (1966)
5. Henderson, R., McCready, E.: Dogwhistles and the at-issue/non-at-issue distinction. In: Gutzmann, D., Turgay, K. (eds.) *Secondary Content: The Linguistics of Side Issues*, pp. 1–21. Brill (to appear)
6. Khoo, J.: Code words in political discourse. *Philos. Top.* **45**(2), 33–64 (2017)
7. Labov, W.: *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC (1966)
8. Machery, E.: Expertise and intuitions about reference. *Theoria* **73**, 37–54 (2012)
9. Machery, E., Mallon, R., Nichols, S., Stich, S.: Semantics, cross-cultural style. *Cognition* **92**, B1–B12 (2004)
10. McCready, E.: Varieties of conventional implicature. *Semant. Pragmat.* **3**, 1–57 (2010)
11. McCready, E.: Emotive equilibria. *Linguist. Philos.* **35**, 243–283 (2012)

12. Potts, C.: *The Logic of Conventional Implicatures*. Oxford University Press, Oxford (2005)
13. Potts, C.: The expressive dimension. *Theor. Linguist.* **33**, 165–198 (2007)
14. Recanati, F.: *Literal Meaning*. Cambridge University Press, Cambridge (2003)
15. Rickford, J.R., King, S.: Language and linguistics on trial: hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language* **92**(4), 948–988 (2016)
16. Stanley, J.: *How Propaganda Works*. Princeton University Press, Princeton (2015)



Transformational Semantics on a Tree Bank

Oleg Kiselyov^(✉)

Tohoku University, Sendai, Japan
oleg@okmij.org

Abstract. Recently introduced Transformational Semantics (TS) formalizes, restraints and makes rigorous the transformational approach epitomized by QR and Transformational Grammars: deriving a meaning (in the form of a logical formula or a logical form) by a series of transformations from a suitably abstract (tecto-) form of a sentence. Unlike QR, each transformation in TS is rigorously and precisely defined, typed, and deterministic. The restraints of TS and the sparsity of the choice points (in the order of applying the deterministic transformation steps) help derive negative predictions and control over-generation.

The rigorous nature of TS makes it easier to carry analyses mechanically, by a computer. We report on such a mechanical, fully automatic application of TS to a tree bank of FraCAS text entailment problems (generalized quantifier section). Set-theoretic logical formulas derived by TS as meanings for input sentences are submitted to an automatic *first-order* theorem prover to decide entailment. A characteristic feature of our approach is the exhaustive enumeration of quantifier and other such ambiguities.

Overall TS proved just as capable as natural logic in inferences involving a variety of generalized quantifiers. Still open is the problem of mechanically dealing with bare plurals.

1 Introduction

We report on an automatic application of Transformational Semantics (TS) [4,5] to the FraCAS bank of textual inference problems [2].

TS, in a word, is a rigorous, rigid and restrained version of the familiar QR [3], well-expressing covert movements and other ‘logical form’ transformations. A transformation from a (logical) abstract form to a semantic set-theoretic formula is composed from several steps. Each step, such as raising a single quantifier, is deterministic, with no wiggle room. The order of the steps, on the other hand, is non-deterministic. TS has been applied to quantifier ambiguity, scoping islands and binding, crossover, topicalization, inverse linking, and non-canonical coordination [4,5].

Because TS is precisely specified, its transformations can be carried out mechanically, by a computer. The current implementation takes the form of a domain-specific language embedded in Haskell. It was originally intended as a

semantic theory design aid: to interactively try various transformations, observe their results or failures. This paper reports on a different, more real-life application: to fully automatically derive meanings of tree bank sentences and their entailments.

The application to the real-world tree bank is not as simple as it may appear. The input is not a clean, abstract (tecto-grammatical) formula; rather it is a Penn-treebank-like annotated tree. It proved rather messy to transform the latter to the former. Another problem is dealing with quantifier and other ambiguities. In TS, they manifest as choices in the order of applying primitive transformation steps. We resolve to systematically enumerate all possible choices, hence, to expose all ambiguities licensed by TS. The final challenge is expressing the meaning of various generalized quantifiers, modal operators and intensional constructions as a *first-order* formula – so to interface with an automatic theorem prover.

Section 2 describes the whole process in detail: the annotated FraCaS and its annotations in Sect. 2.1; the conversion to the tecto-grammatical form in Sect. 2.2; verifying (type-checking) the form in Sect. 2.3; applying the Transformational Semantics to turn the form to the logical formula in Sect. 2.4; and deciding the entailment in Sect. 2.5. The conversion in Sect. 2.2 from an annotated tree to a (tecto-grammatical) form suitable for semantic analyses is not particular to TS and can be used independently, with other semantic theories. Section 3 details the handling of generalized quantifiers like “several”, “many”, “few”, “at least three” – in particular their representation in *first-order* theories. Section 4 describes other challenges: definite descriptions and intensional verbs like “want”. The challenges not yet overcome are discussed in Sect. 5. We then present the results of applying the TS to the generalized quantifier section of FraCaS.

The complete source code is available at <http://okmij.org/ftp/gengo/transformational-semantics/>.

2 From a Sentence to an Entailment Decision

We illustrate the TS process of deriving the meaning formula (and then, entailments) on a simple example, problem 049 from the FraCaS textual inference problem set [2] (actually, from Bill MacCartney’s modified and converted to XML set¹):

A Swede won a Nobel prize.
Every Swede is a Scandinavian.
A Scandinavian won a Nobel prize.

The goal is to decide if the third sentence is entailed by the first two.

We chose this running example because it is so straightforward: it lets us focus on the mechanics of the transformational framework without getting distracted with theoretical semantic problems. The problems are present, in abundance: see Sects. 3, 4 and 5.

¹ <http://www-nlp.stanford.edu/wcmac/downloads/fracas.xml>.

2.1 Parsing

Fortunately, parsing is not our problem. We take as input not raw text but a tree, annotated according to the Penn Historical Corpora system² (also known as the annotation system for the Penn-Helsinki Parsed Corpus of Early Modern English). The FraCAS corpus in this annotated form has been very kindly provided by Butler [1]. The first two sentences of problem 049 look as follows.

```
( (IP-MAT (NP-SBJ (D A) (ADJ Swede))
  (VBD won) (NP-OB1 (D a) (NPR Nobel) (N prize))
  (PU .))
 (ID 86_JSeM_beta_150530))

( (IP-MAT (NP-SBJ (Q Every) (ADJ Swede))
  (BEP is) (NP-OB1 (D a) (ADJ Scandinavian))
  (PU .))
 (ID 87_JSeM_beta_150530))
```

Fig. 1. The first two problem 049 sentences in the tree bank form

All words including punctuation are tagged with their part of speech: N for singular common noun, NPR for singular proper noun, VBD for verb in past tense, D for determiner, ADJ for adjective, Q for quantifier, etc. Special words “be”, “do” and “have” have their own tags, such as BEP for “be” in present tense. Significantly, the syntactic structure is also annotated: e.g., subject noun phrases as NP-SBJ and object noun phrases as NP-OB1. Phrasal structure, however, is only partly annotated: either because the detailed bracketing is easily derivable or less practically useful, or because phrase boundaries are difficult to determine. (This is the case for VP: even in Modern English the attachment of verbal adjuncts is ambiguous). Dropping all annotations gives the original sentence as it was.

The main attraction of using the Penn Historical annotated data as input is that there is a wealth of such data available: not just for Modern and Historical English but also for many other languages; not just FraCAS but also the Bible, “Moby Dick”, newswires, and many other texts [1]. These annotated sources have high quality. This is a treasure trove of empirical material for TS to work on.

2.2 From the Treebank to the Abstract Form

The input to TS is a (tecto-grammatical) *abstract form*, to be described in Sect. 2.3. Looking ahead, Fig. 2 shows the abstract form for the first two sentences in Fig. 1.

² <http://www.ling.upenn.edu/ppche/ppche-release-2016/annotation/index.html>.

```

cl (a_x (swede entity))           cl (every_x (swede entity))
  (won (a_y nobel_prize))         (is_cn (scandinavian entity))

```

Fig. 2. Abstract-form terms

The first task hence is to convert FraCaS problems to this abstract form. Starting from the treebank annotated data saves us parsing – but not completely. The phrasal structure in treebank trees is only partly exposed: the trees are generally rather flat. The trees have to be pre-processed first to fill in the missing structural annotations and make them binary branching. The preprocessing step also removes punctuation and metadata and normalizes parts of speech annotations eliminating irrelevant for semantic analyses distinctions.

The comparison of Figs. 1 and 2 hints that the conversion from the former to the latter is messy. For example, the NP-SBJ branch in Fig. 1 adjoins the ADJ "Swede" to a determiner. In the abstract form, a determiner has the type $N \rightarrow NP$ and cannot take adjective as the argument. Since "Swede" can also be a common noun (category N), one could regard Fig. 1 as mis-annotated. Butler has suggested, however, to assume that such phrases omit the trivial noun, "entity". We have to put it in.

Thus the input annotated tree requires significant preprocessing:

- Normalize all strings to lower case and remove punctuation;
- Abstract away tense: replace both VBP and VBD tags (for present- and past-tense verbs) with just VB;
- Insert the dummy N "entity" as explained earlier;
- Treat "a few", "one of the", etc. as atomic quantifiers: Q "afew";
- Take "The world's greatest" to be the intensified "greatest" rather than a possessive phrase per the original annotation;
- Regard abstract common nouns with the null article (e.g., "charity") as proper-name-like NPs;
- Treat the indefinite, D "a", as a quantifier Q "a";
- Collect noun clusters and mark them as a compound noun: e.g., "nobel_prize";
- In the same manner, turn idiomatic-like phrases such as "really great" and "travel freely" into compounds ADJ "really_great" and VB "travel_freely";
- Simple definite descriptions like "the world" and "the report" function quite like proper nouns, so tag them as such (Sect. 4 discusses definite descriptions in detail);
- Regard "There is NP" as "NP exists" and "There are NP" as "Several NP exist";
- Simplify subject relative clauses by removing traces;
- Collect all noun complements within an NP and introduce new nodes **nc** (for singular common noun with a list of complements) and **ncs** (for plural common noun). Unlike the original Penn Historical annotation tags, ours are in lowercase;

- Recognize copular clauses and tag them with `cop`;
- Regard “some” with a plural restrictor as “several”;
- Introduce nodes `tv` and `tv-app` for a transitive verb with an argument, so to make the tree binary.

All these transformation steps are programmed as top-down macro-tree transducers and applied in the order described. The Haskell code³ closely follows the macro-tree transducer notation and is hence easy to modify and extend. Figure 3 displays the result of the preprocessing: binary branching trees with the minimum needed information for our analyses. The preprocessed trees are straightforward to convert to the abstract terms in Fig. 2.

```
(IP-MAT
  (NP-SBJ (Q a) (nc (adj swede) (N entity)))
  (tv-app (tv won) (NP (Q a) (N nobel_prize))))

(IP-MAT
  (NP-SBJ (Q every) (nc (adj swede) (N entity)))
  (cop (Q a) (nc (adj scandinavian) (N entity))))
```

Fig. 3. Preprocessed treebank sentences

The transformations from the Penn Historical treebank trees to the “tectogrammatized” trees in Fig. 3 and then to the terms in Fig. 2 are not specific to TS and can be used independently, with other semantic theories.

2.3 From Ad Hoc to Meaningful Transformations

The end result of the preprocessing transformations just described, and the starting point of TS transformations in Sect. 2.4 is the abstract form, a tectogrammatized form of a sentence. The abstract form for the first two sentences of the running example was shown in Fig. 2, repeated below as Fig. 4. The preprocessing transformations are ad hoc; little is guaranteed about them. TS transformations in Sect. 2.4, as we soon see, have firmer foundations and a certain degree of correctness. Naturally they demand their input be ‘sane’, that is, well-typed. Hence the next step in our processing chain is type checking the produced abstract form.

The abstract form is defined (see [5]) to be a term in a multisorted algebra, whose sorts are familiar categories. The type of a composite term is determined from the type of its constituents using the typing rule of functional application (i.e., *modus ponens*). As for the primitives in our example, `entity` and `nobel_prize` have the type N , `swede` and `scandinavian` have the type ADJ (equivalent to $N \rightarrow N$), the transitive verb `won` is typed as $NP \rightarrow VP$. Determiners (quantifiers) such as `a_x` and `a_y` have the type $N \rightarrow NP$. They are

³ <http://okmij.org/ftp/gengo/transformational-semantics/Treebank.hs>.

```

c1 (a_x (swede entity))           c1 (every_x (swede entity))
  (won (a_y nobel_prize))         (is_cn (scandinavian entity))

```

Fig. 4. Abstract-form terms

indexed for identification. The predicational copula constant `is_cn` has the type $N \rightarrow VP$; the constant `c1` of the type $NP \rightarrow VP \rightarrow S$ forms a clause.

Implementation-wise, the preprocessing transformations in Sect. 2.2 eventually produce an abstract-form term in Haskell notation, which is then ‘loaded’ into the Haskell interpreter – at which point it is type-checked and verified to have the type S . Although the type checking does not say that the initial treebank trees or our preprocessing steps are ‘correct’, it does check they are not obviously wrong.

2.4 TS Transformations

Even with syntactic details abstracted away, the terms in Fig. 4 are difficult to immediately interpret as logical formulas. The problem is the embedded quantifiers whose scope is not apparent. Making the scope explicit is the job of TS transformations.

Befitting their name, TS transformations transform an abstract term to a form in which the quantifier scope is clearly marked: specifically, by pulling an embedded quantifier (along with its restrictor) in front of the subterm over which it takes scope. TS transformations hence do quantifier raising. Unlike QR [3], however, TS transformations are rigidly and rigorously defined (and also typed). A transformation that raises a quantifier into an inappropriate place or produces an ill-formed logical formula cannot even be written (i.e., accepted by the Haskell type checker). Thus TS assures a degree of correctness. Formally, TS transformations are described in [5]; Fig. 5 illustrates them on our running example.

Input	<code>c1 (a_x (swede entity)) (won (a_y nobel_prize))</code>
Raising <code>a_x</code>	<code>Ex (swede entity) (c1 x (won (a_y nobel_prize)))</code>
Raising <code>a_y</code>	<code>Ex (swede entity) (Ey nobel_prize (c1 x (won y)))</code>

Fig. 5. Steps of TS transformations for the first sentence of the running example

TS transformations are composed from primitive, deterministic steps, such as raising a single quantifier. The raised quantifier like `a_x`, leaving behind the constant `x`, is represented at its new place by the constant `Ex`, applied to the restrictor and the term over which the quantifier takes scope. The landing place is rigidly fixed: right over the closest `c1` subterm. In the final result, the last line of Fig. 5, all quantifier scopes are explicit; such term is straightforward to

interpret as a logical (set-theoretic) formula, as detailed in Sect. 2.5. The two transformation steps can be applied in the opposite order, giving a different but logically equivalent result.

Our interactive Haskell implementation lets the users choose the steps and their order. The automatic implementation is programmed to try all possible steps in every order. Not every sequence of steps is successful (that is, ends in a logical formula): some analyses can be blocked.

2.5 Deciding Entailment

The end result of TS transformations is the abstract form that can be easily interpreted logically – or set-theoretically. Figure 6 shows that interpretation, for all three sentences in our running example (FraCaS problem 049).

The meaning of a sentence is given as a set-theoretic proposition written in first-order logic. For example, “Every Swede is a Scandinavian” is interpreted as the proposition that the intersection of `swede` and `entity` is included in the intersection of `scandinavian` and `entity` (that is, `swede` is a subset of `scandinavian`). Likewise, “A Swede won a Nobel prize” states that the intersection of `swede` and `entity` has an element in relation `won` with `nobel_prize`.

The logical formulas are written in the TPTP format supported by almost all automated theorem provers (<http://www.tptp.org/>). The premises (the first two sentences of FraCaS problem 049) are marked as axioms and the putative entailment as a conjecture. The TPTP formulas are written in Prolog-like syntax, with the variable names capitalized. The universal quantification $\forall x$ is notated as `![X]`: and the existential $\exists x$ as `?[X]`; `&` stands for conjunction and `=>` for implication.

```
fof(s1,axiom,
  ?[Y]: (in(Y,nobel_prize) &
        (?[X]: ((in(X,swede) & in(X,entity)) & rel(Y,won,X)))).

fof(s2,axiom,
  ![X]: ((in(X,swede) & in(X,entity)) =>
        (in(X,scandinavian) & in(X,entity))).

fof(c,conjecture,
  ?[Y]: (in(Y,nobel_prize) &
        (?[X]: ((in(X,scandinavian) & in(X,entity)) & rel(Y,won,X)))).
```

Fig. 6. Problem 049 in TPTP

We submit the Fig. 6 code to E, the automatic theorem prover for the first-order logic with equality [6]. It finds and displays the entailment proof.

3 Generalized Quantifiers in First Order

The story so far of transforming an annotated tree to a set-theoretic proposition has been (save for the preprocessing step, perhaps) easy-going – too easily. We now describe the pitfalls and problems, some of which are still open. The first problem is the generalized quantifiers.

The FraCAS corpus exhibits quantifier phrases that go beyond the mere existence and universality: “several”, “a few”, “few”, “at least three”, “five”, “most”. We describe our treatment of such generalized quantifiers – in particular, their representation in a *first-order* theory. Such a representation lets us use off-the-shelf, mature first-order theorem provers to *automatically* decide entailment. To be sure, there are sentences whose meaning cannot be represented in a first-order theory, but they are not common (in particular, they do not occur in FraCaS).

Expressing the meaning of the whole variety of quantified English phrases in a first-order theory may seem like a hard problem; yet it turned out clear-cut. As far as TS transformations are concerned, generalized quantifiers pose no difficulty: they are raised, along with their restrictors, just like the plain existential and universal quantifiers.

As an example, consider problem 076:

Few committee members are from southern Europe.

Few female committee members are from southern Europe.

The first sentence has the following abstract form:

```
c1 (few_x committee_member) (is_pp (from southern_europe))
```

where `committee_member` is of type N , `from southern_europe` has the type PP and `is_pp` has the type $PP \rightarrow VP$. A TS transformation raises `few_x` as usual, giving:

```
Few_x committee_member (c1 x (is_pp (from southern_europe)))
```

The problem comes from trying to express the raised generalized quantifiers in a first-order theory. The problem is not just writing a formula but also being able to decide its entailments.

Set-theoretically, we interpret `few` as *an uninterpreted relation* – in our case, between the set of committee members and the set of people from southern Europe:

```
fof(s1,axiom,
  (![X]: (in(X,sks11) <=> rel(southern_europe,from,X))) &
  few(committee_member,sks11)).
```

The former set is denoted by the constant `committee_member`; for the latter, we introduce the fresh constant `sks11` accompanied by the postulate that all members of the set `sks11` are also in the relation `from` with `southern_europe`. Likewise, the second sentence of problem 076 is represented by the conjecture:

```

fof(c,conjecture,
  ((([X]: (in(X,skc3) <=> rel(southern_europe,from,X))) &
   ([Xc1]: (in(Xc1,skc2) <=> (in(Xc1,female) & in(Xc1,committee_member))))))
  => few(skc2,skc3)).

```

where the constant `skc2` is the name for the intersection of `female` and `committee_member`, and `skc3` is another name for the set of people from southern Europe (`skc3` and the earlier `sks11` are hence extensionally equivalent).

The relation `few` is uninterpreted: the E prover knows nothing about it. To check if the conjecture holds given the premise `s1`, some properties of `few` are required – such as downward monotonicity in its first argument⁴:

```

fof(few1,axiom,[P,P1,Q,Q1]:
  ((few(P,Q) & seteq(Q,Q1) & imply(P1,P)) => few(P1,Q1))).

fof(imp,axiom, [P,Q]: (imply(P,Q) <=> ([X]: (in(X,P) => in(X,Q))))).
fof(seteq,axiom, [P,Q]: (seteq(P,Q) <=> ([X]: (in(X,P) <=> in(X,Q))))).

```

That is, if `few(P,Q)` holds for some sets `P` and `Q`, and `Q1` is extensionally equivalent to `Q` and `P1` is a subset of `P`, we assert that `few(P1,Q1)` also holds. The axioms `imp` and `seteq` define the subset relation and the extensional equivalence. With these axioms, the E prover easily determines that the conjecture of problem 076 holds.

Other generalized quantifiers are treated similarly, as uninterpreted relations with axioms defining their monotonicity and other properties. As another, somewhat surprising example, let's consider problem 002:

Every Italian man wants to be a great tenor.

Some Italian men are great tenors.

There are Italian men who want to be a great tenor.

When transcribing the original FraCaS corpus to XML, Bill MacCartney added: “Note that second premise is unnecessary and irrelevant”. At first glance, who can doubt it: the fact that some Italian men are great tenors has no bearing on wanting to be one. It is only when we set out to prove the conjecture that we see the subtlety.

The set-theoretic meaning of the three sentences in problem 002 is as follows:

```

fof(s1,axiom,
  ![X]: ((in(X,italian) & in(X,man)) => in(X,want_greattenor))).

fof(s2,axiom,
  ((([X]: (in(X,sks22) <=> (in(X,great) & in(X,tenor)))) &
   ([Xs21]: (in(Xs21,skitalianman) <=>
    (in(Xs21,italian) & in(Xs21,man)))))) &
  several(skitalianman,sks22))).

```

⁴ Bill MacCartney, the author of the XML-annotated FraCaS, noted that the original FraCaS authors must have taken “few” to mean a small absolute number – downward-monotone in the first argument.

```

fof(c,conjecture,
  ((([X]: (in(X,skc2) <=> in(X,exist))) &
    (![Xc1]: (in(Xc1,skitalianmanwant_greattenor) <=>
      ((in(Xc1,italian) & in(Xc1,man)) & in(Xc1,want_greattenor))))))
  => several(skitalianmanwant_greattenor,skc2))).

```

The axiom for several

```

fof(several1,axiom,! [P,P1,Q,Q1]:
  ((several(P,Q) &
    (![X]: ((in(X,P) & in(X,Q)) => ((in(X,P1) & in(X,Q1))))))
  => several(P1,Q1))).

```

states that if `several(P,Q)` holds for two sets `P` and `Q`, then it holds for any other sets `P1` and `Q1` that have at least the same intersection. Without the second sentence of problem 002, the prover was unable to make the inference. A moment of thought tells why: the premise “Every Italian man wants to be a great tenor.” does not actually imply that there exist at least several Italian men (even if we are to assume that “every” has existential import). It is the seemingly irrelevant “Some Italian men are great tenors.” that asserts their existence, which is needed to reach the conclusion in the problem.

4 Further Challenges

Generalized quantifiers are not the only and not the biggest challenge that we have encountered so far. We now describe other problems in trying to represent the meaning of FraCaS sentences as first-order set-theoretic propositions.

Problem 002 discussed in Sect. 3 has another complication. Here its first sentence again:

Every Italian man wants to be a great tenor.

showing off the intensional “want” with the infinitival complement. Intensionality is a thorny subject. We tackle it rather unconventionally: syntactically, without resorting to possible worlds. To wit, we represent the predicate of our sentence as an uninterpreted constant `want_greattenor`, spelling out the wanted property. Clearly it does not imply (or implied by) `want_great` and `want_tenor`, exhibiting the desired referential opacity.

Definite descriptions are common among FraCaS problems as they are in English in general. A simple definite description like “the report” is regarded as an uninterpreted constant `the_report` – quite like a proper name. There are many more complicated cases in FraCaS, for example, problem 017:

An Irishman won the Nobel prize for literature.

An Irishman won a Nobel prize.

which we understand as

```

fof(s1,axiom,?[X]:
  ((in_prominent(the_sks12,sks12) &

```

```
(![Xs11]: (in(Xs11,sks12) <=>
  (rel(literature,for,Xs11) & in(Xs11,nobel_prize)))) &
(in(X,irishman) & rel(the_sks12,won,X))).
```

```
fof(c,conjecture,
  ?[X]: (in(X,irishman) & (?[Y]: (in(Y,nobel_prize) & rel(Y,won,X))))).
```

introducing two uninterpreted constants `sks12` and `the_sks12`. The first sentence of problem 017 asserts that `sks12` is extensionally equivalent to the intersection of `for-literature` and `nobel_prize`, and that `the_sks12` is the prominent element of that set. The prominent membership has the expected properties of unique membership:

```
fof(prominent1,axiom,! [X,P]: (in_prominent(X,P) => in(X,P))).
fof(prominent2,axiom,! [X,Y,P]:
  ((in_prominent(X,P) & in_prominent(Y,P)) => (X=Y))).
```

5 Open Questions

Although the generalized quantifier section of FraCaS contains rather simple sentences, they already pose problems that we are yet to solve. The most prominent, and the frequent, is plurality – especially bare plurals. In phrases like “there are Italian men” one may assume an implicit quantifier “several”. However, such assumption fails, for example, for problem 013:

Both leading tenors are excellent.

Leading tenors who are excellent are indispensable.

Both leading tenors are indispensable.

If we are to treat “leading tenors” as “several leading tenors”, the third sentence clearly does not follow from the other two – yet native speakers report the entailment. The example illustrates the well-known (and well-argued about) ambiguity of bare plurals between generic and existential readings. Literature has no shortage of proposals for analyzing bare plurals. What is lacking is the robust method that can let us reliably do textual inferences without any human intervention.

6 Evaluation

We have applied TS to the entailment problems in the generalized quantifier section of FraCaS, which has 80 problems. So far⁵ we have handled 36 of them – in all the cases obtaining the agreement with the FraCaS entailment results. Specifically, out of 16 problems of FraCaS Subsect. 1.1 (conservativity) we successfully handled 12.

Currently we cannot deal with many cases of mass nouns, bare plurals or definite plural descriptions, as discussed in Sect. 5. For the remaining problems, it is just the matter of writing axioms for generalized quantifiers.

⁵ As of February 2018.

7 Conclusions and Reflections

We have described the transformational-semantics-based approach to text entailment. It transforms an input Penn treebank-like annotated tree to a tecto-grammatical form, and then to a set-theoretic logical formula submitted to a first-order theorem prover. The approach is fully automatic and exhaustively enumerates all ambiguities licensed by TS.

We have applied TS to the generalized quantifier problems of FraCaS. For all the problems we can currently apply TS to, we have obtained the expected entailments or the lack of entailments. Thus, TS does work on a tree bank, and is capable of natural logic inferences with generalized quantifiers.

Applying TS to the FraCaS corpus has been a humbling experience. Facing a bank of sentences rather than a few examples – and analyzing them automatically, and being able to perform inferences – turns out quite a bigger challenge than I could imagine. Disappointingly, the quantifier ambiguity, the original motivation for TS, was conspicuously absent in the FraCaS fragment at hand. Quantifier puzzles are indeed uncommon. Plurals, on the other hand, appear all the time.

Plurals and mass nouns, hence, are the most pressing problems for the future work. We would also like to extend TS to event semantics, so to handle tense and aspect-related parts of FraCaS.

Acknowledgments. I am very grateful to Alastair Butler for providing the FraCaS data in the treebank form and explaining the annotation system – and for his many encouragements.

References

1. Butler, A.: The treebank semantics parsed corpus (2017). <http://www.compling.jp/tspc/>
2. Cooper, R., Crouch, D., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., Konrad, K.: Using the framework. Deliverable D16, FraCaS Project (1996)
3. Heim, I., Kratzer, A.: *Semantics in Generative Grammar*. Blackwell Publishers, Oxford (1997)
4. Kiselyov, O.: Applicative abstract categorial grammars in full swing. In: Otake, M., Kurahashi, S., Ota, Y., Satoh, K., Bekki, D. (eds.) *JSAI-isAI 2015*. LNCS, vol. 10091, pp. 66–78. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-50953-2_6
5. Kiselyov, O.: Non-canonical coordination in the transformational approach. In: Kurahashi, S., Ohta, Y., Arai, S., Satoh, K., Bekki, D. (eds.) *JSAI-isAI 2016*. LNCS (LNAI), vol. 10247, pp. 33–44. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61572-1_3
6. Schulz, S.: System description: E 1.8. In: McMillan, K., Middeldorp, A., Voronkov, A. (eds.) *LPAR 2013*. LNCS, vol. 8312, pp. 735–743. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-45221-5_49



Derived Nominals and Concealed Propositions

Ilaria Frana¹(✉) and Keir Moulton²

¹ University of Enna “Kore”, Enna, Italy
ilaria.frana@unikore.it

² Simon Fraser University, Burnaby, Canada

1 Introduction

Vendler [19] described derived nominals (DNs) like *the collapse/ing of the Germans* as ambiguous between event denoting expressions and proposition denoting expressions. DNs can combine with event-selecting predicates (1a), like *gradual*, which bona fide propositional *that*-clauses or fact-denoting expressions cannot (1c), and have event-readings. ((1a) can be paraphrased as ‘the event of the Germans collapsing was gradual’.) DNs can also combine with proposition-selecting predicates like *aware of* (2a) which always also allow finite complements (2b) and, in such cases, have propositional readings—(2a) and (2b) are synonymous. We call DNs in the latter cases Concealed Propositions (ConPs), and we defend the idea that they are analogous in important respects to concealed questions (CQs).¹ Here we argue against Vendler’s Ambiguity Hypothesis (3) and defend an analysis of DNs in which they uniformly denote (or quantify over) events. In doing so, we overcome a challenge, discovered by Zucchi [20], to the unambiguous event approach, and provide an analysis to both definite and quantified DNs. We show that a copy-theoretic account overcomes the problem and aligns ConPs with concealed questions (CQs) in the analysis of Frana [4, 5].

- (1) a. The collapse/ing of the Germans was gradual/sudden/fast.
b. #(The fact) that the Germans collapsed was gradual/sudden/fast.
- (2) a. John knew/was informed /was aware of the collapse/ing of the Germans. (ConP)
b. John knew/was informed/was aware (of the fact) that the Germans collapsed.
- (3) The Ambiguity Hypothesis
Derived nominals are ambiguous between eventualities and propositions.

This project was supported in part by a Social Science and Humanities Research Council of Canada Insight Grant (#435-2015-0454) awarded to Junko Shimoyama and Keir Moulton.

¹ One must be careful with the gerundive forms. As Vendler points out, the verbal gerunds (with an accusative-case marked object rather than *of*) are not possible as arguments of *occur* and *slow* e.g. *John’s singing *(of) the Marseillaise was slow.* ([12, p. 90]). Our discussion is limited to derived nominalizations and nominal gerunds.

Depending on the selection properties of the embedding predicate, either one or the other interpretation is available (or both, if the predicate selects for both propositions and events).

2 Problems for the Ambiguity Hypothesis/Propositional Approach

2.1 The Overgeneration Problem

Zucchi raises an important empirical point against the ambiguity hypothesis. If, as argued by Vendler, DNs were ambiguous between a proposition interpretation and an event one, then we would expect that whenever a DN appears as the object of a proposition-selecting predicate, it should have an interpretation that is semantically equivalent to that of a finite clause. As Zucchi shows, this is not the case with verbs like *remember*: while (5) is compatible with a scenario in which John did not witness the event, but was simply informed about it, (4) is not. In other words, if DNs were ambiguous between denoting events and propositions, then (4) should have a reading synonymous to (5), but that reading is missing.

- (4) John remembers Mary's arrival. (Zucchi 1993)
 (5) John remembers (the fact) that Mary arrived (because he was told so).

If DNs uniformly denote events, and event-selecting *remember*, unlike proposition-selecting *remember*, carries witness requirements, then the contrast in (5) follows. The overgeneration problem extends to a range of verbs type. It is known that verbs of perception reports, like *see* or *hear*, can deliver direct or indirect perception reports, depending on the type of complement [1]. If DNs were ambiguous between proposition and event meanings, then perception verbs should deliver indirect perception reports with DNs, contrary to fact:

- (6) a. Julia saw that the package arrived from Amazon, but she didn't witness the arrival (e.g. she saw a box sitting on her doorstep)
 b. John heard that the doorbell rang (his kid had to tell him).
 (7) a. Julia saw the arrival of the package from Amazon, #but she didn't witness the arrival.
 b. John heard the ring of the doorbell (#his kid had to tell him).

The problem extends to other predicates which don't allow DNs to 'mimic' propositional meanings. As discussed in [11] the two sentences below are not truth-conditionally equivalent: while (8) conveys that Nora offered an explanation to the fact that Fido barked (i.e. to the question "Why did Fido bark?"), (9) can be used to report a situation in which Nora uttered "Fido barked" in response to another question. In this case, "Fido barked" is not the thing explained (the explanandum), rather it is what Nora said in the course of explaining something else (e.g. why the burglar ran off).

- (8) Nora explained the fact that Fido barked. *explanandum*
 (9) Nora explained that Fido barked. *explanans*

If DNs were ambiguous between denoting events and propositions, then they should have a propositional/explanans interpretation, contrary to fact:

- (10) Why does everyone look so happy?
 a. John explained that Sally won, but not how.
 b. John explained the fact that Sally won, #but not how.
 c. John explained Sally's win, #but not how.

2.2 Factivity

We add another empirical point against the ambiguity analysis. Let's start with the observation that, unlike predicates like *know* and *be aware of*, verbs like *tell* and *inform* are not factive when their complement is a *that*-clause, i.e., the proposition expressed by their complement does not have to be true in order for the whole sentence to be true. Thus, while (11a) feels contradictory, (11b) and (11c) do not.

- (11) a. Julia knew that Cicero died, #when in fact he was alive.
 b. Julia was informed that Cicero died, when in fact he was alive.
 c. Antonio told Cicero that Julia arrived, when in fact she hasn't arrived.

Interestingly, when the complement of *tell* and *inform* is a DN with a ConP-reading, the sentence carries a factive commitment, as shown by the fact that the examples below feel contradictory:

- (12) a. Julia was informed of Cicero's death, #when in fact he was alive.
 b. Antonio told Cicero of Julia's arrival, #when in fact she hasn't arrived.

The fact that *tell* and *inform* are factive when they occur with DNs, but not factive when they occur with propositional *that*-clauses is a mystery if ConPs were propositions, as defended by the ambiguity analysis.

One might wonder whether the factivity effect observed with ConPs is due to the existence presupposition carried by the definite, projecting out of the intensional context. We have several reasons to believe that is not so. First, this would not explain the contrast between (13a) and (13b). If the markedness of (13b) were due to the fact that the presupposition of existence of the definite projects out of the intensional context, thus leading to a contradiction with the continuation, the same should be true of (13a).

- (13) a. Romeo was informed that the delivery of his love letter (to Juliet) went through, but in fact that never happened.
 b. Romeo was informed of the delivery of his love letter (to Juliet), #but in fact that never happened.

Second, as noted in the literature (e.g., [9]) presuppositions that project from attitudes (14a) can be canceled as in (14b):

- (14) a. Mary believes Smith’s murderer escaped.
 [presupposition: there is a unique individual who murdered smith]
 b. Mary mistakenly believes that someone murdered Smith, and she believes that Smith’s murderer escaped.

If the “factivity” effect is just due to the definite we should be able to cancel it in a way analogous to (14b). That this is not true is demonstrated in (15).

- (15) Mary believes that Jocasta arrived (when in fact she hasn’t). She then told me of Jocasta’s arrival.

The implication that Jocasta arrived does not seem to be canceled here, suggesting that the factivity doesn’t come from the definite, but from something else. To conclude, if ConPs were propositions, we would not expect the lexical entries of the embedding predicate to encode different requirements (witness/factivity) for *that*-clauses and DNs, given that, at a level of semantic interpretation, these two types of syntactic complements would be mapped to the same type of semantic object, namely a proposition. If, on the other hand, ConPs were not propositions, then the observed differences would no longer be a mystery: the lexical entry of the predicate could encode different requirements, depending on the type of semantic object it composes with (a proposition vs. an event).

3 The Event Approach

3.1 Zucchi’s Proposal

The challenge, then, is to understand how certain predicates can take DNs and “mimic” a propositional interpretation. Zucchi’s answer to that challenge is that DNs uniformly denote events and they come to “mimic” propositional interpretations—to denote ConPs in our terms—by a manipulation in the entry of the selecting verb. For instance, the entry for event-selecting *inform* in (16) allows Zucchi to derive a propositional interpretation without assuming that DNs denote propositions.

- (16) a. John is informed of Mary’s arrival. (= J. is informed that M. arrived)
 b. $\llbracket \text{be informed of}_E \rrbracket = \lambda e. \lambda x. x \text{ is informed (OCCUR}(e))$
 c. $[\text{The event of M’s arrival}]_i$ is such that J. is informed that e_i occurred

At the core of *be informed of_E* is still the meta-language relation *inform*, which describes a relation between individuals and propositions just in the way English CP-taking *inform* does. It is just that the proposition is derived by applying the individual event argument to the predicate OCCUR. Zucchi suggests an analogous shift for predicates like *be aware of*.

3.2 The Problem of Co-extensional Events

Zucchi's analysis treats the DN complement of the verb as saturating an individual event argument. In that respect, it will be a transparent position, just like the internal argument of direct perception verbs.² Recall that [1] (see also [8]) showed that we can capture the epistemic neutrality of direct perception complements on the hypothesis that they saturate an individual argument slot. That will capture the fact that replacement of extensional equivalents preserves truth in direct perception reports:³

- (17) a. Caius witnessed the death of Caesar.
 b. The death of Caesar is the murder of Caesar.
 c. \Rightarrow Caius witnessed the murder of Caesar.

If (17a) and (17b) are true, you have to assent to the conclusion in (17c). Verbs like *witness* select an individual event regardless of that event's description (as a death or murder), and so if the subject sees that event truth is ensured. This, of course, is not true of opaque environments. It turns out that ConPs sit in opaque environments, as shown in (18) (modeled after [13]).

- (18) a. Caius was informed of the death of Caesar.
 b. The death of Caesar is the murder of Caesar
 c. \nRightarrow Caius was informed of the murder of Caesar.

Zucchi himself discusses entailment patterns analogous to (18) as a fatal problem for his analysis. Given that Caesar was murdered, then the murder of Caesar and the death of Caesar are the same event.⁴ However, there is a possible interpretation under which (18a)–(18b) do not entail (18c). Zucchi's analysis does not give justice to such intuition:

² Here, we use the terms transparent/opaque in the sense of [2]: an expression is said to be transparent if its descriptive content is evaluated at the utterance world.

³ Barwise discusses examples in which the type of complement taken by perceptual *see*, whether a naked infinitive or a *that*-clause, disambiguates between epistemic and non-epistemic readings of the predicate, with only the former allowing for non-epistemic interpretations:

- (i) a. Ralph saw a spy hiding a letter under a rock, but thought she was tying her shoe.
 b. Ralph saw that a spy was hiding a letter under a rock, #but thought she was tying her shoe.

⁴ For people unconvinced by this premise, Zucchi offers the following example:

- (i) a. Oedipus was informed of the arrival of Jocasta.
 b. Unbeknownst to Oedipus, Jocasta is his mother. Hence, the arrival of Jocasta is the arrival of Oedipus mother.
 c. Oedipus was informed of the arrival of his own mother.

- (19) [The actual event of Caesar’s death]_i is such that Caius is informed that e_i occurred

On the other hand, if DNs could denote propositions, then the lack of entailment would follow: Caius was informed that Caesar died does not entail that Caius was informed that Caesar was murdered.

Zucchi himself discusses a hypothetical easy fix for this issue, namely to change the lexical entry of event-selecting *inform* so that it combines with a generalized quantifier. As shown below, the event descriptor is no longer quantified in, hence the entailment is no longer predicted.

- (20) Revised entry for *inform* (to be revised)
 a. $\llbracket \text{be informed of}_E \rrbracket = \lambda Q.\lambda x. x \text{ is informed (Q OCCUR)}$
 b. Caius is informed that [the event of Caesar’s death] occurred

However, Zucchi shows the problem runs deeper and entailment patterns analogous to (13) can be reproduced with quantified DNs as well. A quantified DN can take wide scope in terms of its quantification force but it is still opaque on the event description. Zucchi demonstrates this with nouns quantified by *only three*.

- (21) John was informed of only three arrivals of Mary.

In this case though, we don’t want to interpret the quantified event description in the scope of *inform* because (21) doesn’t mean:

- (22) J. is informed that [only three arrivals of Mary] occurred

(21) can be true if John was never told of the number of arrivals of Mary; rather, it means that for (only) these three arrivals, was he informed of them. So we want the quantificational force to scope out as shown in (23) (where X ranges over whatever semantic type *only three ...* denotes).

- (23) [Only three arrivals of Mary] λX [J. is informed that [occurred(X)]]

The problem though is that this takes the event description out of the intensional scope of the matrix verb, predicting that substitution of extensional equivalents will be possible. This is not correct as we now show by demonstrating that even when the quantificational force is interpreted outside the scope of the intensional operator, the event description is interpreted inside the intensional operator, i.e. must be opaque. Zucchi himself concluded as much but we want to demonstrate this fact with examples that differ in two ways from Zucchi’s original. First, we are going to use the universal *every* in what follows, since its contribution is easier to encode than numerals modified by *only*. Second, our test of opacity

will focus squarely on the event description itself (as we did above with the co-extensional *death of Caesar* and *murder of Caesar*).⁵

Assume that Charlie is attending a magic show. During the show, he sees the magician make a rabbit disappear several times. Each disappearance of the rabbit actually consists of a quick jump of the rabbit inside a box, which his eyes do not register (and there were no other jumping-inside-the-box events). In this scenario, (24a)–(24b) do not entail (24c). However, given that each ‘disappearing event’ is also a ‘jumping event’ (and vice-versa), the event-analysis predicts the entailment.

- (24)
- a. Charlie knew of/was aware of every disappearance of the rabbit.
 - b. Every disappearance of the rabbit was a jumping of the rabbit inside the box.
 - c. \nRightarrow Charlie knew of/was aware of every jumping of the rabbit inside the box.

We have reached an impasse. On the one hand, we have evidence against the ambiguity approach (the overgeneration problem and our argument with factivity); on the other hand, there seems to be a fatal problem with the unambiguous event-approach. To summarize the problem is the following. We need the event description to be **part of the propositional content** of the argument of the verb, to account for the failure of substitution above. But, at the same time, we want the nominalization to scope out for the purposes of its quantificational determiners.⁶

4 Toward a Solution to the Impasse

In this section we propose a solution to the impasse that preserves Zucchi’s original proposal that DNs denote events, rather than propositions and, at the same time, resolves the tension between the need to QR the DN and the fact that the DN must receive an opaque interpretation. The solution extends Frana’s [4,5] analysis of CQs to the domain of ConPs.

4.1 Concealed Questions

CQs are nominal arguments of (certain) question-embedding verbs that can be paraphrased as questions/propositions ([5,6,10,15], a.o.). Some examples and

⁵ In some examples, Zucchi contrasts *arrival of Jocasta* with *arrival of Oedipus’ mother*. These are also co-extensional event descriptions, in virtue of the co-extensionality of *Jocasta* and *Oedipus’ mother*. But this co-extensionality could arise by interpreting these object nominal expressions transparently (i.e. “at the utterance evaluation world”). We really need to check the event description, since generally the “main” predicate must be interpreted opaquely in opaque environments (i.e. its world argument can’t be supplied by the utterance context).

⁶ There is a literature of so-called wide scope, opaque interpretations [18]. We may be seeing instances of such thing, although we leave this for future research.

their paraphrases are given below (assume that the actual price of the new iPhone is \$800 and that the kind of wine Clara likes the most is Pinot Grigio).

- (25) a. Clara knows *the price of the new iPhone*.
 b. Clara knows what the price of the new iPhone is/that the new iPhone costs \$800.
- (26) a. Gianni can't remember *the kind of wine Clara likes the most*.
 b. Gianni can't remember what kind of wine Clara likes the most/that the wine Clara likes the most is Pinot Grigio.

CQs display interesting similarities with ConPs: although syntactically DPs, they can serve as arguments of certain question/proposition selecting verbs and can be paraphrased by questions/propositions; just like ConPs, when they occur with verbs like *tell* or *inform* they impose a factivity commitment (“Mary told John the place where Luisa had gone, #but she turned out to be mistaken”); they also occupy intensional argument positions, thus not allowing for substitution of equivalents. As shown in (27), knowing what the price of the new iPhone is does not entail knowing what the price of a 4-year membership at the local gym is, even if the two definite descriptions happen to be co-extensional at the actual world.

- (27) a. Clara knows the price of the new iPhone.
 b. The new iPhone costs the same as a 4-year membership at the local gym.
 c. \nRightarrow Clara knows the price of a 4-year membership at the local gym.

Thus, CQs are intensional objects. The question is what kind of intensional objects are they? One popular answer to this question, which traces back to [6], is that CQs denote intensions of individuals, i.e., individual concepts (ICs), functions from possible worlds into individuals:

- (28) *the price of the new iPhone* _{$\langle s,e \rangle$}

$$\begin{bmatrix} w_0 \rightarrow \$800 \\ w_1 \rightarrow \$900 \\ w_2 \rightarrow \$1200 \\ \vdots \quad \quad \quad \vdots \end{bmatrix}$$

Informally, a sentence ‘a knows/is aware of the CQ’, construed as an IC, gives us true iff the value that the concept yields at the actual world equals the value that the concept yields at all of the attitude holder doxastic alternatives (for instance, ‘John knows/is aware of the price of the new iPhone’ is true iff the concept in (28) outputs the same value at the actual world and at each of John’s belief worlds). Since the price of the new iPhone and the price of a 4-year membership at the local gym are not the same concept (they only share the same extension at the actual world), then the lack of entailment in () is derived. The first step of our proposal is to generalize this solution to the domain of events.

4.2 Analysis of Definite DNs with Event-Concepts

We assume that the nominal predicate of a DN denotes a property P of events, whereas a definite DN denotes the (uniquely salient) event e that satisfies P in w (29). The intension of e (an event concept) is a function from possible worlds w' to events. The ConP reading of a simple sentence, such as *Caius knew of the death of Caesar* is derived in (31), which employs [15]’s entry for individual-concept selecting *know*, which we extend to event-concepts in (30). In (28), the denotation of *know* combines with the event-argument (type E) via intensionalized functional application [7]:

- (29) $\llbracket \text{the DOC} \rrbracket^w = \iota e: e \text{ is a death of Caesar in } w$
- (30) $\llbracket \text{know of} \rrbracket^w = \lambda f_{\langle s, E \rangle} . \lambda x_e . \forall w' \in \text{Dox}_x [f(w') = f(w)]$
- (31) $\llbracket \text{know of} \rrbracket^w (\lambda w' \llbracket \text{the DOC} \rrbracket^{w'}) (\text{Caius})$
 $= \forall w' \in \text{Dox}_C (\lambda w' \llbracket \text{the DOC} \rrbracket^{w'} (w') = \lambda w' \llbracket \text{the DOC} \rrbracket^{w'} (w))$
 $= \forall w' \in \text{Dox}_C (\llbracket \text{the DOC} \rrbracket^{w'} = \llbracket \text{the DOC} \rrbracket^w)$
 $= \forall w' \in \text{Dox}_C [\iota e: e \text{ is a death of } C. \text{ in } w' = \iota e: e \text{ is a death of } C. \text{ in } w]$

The formula in (31) is true in the world of evaluation w iff in each one of Caius’ belief worlds w' (at w), the event of Caesar’s death in w is the event of his death in w'. Even if the death of Caesar and the murder of Caesar are the same event in w—say, e₄₀—(32a) is true iff (the counterpart of) e₄₀ is a dying of Caesar in all w'—not necessarily a murdering of Caesar. Thus, the entailment below does not go through:

- (32) a. Caius knew of the death of Caesar.
- b. The death of Caesar is the murder of Caesar
- $\not\Rightarrow$ Caius knew of the murder of Caesar.

Thus, event concepts allow us to derive the lack of substitution of equivalents for attitude predicates, such as *know of*, without invoking QR.⁷

We also derive the factivity of ConPs naturally: (31) entails that the death of Caesar happens in the actual world. This is a consequence of the analysis modeled after CQs: it captures both the opacity and factivity. As we saw above, when such verbs select propositions they are not necessarily factive.⁸

⁷ An event-concept analysis could also be given for verbs of communication; we won’t do this here, however, for reasons of space.

⁸ We are aware, however, that the derived truth conditions feel a little too strong for ConPs. For instance, it seems in some cases that all one needs to know if one knows of the death of Caesar is to know that such an event occurred, not that one knows which event is a death of Caesar. While we ourselves share this reservation, we think rejecting a concept approach as above might be too hasty, since similar issues arise for a concept-based analysis of CQs. After all, for one to know what the capital of Italy is, it would be sufficient, in most cases, to know its name, e.g. one may know the capital of Italy in the sense that they know its name is Rome, without necessarily being able to recognize the capital of Italy in any other way. Since events, do not bear names, it is harder to imagine under which mean of presentations these events are identified. We leave this issue and the issue of the way events are located and identified across worlds for future research.

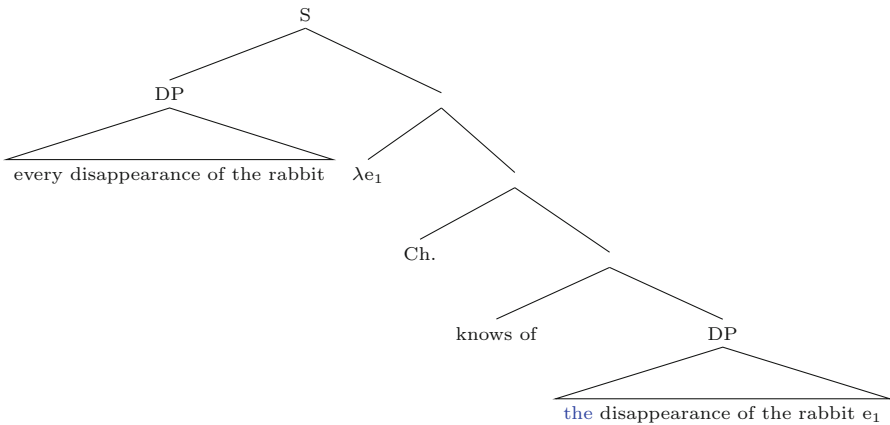
4.3 Solution to the Problem of Quantified DNs

The above approach, however, will not address the quantified ConPs which show the puzzling behavior of being wide-scoping in terms of their quantificational force but are nonetheless opaque in terms of their descriptive (or functional) content. In our analysis, quantified DNs are not suitable arguments for the embedding verb, since they do not denote (intensions) of events. Thus, they must undergo QR. However, if we allow QR, the event descriptor (i.e., the nominal predicate inside the quantified DN; in this case *disappearance of the rabbit*) ends up being evaluated at the actual world—just as in Zucchi’s analysis. Worse than that the embedded proposition ends up being the tautological and meaningless equivalence “ $e = e$ ”:

- (33) a. [Every disappearance of the rabbit] λe . Charlie knows e
- b. $\forall e$ [e is a rabbit-disappearance at w_0 (= rabbit-jumping at w_0)] \rightarrow
 $\forall w' \in \text{Dox}_C(\lambda w' \llbracket e \rrbracket^{w'}(w') = \lambda w' \llbracket e \rrbracket^{w'}(w))$
 $= \forall e$ [e is a rabbit-disappearance at w_0 (= rabbit-jumping at w_0)]
 $\rightarrow \forall w' \in \text{Dox}_C(e = e)$

We show that the problem disappears if we extend [4,5]’s analysis of CQs with quantified DPs to CPs with quantified DNs. Her analysis assumes the copy-theory of movement and a mechanism that converts the lower copy to a definite description containing a bound variable [3]. Thus, instead of (33a), we have (34).

- (34) *Copy Theoretic Structure*
 Charlie knows of every disappearance of the rabbit.



In Fox’s system the value of a descriptive trace is provided by the variable assignment function, but with the presupposition that this value is in the extension of the NP-predicate at the world of evaluation. However, in order to avoid false sentences coming out undefined when the attitude holder has false beliefs regarding the extension of the NP-pred. at w , Frana replaces the definite determiner with the maximality operator in (35). This operator is a standard

maximality operator (c.f. [16], for example), with an additional clause designed to deal with empty sets. According to the first clause, when the_{max} applies to the extension of a predicate in a given world (a set), it returns the maximal element of that set (an individual when the set is a singleton set, or an individual sum when the set consists of more than one individual). The second clause says that if the set picked out by the predicate is empty, then $the_{max}(A)$ returns the null individual.⁹ Descriptive traces are then interpreted as in (36); here $Pred$ is a predicate of individuals or events, and x ranges over individuals or events.

- (35) For any set A (i.e. the extension of a predicate NP in w)
 if $A \neq \emptyset$ then,
 $the_{max}(A) = \iota x [x \in A \wedge \forall x' \in A [x' \leq x]]$;
 if $A = \emptyset$ then,
 $the_{max}(A) = *$ (the null individual, which is not in any Natural Language denotation, c.f. [1])
- (36) $\llbracket \mathbf{the}_{max} Pred x_i \rrbracket^{w,g} = g(i)$ if $\llbracket Pred \rrbracket^{w,g}(g(i))$,
 otherwise $\llbracket \mathbf{the}_{max} Pred x_i \rrbracket^{w,g} = *$
- (37) For any constituent α and variable assignment g ,
 $\llbracket \lambda_{iT} \rrbracket^{w,g} = \lambda x_T. \llbracket \alpha \rrbracket^{w,g[i/x]}$

The structure (34) is interpreted as in (38), where DIS abbreviates the predicate of events *disappearance of the rabbit*.

- (38) $\llbracket (34) \rrbracket^{w,g} =$
- $$\forall e (DIS(e) \text{ in } w \rightarrow \forall w' \in \text{Dox}_C(w) [\lambda w_2. \llbracket \mathbf{the}_{max} DIS e_1 \rrbracket^{w_2, g[1/e]}(w') \\ = (\lambda w_3. \llbracket \mathbf{the}_{max} DIS e_1 \rrbracket^{w_3, g[1/e]}(w))] =$$
- $$\forall e (DIS(e) \text{ in } w \rightarrow \forall w' \in \text{Dox}_C(w) [\llbracket \mathbf{the}_{max} DIS e_1 \rrbracket^{w', g[1/e]} = \\ \llbracket \mathbf{the}_{max} DIS e_1 \rrbracket^{w, g[1/e]}]]$$

Even if every rabbit-disappearing event is a rabbit-jumping event at w , according to the formula above, (24a) is true iff for every actual disappearing event e , (the counterpart) of e is also a disappearing event—not necessarily a jumping event—in all of Charlie’s belief worlds w' . Thus, the conclusion in (24c) does not follow from the premises.

To conclude, building on existing analyses of CQs, we provide an analysis of ConPs in which definite DNs denote (intensions of) events, thus solving the problem of co-extensional events, without assuming that DNs denote propositions. The analysis is also extended to cover ConPs with quantified DNs ? which do not denote event concepts—within Frana’s copy-theoretic account. In such cases, it is the copy-trace left by the QR-ed DN that supplies the event-concept argument to the verb.

⁹ Null or absurd individuals have been employed in the choice function literature to resolve the empty NP-restrictor problem.

5 Summary and Concluding Issues

We have presented an account of DNs with ConP-readings that builds on existing analyses of individual-denoting DPs with concealed question-readings. The gist of our proposal is to extend an individual-concept account of CQs to ConPs, employing event concepts. This allowed us to resolve Zucchi's dilemma: ConPs' quantificational force scopes wide but its descriptive content is opaque. We used Frana's copy-theoretic analysis of CQs to capture this behavior. Another welcome benefit of the analysis is that it captures the factivity of ConPs with verbs that are otherwise not factive.

There is a further respect in which ConPs resemble CQs, and thus constitutes further evidence for our CQ-like analysis of ConPs. One issue that has arisen in the CQ literature is distinguishing between DPs complements of question-taking verbs that are simply so vague in admitting readings that are compatible with questions and those DP complements that deliver genuine CQ-meanings. One way to distinguish CQs from other question-like readings is their lack of vagueness.

- (39) Context (Part 1): The panel picked Mr. P as the winner of the writing contest.
I was informed of the winner of the contest. = CQ
- (40) Context (Part 2): Roger learned later that Mr. P plagiarized his work.
 a. #I was informed of the winner of the contest (. . . i.e. he plagiarized.)
 b. I was informed about the winner of the contest.

While any kind of contextually salient property/fact is something you can be informed *about*, *of*+CQ is not vague in the same way: it only delivers a meaning that resembles the identity question *who the winner was*. [14] has shown that *about*-phrases attached to sentence-embedding verbs are quite vague about the role of the DP.

Now, we predict that ConPs are likewise *not* vague. To test this we are going to compare canonical ConPs introduced by *of* with DNs introduced by *about*, as in (41):

- (41) Context: I was told that Mary resigned and that it was because she took another job. Then John said the truth of the matter is that she resigned because she stole cash.
 a. Did John inform you about Mary's resignation?
 b. #Did John inform you of Mary's resignation?

Inform of is odd in this context: it simply cannot convey any salient propositional/property available in the context holds of Mary's resignation, unlike *inform about*.¹⁰ We can conclude, then, that ConP interpretations are not just among a set of possible interpretations owing to vagueness. The event-concept

¹⁰ Zucchi argues that *surprise* allows contextually-supplied properties in place of OCCUR.

analysis, coupled with Frana's copy-theoretic implementation for quantified DNs, accounts for this.

References

1. Barwise, J.: Scenes and other situations. *J. Philos.* **78**, 369–397 (1981)
2. Fodor, J.D.: The linguistic description of opaque contexts. Ph.D. thesis, MIT (1970)
3. Fox, D.: Antecedent-contained deletion and the copy theory of movement. *Linguist. Inq.* **32**(1), 63–96 (2002)
4. Frana, I.: Quantified concealed questions. *Nat. Lang. Seman.* **21**(2), 179–218 (2013)
5. Frana, I.: *Concealed Questions*. Oxford University Press, Oxford (2017)
6. Heim, I.: Concealed questions. In: Bauerle, R., Egli, U., von Stechow, A. (eds.) *Semantics from Different Points of View*, vol. 6, pp. 51–60. Springer, Berlin (1979). https://doi.org/10.1007/978-3-642-67458-7_5
7. Heim, I., Kratzer, A.: *Semantics in Generative Grammar*. Blackwell, Malden (1998)
8. Higginbotham, J.: The logic of perceptual reports: an extensional alternative to situation semantics. *J. Philos.* **80**(2), 100–127 (1983)
9. Karttunen, L.: Presupposition of compound sentences. *Linguist. Inq.* **4**(3), 169–193 (1973)
10. Nathan, L.: On the interpretation of concealed questions. Ph.D. thesis, MIT (2006)
11. Pietroski, P.: On explaining that. *J. Philos.* **97**(12), 655–662 (2000)
12. Portner, P.: *Situation theory and the semantics of propositional expressions*. Ph.D. thesis, University of Massachusetts-Amherst (1992)
13. Ramsey, F.P.: Facts and propositions. In: *Proceedings of the Aristotelian Society*, Suppl. vol. VII (1927)
14. Rawlins, K.: About *about*. In: Snider, T. (ed.) *Proceedings of Semantics and Linguistic Theory XXIII*, pp. 336–357. CLC Publications, Ithaca (2013)
15. Romero, M.: Concealed questions and specificational subjects. *Linguist. Philos.* **28**(6), 687–737 (2005)
16. Rullmann, H.: *Maximality in the semantics of Wh-constructions*. Ph.D. thesis, University of Massachusetts-Amherst (1995)
17. von Stechow, A.: *Some Remarks on choice functions and LF movement*. University of Tübingen, MS (1996)
18. Szabó, Z.G.: Specific, yet opaque. In: Aloni, M., Bastiaanse, H., de Jager, T., Schulz, K. (eds.) *Logic, Language and Meaning*. LNCS (LNAI), vol. 6042, pp. 32–41. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14287-1_4
19. Vendler, Z.: *Linguistics in philosophy*. Cornell University Press, Ithaca (1967)
20. Zucchi, A.: *The Language of Propositions and Events*. Kluwer, Dordrecht (1993)



Denials and Negative Emotions: A Unified Analysis of the Cantonese Expressive *Gwai2*

Grégoire Winterstein^(✉), Regine Lai, and Zoe Pei-sui Luk

Department of Linguistics and Modern Language Studies,
The Education University of Hong Kong, Tai Po, Hong Kong
{gregoire,ryklai,psluk}@eduhk.hk

Abstract. This paper deals with the Cantonese morpheme *gwai2* (鬼, lit. ‘ghost’) which, besides its spooky nominal use, also conveys expressive meaning when modifying a wide range of expressions: adjectives, verbs, wh-pronouns, etc. We begin by reviewing the empirical domain of *gwai2* and different claims of the literature concerning its dual nature as an intensifier and a mixed-expressive conveying at-issue negation. We discuss both of these claims, showing that *gwai2* cannot be treated as an intensifier in the usual sense, and that it does not contribute a truth-conditional negation, but rather a form of denial. We then propose a unified analysis of the morpheme based on the assumption that it indicates a negative attitude of the speaker towards its argument, notably by showing how to derive denials from this negative attitude.

Keywords: Cantonese · Expressive content · Denial · Intensification

1 Introduction

This paper deals with the semantic contribution of the Cantonese morpheme *gwai2* (鬼). Literally, the term means “ghost” and can be used as a noun with this meaning. Besides that spooky nominal use, *gwai2* also modifies a wide range of expressions (adjectives, verbs, wh-pronouns, etc.) In those uses, which are the ones of interest to us, *gwai2* conveys an expressive meaning with hazy contours: it apparently either intensifies the expression it modifies or negates it.

Gwai2 has already attracted attention in the literature. Lee and Chin (2007) provide a detailed description of the syntactic distribution of *gwai2* and the meaning conveyed by *gwai2* in the various positions it can occupy. Beltrama and Lee (2015) show that *gwai2* conveys expressive meaning à la Potts (2005,

The authors would like to thank Andy Chin, Shin Kataoka and David Li, as well as the audience at LENLS 14 and the third Okinawan Semantics Workshop organized under the auspices of the Asian Semantics Society, especially C. Davis, Y. Hara, E. McCready, M. Yoshitaka Erlewine and L. Rieser for their comments and discussions.

2007), and analyze one of its uses as a form of mixed expressive (McCready 2010).

These works agree on distinguishing two main uses for *gwai2*: one in which *gwai2* is usually described as an *intensifier*, and another in which it is a *negator* (in addition to its nominal, literal use which we ignore here). While these two uses appear to be in near complementary distribution, they have in common the fact that *gwai2* always conveys a form of heightened emotion of the speaker. This expressive component is not perceived as particularly rude (e.g. *gwai2* is heard on public radio), and is comparable to the English expressive (*like*) *hell*.

Besides *gwai2*, other elements in Cantonese have similar distributions and contributions. Notable elements (because of their frequency) are *lan2* (關/撚, lit. ‘*dick/penis*’) and *gau1* (閘/鳩, lit. ‘*cock/penis*’), which share most of the characteristics of *gwai2*, but in a much ruder and marked register. The letters *X* and *Q* are also frequently substituted for *gwai2/lan2* (Matthews and Yip 2011).

Our goal in this paper is to provide a unified analysis of *gwai2*, which so far has not been attempted. We begin by reviewing both of the readings commonly attributed to *gwai2*. In Sect. 2, we discuss the “negator” use of *gwai2* and show that it does not convey a simple at-issue negation, as previously claimed, but is rather an instance of denial. Section 3 focuses on the so-called “intensifier” reading. There we contend that intensification is not the core contribution of *gwai2*, which we analyze as a pure expressive. Section 4 brings the observations of the preceding sections together and we propose that the two readings are manifestations of the expressive component of *gwai2*. We argue that this meaning conventionally encodes a form of negative affect. Depending on the nature of the element modified by *gwai2*, this affect is interpreted differently. We conclude by comparing *gwai2* to expressives in other languages.

2 *Gwai2* as a Marker of Denial

In this section we begin by reviewing the distribution of *gwai2* in what is customarily called its “negator” reading (Sect. 2.1). There the main effect of *gwai2* seems to be to negate the content of its prejacent (i.e. its host sentence). A basic contrast is shown in (1)–(2).

- | | | | |
|-----|---|-----|---|
| (1) | keoi5 sik1.
s/he know
S/he knows. | (2) | keoi5 gwai2 sik1.
s/he GWAI know
Like hell s/he knows.
= s/he doesn't know |
|-----|---|-----|---|

The translation of (2) reflects the hypothesis we defend in Sect. 2.2: while it is correct that *gwai2* conveys a form of negation there, it is best seen as a denial of a previous statement rather than a standard descriptive negation.

2.1 Empirical Domain

The negator reading of *gwai2* is typically observed when *gwai2* is prefixed to a verb phrase, or suffixed to a bare predicate (verbal or adjectival) in the

- an expressive content indicating the heightened emotional state of the speaker (similar to the one described by Potts 2005)
- a standard, descriptive at-issue negation.

We agree with their claims about the expressive content, and the arguments they use (i.e. it meets the usual tests of scopelessness, impossibility to be bound, behavior with denials and general ineffability). However, we argue that the negation conveyed by *gwai2* is a form of non at-issue dialogical denial, a hypothesis already evoked by Lee and Chin 2007.

First, the negation contribution by *gwai2* is not affected by usual truth-conditional affecting environments, e.g. questions, antecedents of conditionals, or modal operators. In those environments, the standard marker of descriptive negation, *m4*, is felicitous. If *gwai2* contributed a standard negation, we would expect these environments to license it, only adding its expressive component in the picture. Instead, it can be shown that *gwai2* cannot be embedded in any of these environments: (9).

- (9) a. *keoi5 hai6m4hai6 gwai2sik1 aa3?
 s/he is-not-is GWAI-know SFP
 (int.) *Doesn't he (goddamn) know ?*
- b. *jyu4gwo2 keoi5 gwai2 sik1, nei5 zau6 jiu3 gong2 bei2 keoi5 zi1.
 if s/he GWAI know, you then need tell give s/he know
 (int.) *If he doesn't goddamn know, you need to tell him/her.*
- c. *waak6ze2 keoi5 gwai2 sik1.
 maybe s/he GWAI know
 (int.) *Maybe s/he doesn't goddamn know.*

In addition, an utterance of the form *gwai2_{neg}p* is only possible if *p* has been previously evoked in the discourse. Thus, in (10), even though there is a general assumption that the coffee served at a coffee place should be hot, negator-*gwai2* is not licensed, whereas the standard negation *m4* is:

- (10) [At a coffee place, the speaker just picked up his cup.]
 a. ni1 buil gaa3fel (m4/# gwai2) jit6 ge2!
 DEM CL coffee NOT/GWAI hot SFP
 This coffee's not hot.

On the other hand, if the content has been previously conveyed (either in an at-issue or not way), it can be targeted by *gwai2*:³

³ Note that the content target can itself involve a negation, e.g. (i) (suggested by a reviewer).

- (i) A: Siu-ming, who is not a linguist, could not understand the importance of his own dialect.
 B: keoi5 hai6 gwai2 m4 hai6 linguist.
 he is GWAI NEG is linguist
 Like hell he's not a linguist.

- (11) A: Siu-ming, the linguist, came to the party.
 B: keoi5 hai6 gwai2 linguist.
 he is GWAI linguist
 Like hell he's a linguist.

This behavior of negator-*gwai2* seems to place it in the category of “bullshit” operators (Spender and Maier 2009) (hence our choice of translation by *like hell*). However, there appears to be a restriction on the ability of *gwai2* to target some conversational implicatures. While *gwai2* can deny quantity implicatures, it has more difficulties targeting other conversationally conveyed content, notably manner implicatures (12).

- (12) A: keoi5 dou6zi3 tai1jan4 sei2mong4 wo5.
 s/he caused other death EVI-SFP
 I heard s/he caused the death of someone.
 B: #hai6 gwai2, keoi5 hai6 mau4saat3 aa3
 is GWAI s/he COP murder SFP
 (int.) *Like hell s/he did, s/he murdered someone.*

We assume that these effects are related to the larger question of the accessibility of these conversational implicatures and other conversational features, and not inherent to *gwai2*, and we will therefore not deal with these facts here.

Gwai2 can also come as an answer to a question, biased or not. In (13), A's question can be neutral (marked with the SFP *aa4*), biased towards a positive answer (SFP *ho2*) or a negative one (SFP *me1*, see Hara 2014 for an overview of biased questions in Cantonese), and allow B as an answer.

- (13) A: keoi5 jau5 cin4 me1/aa3 ho2/aa4?
 he have money SFP
 He has money, (does he/doesn't he)?
 B: jau5 gwai2.
 have GWAI
 like hell he does (= he has no money)

Finally, *gwai2* does not interact with the interrogative Sentence Final Particle (SFP) *me1* as regular negation does (Lee and Chin 2007). The SFP *me1* turns declarative sentences into interrogatives, conveying in addition the low belief of the speaker in a positive answer (Kwok 1984; Matthews and Yip 2011; Hara 2014). When *me1* is used with a standard negation of a content *p*, it thus conveys the belief of the speaker in *p* (14), but the opposite happens with (15). There *gwai2* only seems to convey an emotional content and the speaker is understood to believe $\neg p$.

- | | |
|--|--|
| <p>(14) keoi5 m4-zi1 me1?
 s/he NEG-know SFP
 <i>He knows, doesn't he?</i></p> | <p>(15) keoi5 gwai2-zi1 me1?
 s/he GWAI-know SFP
 <i>He wouldn't know, would he?</i></p> |
|--|--|

The contrast in (14)–(15) is accounted for by the fact that both denial-*gwai2* and the SFP *me1* require a previous evocation of their prejacent in order to

indicate that the speaker does not believe in it. Therefore, instead of involving a case of double negation, (15) is rather an example of harmony between the constraints of *gwai2* and *me1*.

Taking stock, in its denial uses:

- *gwai2* takes scope over a whole utterance denoting a state
- the content of the prejacent must be *echoic*: it must have been previously evoked in the discourse (conveyed by a speaker, or evoked via a previous question)

On a final note, the echoic property of the denial cases is reminiscent of what Carston (1996) considers to be the central feature of (English) metalinguistic negation. The parallel between the denial conveyed by *gwai2* and metalinguistic negation appears sensible enough, but they differ in several aspects. First, the behavior of *gwai2* as an answer to questions differs from that metalinguistic negation, and second, *gwai2* is unable to target aspects like pronunciation which are accessible to metalinguistic negation.

3 *Gwai2* as a Pure Expressive

In its non-negator uses *gwai2* has been described as an *intensifier*. In Sect. 3.1 we illustrate the different environments in which *gwai2* can appear with this reading. Then we argue that *gwai2* does not necessarily convey a form of intensification in those environments (Sect. 3.2). Instead, we show it is better treated as a pure expressive.

3.1 Empirical Domain

Lee and Chin (2007) describe *gwai2* as an intensifier when it appears between an adverb and an adjective (16), is used in verbal compounds (between the verb stem and affixes) (17).

- (16) go3 pi1sa4 hou2 gwai2 hou2sik6!
 CL pizza very GWAI delicious
The pizza is damn delicious.
- (17) keoi5 sik6 gwai2 zo2 ngo5 di1 tong2
 he eat GWAI PFV my CL candy
He fucking ate my candy.

That interpretation of *gwai2* is also triggered when *gwai2* is infixated in interrogative words (18)–(19), quantifier phrases (20)–(21) and in some adjectives (22) (compare with (4) above, and refer to Sect. 4.1 for more details about the cases of infixation).

- (18) bin1-gwai2-go3 lai2-zo2 aa3?
 who-GWAI come-PFV SFP
Who the hell came?

- (19) dim2-gwai2-joeng2 zou6 ga3?
How-GWAI do SFP
How the hell do you do it?
- (20) mou5-gwai2-jan4 lei4
nobody-GWAI come
Not a soul came.
- (21) ni1 gaan1 fong2 zeoi3-gwai2-do1 ho2ji3 co5 sei3sap6 go3 jan4.
this CL room most-GWAI can sit forty CL people
Forty people at (damn) most can sit in this room.
- (22) hou2 maa4-gwai2-faan4!
very GWAI-annoying
[This/He] is damn annoying.

3.2 *Gwai2* as a Pure Expressive

A standard view on intensifiers is that they are “linguistic devices that boost the meaning of a property upwards from an assumed norm” (Quirk et al. 1985), or that they require a scalar dimension they can modulate by indicating some higher-than-usual degree on the scale (Eckardt 2009).

Some of the expressions above do not readily involve a scalar dimension which could be manipulated by *gwai2*, and do not trigger a denial reading either. A case in point is the infixation in interrogative words (18)–(19). There, it is not clear which degree should be intensified. Instead, *gwai2*’s main contribution is the indication of the emotional agitation of the speaker. Similarly, *gwai2* can also modify non-gradable elements. In (23), the use of *gwai2* is again limited to a display of emotion by the speaker, but does not convey (for example) an indication of great age of A-Wai.

- (23) A3-Wai5 sing4-gwai2-zo2-nin4 laa3.
A-Wai of-age-GWAI-PFV SFP
A-Wai is goddamn of age.

By itself, the fact that *gwai2* associates with non-gradable predicates is not proof that it is not an intensifier. English *totally* and *very* also have this property; when they are used with a non-gradable item, they are able to operate on some non-lexical, pragmatically obtained scale. In their non-gradable uses, these elements are usually described as slack regulators (cf. *very*: Bylinina and Sudo 2015) or as indicating a form of strong commitment about an open issue (cf. *totally*: Beltrama 2016). These analyses however do not apply to *gwai2*. For one, *gwai2* in (23) does not convey a sense that its argument is a clear prototypical case of the property in question (as in *very first time*), or some notion of precisification (as in *very center*) or any comparable value. *Gwai2* also does not seem amenable to an analysis that would treat it like *totally*: it cannot come as a reply to a question, nor to confirm the prior assertion of a subjective property (as described by Beltrama 2016).

Therefore, we will consider that in the uses discussed in these sections, *gwai2* is a pure expressive which conveys the heightened emotion of the speaker. This emotion is specific to the argument of *gwai2*, which we illustrate in (24).

- (24) a. bin1-gwai2-go3 jam2-zo2 ngo5 zi1 be1zau2 aa3?
 who-GWAI drink-PFV my CL beer SFP
Who the hell drank my beer?
- b. bin1go3 jam2-gwai2-zo2 ngo5 zi1 be1zau2 aa3?
 who drink-GWAI-PFV my CL beer SFP
Who fucking drank my beer?

In (24-a) the speaker is understood to be angry at the person who drank their beer, while in (24-b) the emotion of the speaker is related to the fact that his beer was drunk, i.e. those are not cases of isolated conventional implicatures (in the terms of Potts 2005).

To summarize, we have argued that what ties all the uses of *gwai2* considered here is its expressive component, rather than a form of intensification. The next section investigates the content of this component in more detail.

4 Unifying *Gwai2*

In this section we propose a unified analysis for *gwai2*. In a nutshell, we argue that the interpretation of *gwai2* is due to the nature of its expressive content. Our analysis will scavenge and adapt bits and pieces from other approaches to expressives and their affective (or emotive) orientation.

In terms of semantic contribution, our proposal is not very different from the original proposition by Potts (2005, p. 167) for expressives like the English *damn/fucking*, i.e. we assume a representation as in (25).

$$(25) \quad \llbracket gwai2 \rrbracket = \lambda X. \mathbf{negAffect}(\cap' X) : \langle \langle \tau^a, t^a \rangle, t^c \rangle$$

The \cap' operation shifts the type of the argument to its ideal, i.e. an element of the appropriate type for the evaluation conveyed by the predicate **negAffect**. This operation is similar to the one used by Potts, but needs to be slightly more versatile. Minimally it should allow the shift from the denotatum of wh-pronouns, and also recognize echoic statements as a type in its own right (cf. below). Since our goal lies more in the constraint encoded by *gwai2* than its compositional properties, we will leave those details aside (for the modification of echoic propositions, see for example the propositions of McCready 2008 about *man* and the modification of contextually salient propositions).

The description in (25) relies on the predicate **negAffect**. This predicate is meant to indicate a (default) negative attitude of the speaker towards the argument of *gwai2*. We argue for this analysis in Sect. 4.1. This is one departure from the usual view on expressives, which are often seen as underspecified for the emotion they encode, and paraphrased as “indicating the speaker’s heightened emotional state” (Potts 2005; Constant et al. 2009; McCready 2012). Assuming that *gwai2* lexically encodes a negative attitude accounts for a number of

its properties, notably its denial reading. We show how to go from a negative attitude to denial in Sect. 4.2. It however opens one issue: in some cases, that negative affect reading of *gwai2* is absent, and *gwai2* is rather understood as positive. We deal with those cases in Sect. 4.3, where we use the approach of McCready (2012) based on default logic to explain them.

4.1 *Gwai2* Encodes Negative Affect

When dealing with the affective orientation of expressives like English *fucking*, McCready (2012) argues that this orientation is underspecified: depending on contextual elements, it can be either positive or negative. For the case of *gwai2*, we will argue that this orientation is lexically biased towards the negative. This is based on the observation that in several contexts, *gwai2* can only be used to indicate a degree of negative emotion. Infixation in wh-words is a case in point: (26) can be uttered by a speaker at their birthday party only after opening an unpleasant/joke gift, but not to show genuine delight.

- (26) ni1 joeng6 bin1-gwai2-go3 sung3 gaa3?
 DEM CL who-GWAI offer SFP
Who the hell got me this one?

To further show the affinity of *gwai2* with negative affect, we investigated the effects of the infixation of *gwai2* in Cantonese disyllabic adjectives. As mentioned above, in such cases *gwai2* can either convey the denial of its prejacent, or the more simple pure expressive content. Lee and Chin (2007) observe that there seems to be a correlation between the affect associated with the adjective and the interpretation of infix-*gwai2*. Adjectives with positive connotations tend to get denied, whereas negatives ones do not. The pair in (27)-(28) illustrates this: the positive sounding adjective *useful* with infixed *gwai2* is interpreted as the denial of a previous statement, contrary to infixation in *useless*.

- | | | | |
|------|---|------|---|
| (27) | jau5-gwai2-jung6!
useful-GWAI
<i>Like hell it's useful.</i> | (28) | mou5-gwai2-jung6!
useless-GWAI
<i>This is damn useless!</i> |
|------|---|------|---|

We asked 11 native Cantonese speakers to annotate a list of 2047 disyllabic adjectives (extracted from a MOR grammar for CHAT Data obtained at <http://talkbank.org/morgrams/>). One group was instructed to indicate the effect of the infixation of *gwai2* in the adjective as either: a denial, an intensification (used as way to refer to non-denial cases), both, or an ungrammatical result. Another group had to indicate whether they thought the adjectives have a positive, negative, or neutral connotation. Each adjective was annotated by two annotators in each task. 407 received concordant annotations, the rest is ignored here, but our theoretical solution offers a way to account for the discrepancies in annotation on those ignored items. The results are summarized in Table 1.

The results show that where annotators agree on the effect of *gwai2*, the non-denial readings of *gwai2* mostly involve adjectives with no clear positive polarity i.e. neutral and negative ones. The exceptions in the table are:

Table 1. Correlation of subjective adjective connotation and effect of *gwai2*-infixation

	Neg. connotation	Pos. connotation	Neut. connotation	Tot.
Intensification	116 (81.7%)	4 (2.8%)	22 (15.5%)	142
Denial	12 (5.2%)	164 (71.6%)	53 (23.1%)	229
Both	10 (27.8%)	17 (47.2%)	9 (25.0%)	36

- Adjectives like *waan4koeng4* (‘tenacious’), *daai6lik6* (‘strong’), *haak3hei3* (‘polite’), *hou2je5* (‘excellent’) that have positive connotation and get intensified rather than denied. However the effect of *gwai2* there is not perceived as positive, rather its interpretation is that the property holds at a too high degree (e.g. “*too polite*”).
- Adjectives like *ciu4sei5* (‘haggard, gaunt’), *hung1heoi1* (‘hollow, void’) which are negative and do not get intensified. Those are not very colloquial adjectives, and the annotators (along with these authors) recognize their intuitions are vague about them.

Beyond infixation, non-denial cases also normally involve a negative attitude of the speaker. This is of course the case when the argument of *gwai2* is negatively connotated. This is also in a case like (29), where, even though pallor can be seen as a positive attribute in Chinese culture, it can also be a sign of poor health and this reading appears more prevalent in combination with *gwai2*.

- (29) keoi5 hou2 gwai2 baak6!
 s/he very GWAI white
S/he’s damn white.

To summarize: when *gwai2* modifies elements that have no intrinsic connotation (such as wh-pronouns or neutral adjectives), it necessarily conveys a negative attitude of the speaker. It does so too when its argument is negative. In the case of adjective infixation, if its host is positively connotated, the most obvious reading of *gwai2* is one of denial. In the next section, we argue such denials can be derived from the negative attitude of the speaker, thus supporting our hypothesis that by default *gwai2* encodes such a negative attitude.

4.2 From Negative Attitude to Denial

The general picture we drew is that *gwai2* conveys a denial when it scopes over a whole utterance (as an affix on the main predicate of a sentence) and when that utterance is echoic, i.e. has been evoked previously in the discourse by an agent different from the current speaker. The denial reading also conveys an expressive component, akin to the one conveyed in the non-denial cases. It thus seems reasonable to try and see whether the perceived intensification and negation cases can be derived from this expressive component.

The cases of denial discussed in Sect. 2.2 have one thing in common: they all involve a proposition *C* such that a speaker *S*₁, distinct from the *gwai2*-speaker

S_2 , has a non-null degree of belief in C , which we will write as $P_{S_1}(C) > 0$ (we equate degrees of belief with probabilities, in typical Bayesian fashion, see e.g. Jeffrey 2004). We analyze the echoic property by considering that S_1 has made a conversational move that involves *grounding* the possibility that C is true, i.e. that $P(C) > 0$ (see Clark 1996 and Ginzburg 2012 a.o. for elaborate considerations). Such a move is trivial in the case of assertions and with any content conventionally conveyed by S_1 (though conveying non at-issue content does not usually involve a call on the addressee to ground the content in question). In those cases the belief of S_1 is usually much higher than 0, but not necessarily equal to 1. The case of questions also involves such a move. If S_1 asks whether C is true, they are pushing $C?$ on the stack of Questions Under Discussion (QUD) and ask the addressee to do the same (Ginzburg 2012). Doing so entails recognizing that both C and $\neg C$ are possible. So both questions and assertions have in common that a content C has been uttered in a way that calls on the addressee to recognize that $P(C) > 0$ (at least before S_2 makes their move).

In the denial cases *gwai2* therefore takes as its argument a content like $C' = \text{GROUND}(S_1, P(C) > 0)$, which we mean to denote a move made by S_1 to add $P(C) > 0$ to the Common Ground.

What does it mean to have a negative attitude towards such a content? By itself, a move to ground content calls for two possible actions: acceptance/grounding by the hearer or a refusal to do so. Under this assumption, a negative attitude of the speaker is best interpreted as a signal for the second option: the speaker (emphatically) refuses C' , i.e. to accept C as part of the common ground.

The case of the different types of questions introduced in (13) helps to illustrate this. In those examples speaker A is our S_1 and B is S_2 . S_1 is asking whether C is the case, where $C =$ “he has money”.

Depending on which question particle S_1 uses, their beliefs in C will be of various strengths, but always allowing room for C to be true or false. More precisely:

- The unbiased question particle *aa4* indicates comparable beliefs in both options: $P_{S_1}(C) \sim P_{S_1}(\neg C) \sim 0.5$
- The particle *me1* is biased towards a negative answer: $P_{S_1}(\neg C) > P_{S_1}(C) > 0$
- The particle *ho2* is biased towards a positive answer: $P_{S_1}(C) > P_{S_1}(\neg C) > 0$

Thus irrespective of the bias of S_1 , their questioning move always involves adding the possibility of C to the Common Ground as part of the accepting the question. Of course S_2 might have beliefs about $P(C)$ and will convey it by answering the question, but before doing it, S_2 needs to ground the question and what comes with it. This is what S_2 refuses to do when using *gwai2*. Note that S_2 cannot deny $\neg C$ because $\neg C$ would not be echoic in that example i.e. has no linguistic reflex in S_1 's utterance.

If *gwai2* can deny the commitments conveyed by an assertion or a question, one might wonder about other illocutionary moves. So far we only considered adding a non null belief to the common ground, but it seems that *gwai2* may

also target other contents. While *gwai2* cannot be used to refuse a direct order (30), it is felicitous when refusing an invitation or suggestion (31).

- (30) a. zap1 fong2!
 clean room
 Clean your room!
- b. #gwai2 zap1.
 Gwai clean
 (int.) *Like hell I will*
- (31) a. jat1cai4 waan2 laa1
 together play SFP
 Let's play together
- b. gwai2 tung4 nei5 waan2
 Gwai with you play
 Like hell I'll play with you.

Both moves above involve a commitment to an *outcome* (Ginzburg and Sag 2000), i.e. the future realization of a propositional content. The contrast in (30)–(31) suggests that the content of the outcome is only accessible in the case of invitations. One way to explain the contrast is to consider that beyond the commitment to an outcome, invitations also involve a call on addressee to answer the invitation, whereas orders do not convey this. To capture that difference between invitations and direct imperatives, one could use a Dialogue Game Board approach in the vein of that of Ginzburg (2012) and predecessors. We will not pursue that line of inquiry here and leave it to further work.

Taking stock we have seen how to derive denials from the negative attitude of the speaker encoded by *gwai2*. That reading is triggered only in echoic cases. Going back to the case of infixation in disyllabic adjectives, we can explain the results in a new light. The scope of *gwai2* in the infixation cases is ambiguous: it can either (i) take scope on the predicate alone, in which case the utterance will convey its prejacent along with a negative expressive component, or (ii) convey a denial if the host of *gwai2* is echoic. Case (i) is not readily compatible with positive adjectives: there is a clash between the positive connotation of the adjective and the constraint conveyed by *gwai2*. This explains the preference to read those cases as instances of denial. When the adjective is not overtly positive, the non-denial readings are accessible to intuition, which accounts for the results.

4.3 *Gwai2* and Positive Attitudes

To finish this section we will look at cases that involve the use of *gwai2* without conveying a negative attitude of the speaker. Those are potential counterexamples to our claim that this negative attitude is the core contribution of *gwai2*. An example of that sort is given in (32).

- (32) keoi5 hou2 gwai2 leng3!
 she very GWAI pretty
She's damn pretty.

In (32), the use of *gwai2* does not necessarily convey any sort of negative attitude of the speaker regarding the prettiness in question. It therefore seems to behave there much as English *damn* or *fucking* would. Note that a denial reading is not accessible here because of the predicate-internal position occupied by *gwai2* which prevents it to scope over the whole utterance.

We argue that in a case like (32), the content of *gwai2* clashes with the connotation attributed to prettiness. A similar issue has been addressed by McCready (2012) who shows that some expressives are underspecified in terms of the emotion they convey (e.g. *fucking*, *damn*, or Japanese *kuso*). To model how the affective orientation of such expressives is determined, McCready uses a mixture of nonmonotonic inference and game-theoretical considerations on how communication proceeds. What is of interest here is his hypothesis that lexically encoded information supersedes other sources from which to infer affective information, while remaining defeasible by other information.

In a case like (32), we then assume that two indications of affect are at odds: the negative one marked by *gwai2*, and the positive indication that comes with *pretty*. We assume that in such cases, the “stronger” of the two survives, meaning that if the positive affect associated with the argument of *gwai2* is strong enough, it can override the negative bias of *gwai2*.

This predicts a number of things. First, there should be predicates that are only weakly positive, i.e. whose positive constraint “loses” against *gwai2*. A case like (33) is such a case: if uttered, it will convey a degree of scorn of the speaker towards the elegance rather than a fully positive appraisal as in (32) (see also the cases discussed at the end of Sect. 4.1). In other words, the predicate *elegant* is intensified in (33), but this is not understood as a positive thing.

- (33) keoi5 hou2 gwai2 gou1gwai3
 he very GWAI elegant
He's damn elegant.

Second, the same positive override should be observable when *gwai2* appears in other positions. This is indeed the case: see for example (34) which involves a highly positive property and a correspondingly positive attitude of the speaker conveyed by *gwai2*.

- (34) ngo5 zung3 gwai2 zo2 luk6hap6coi2
 I won GWAI ASP lottery
I fucking won the lottery.

Finally, we should find speaker variation in the interpretation of some examples, since different speakers might attribute different degrees of positivity to the same predicate. This could lead to some miscommunication problems. The annotation task mentioned previously supports this prediction: the annotators disagreed on a number of cases, and cases like (33) are not clear-cut for some speakers. Some

understand it as a negative thing, others, probably more sensitive to matters of elegance, see it as a positive statement.

5 Conclusion: Beyond *Gwai2*

We have offered an analysis of *gwai2* that rests on the idea that it encodes a (strong) default indication that the speaker is feeling negative. We argued that when *gwai2* scopes over a whole echoic utterance, that negative attitude amounts to a denial. Though we argue that this negativity is conventionally attached to *gwai2*, distinguishing it from expressives such as English *fuck*, it can be overridden when the argument of *gwai2* encodes a strong positive affect. This override only happens when *gwai2* modifies sub-sentential elements. When its argument is echoic, its scope is at the speech-act level and there is no sense in which a positive aspect could override the negativity of *gwai2*.

The particle *gwai2* is by far not unique in the world of expressives: it intensifies in the same way as other expressives like English *damn*, and it negates in the same way as other expressives such as *bullshit*. In some respect *gwai2* resembles the (equally netherworldly) expressive *like hell* which is also described as an intensifier (*It hurt like hell*) and has denying properties when used as a reply. It is however not clear whether *like hell* also matches the lexically encoded negative attitude that we argue characterizes *gwai2*.

There are elements with such inherent negative properties, but they often have a more restricted distribution than *gwai2* (e.g. slurs typically encode negative attitudes about their referent McCready 2010, 2012). One element that appears similar to *gwai2* is the family of French expressives derived from the adjective *sale* ('dirty'). The adjective can only modify nouns by indicating a negative attitude of the speaker (e.g. *sale flic* 'damn cop', *sale prof* 'damn teacher', or *sale ami* 'fucking friend' which can only have a negative reading).⁴ The derived adverb (*salement*, 'dirtily') however can modify properties in a positive way (e.g. *salement bon*, 'damn good (to eat)') much like *gwai2* can. In spite of this, *sale* (or its derived forms) cannot be used for denials.

One element that allows denials and which could be problematic for our analysis is English *fuck*, which is not considered to be inherently negative. While the denial use of *fuck* has not been dealt with (to our knowledge) in the literature before, this element has similar properties to *gwai2* in an utterance like (35) (taken from the British National Corpus).

- (35) "Why?" Hitch shrugged. "Fuck knows. Like I said, I'm just doing what I'm told."

It is not clear that this use extends beyond the verb *know*, and how widespread that use is. It appears to be mostly a British English phenomenon: there are 9 occurrences in the British National Corpus, against only 3 in the Corpus of

⁴ Compare with the case of French *vache/vachement* ('bovine/cowish' and its derived adverb) which are underspecified in a similar way as English *fucking*.

Contemporary American English which is 5 times bigger. The use in question could conceivably be traced back to a time when *fuck* was negative, only to be bleached at a later time, or that use could have a different origin altogether.

This all suggests dimensions along which expressives can be compared cross-linguistically, namely their inherent connotation (or absence of), and their scopal properties notably their ability to modify echoic content. We leave such considerations to future work.

References

- Beltrama, A.: Bridging the gap: intensifiers between semantic and social meaning. Ph.D. thesis, University of Chicago, Chicago (2016)
- Beltrama, A., Lee, J.L.: Great pizzas, ghost negations: the emergence and persistence of mixed expressives. In: Csapak, E., Zeijlstra, H. (eds.) *Proceedings of Sinn und Bedeutung 19*, Göttingen, Germany, pp. 143–160 (2015)
- Bylina, L., Sudo, Y.: Varieties of intensification. *Nat. Lang. Linguist. Theory* **33**(3), 881–895 (2015). <https://doi.org/10.1007/s11049-015-9291-y>
- Carston, R.: Metalinguistic negation and echoic use. *J. Pragmat.* **25**, 309–330 (1996). [https://doi.org/10.1016/0378-2166\(94\)00109-X](https://doi.org/10.1016/0378-2166(94)00109-X)
- Clark, H.H.: *Using Language*. Cambridge University Press, Cambridge (1996)
- Constant, N., Davis, C., Potts, C., Schwarz, F.: The pragmatics of expressive content: evidence from large corpora. *Sprache und Datenverarb.* **33**(1–2), 5–21 (2009)
- Eckardt, R.: APO: avoid pragmatic overload. In: Mosegaard, M.B., Visconti, J. (eds.) *Current Trends in Diachronic Semantics and Pragmatics*, pp. 21–41. Emelard, Bingley (2009)
- Ginzburg, J., Sag, I.A.: *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. CSLI Lecture Notes, vol. 123. CSLI Publications, Stanford (2000)
- Ginzburg, J.: *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford (2012)
- Hara, Y.: Semantics and pragmatics of cantonese polar questions: an inquisitive approach. In: Aroonmanakun, W., Boonkwan, P., Supnithi, T. (eds.) *Proceedings of PACLIC 28*, Phuket, Thailand, pp. 605–614 (2014)
- Jeffrey, R.: *Subjective Probability, The Real Thing*. Cambridge University Press, Cambridge (2004)
- Kwok, H.: *Sentence Particles in Cantonese*. Center of Asian Studies, University of Hong Kong (1984)
- Lee, P.P.L., Chin, A.C.O.: A preliminary study on cantonese gwai ‘ghost’. In: Sze-wing, T., Sio, J. (eds.) *Studies in Cantonese Linguistics*, vol. 2, pp. 33–54. Linguistic Society of Hong Kong, Hong Kong (2007)
- Matthews, S., Yip, V.: *Cantonese: A Comprehensive Grammar*, 2nd edn. Routledge, Abingdon (2011)
- McCready, E.: What man does. *Linguist. Philos.* **31**, 671–724 (2008). <https://doi.org/10.1007/s10988-009-9052-7>
- McCready, E.: Varieties of conventional implicature. *Semant. Pragmat.* **3**(8), 1–57 (2010). <https://doi.org/10.3765/sp.3.8>
- McCready, E.: Emotive equilibria. *Linguist. Philos.* **35**(3), 243–283 (2012). <https://doi.org/10.1007/s10988-012-9118-9>

- Potts, C.: *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford (2005)
- Potts, C.: The expressive dimension. *Theor. Linguist.* **33**, 165–198 (2007). <https://doi.org/10.1515/TL.2007.011>
- Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: *A Comprehensive Grammar of the English Language*. Longman, Harlow (1985)
- Spenader, J., Maier, E.: Contrast as denial in multi-dimensional semantics. *J. Pragmat.* **41**, 1707–1726 (2009). <https://doi.org/10.1016/j.pragma.2008.10.005>



Evidentials in Causal Premise Semantics: Theoretical and Experimental Investigation

Yurie Hara^{1,3}(✉), Naho Orita², and Hiromu Sakai¹

¹ Waseda University, Tokyo, Japan

yuriehara@aoni.waseda.jp, hsakai@waseda.jp

² Tokyo University of Science, Noda, Chiba, Japan
orita@rs.tus.ac.jp

³ Hokkaido University, Sapporo, Japan

Abstract. We formalize the causal component of Davis & Hara's (2014) analysis of Japanese evidentiality, which defines "indirect evidence" as an observation of the *effect* state of the cause-effect dependency. The analysis correctly predicts that uttering *p-youda* only commits the speaker to 'if *p*, *q* must be true' but not to the prejacent *p*, and successfully derives the asymmetry between the prejacent *p* and the evidence source *q*. Also, the results of the rating study and the corpus study show that the interpretation and the distribution of evidentials are subject to the cause-effect dependencies.

Keywords: Evidentiality · Causality · Modality
Causal premise semantics · Naturalness rating experiment
Corpus study · Implicature · Causal network

Japanese has an indirect evidential morpheme *youda*, which gives rise to (at least) two messages:

- (1) Ame-ga futta youda.
rain-NOM fell EVID
'It seems that it rained.'
Message 1: "It rained." (M1)
Message 2: "The speaker has indirect evidence for 'it rained'." (M2)

Formal studies of evidentiality center around the following two questions: Q1. What are the statuses of the two messages? Q2. What is indirect evidence? Davis & Hara (2014) (D&H, hereafter) argued that unlike previous studies, M1 in (1) is an implicature while M2 is the assertional content of (1). Furthermore, D&H claim that indirect evidence for *p* is some state *q* which is usually caused by *p*. Thus, the at-issue content of (1) is that the speaker observed some state (say, wet streets) which is usually caused by "it rained". Although D&H's analysis overcomes the problems of the previous studies, the notion of causality is left as

a primitive. The goal of this paper is to formally model the causality component in the interpretation of evidentials in the framework of causal premise semantics (Kaufmann 2013). We also report the rating study which supports the idea that causal dependency affects the felicitous use of the evidential morpheme.

This paper is structured as follows: We first review D&H's analysis which defines evidentiality as an observation of the effect state of the asymmetric causal relation in Sect. 1. To formalize the causality component in D&H's analysis, we review Kaufmann's (2013) causal premise semantics and demonstrate how we derive the evidential interpretations in Sect. 2. Sections 3 and 4 report the rating study and the corpus study, respectively. Both studies support the proposal that interpretation and availability of evidentials are subject to the causal dependencies. Section 5 concludes the paper.

1 Davis and Hara (2014)

1.1 *p* in *p-youda* Is Not an Epistemic Commitment

The previous studies on evidentials (Izvorski 1997; Matthewson et al. 2006; McCready and Ogata 2007) argue that evidentiality is a kind of modality. That is, *Evid(p)* entails *Modal(p)*. In other words, M1 in (1) is the at-issue commitment, while M2 is a presupposition. However, Davis and Hara (2014) (D&H, hereafter) show that this treatment cannot be maintained since the prejacent *p* in *p-youda* is cancellable as in (2).¹

- (2) Ame-ga futta youda kedo, jitsu-wa futte-nai.
rain-NOM fell EVID but fact-TOP fall-NEG
'It seems that it rained, but in fact it didn't.'

In (3) and (4), in contrast, both a bare assertion *p* and *Modal(p)* commit the speaker to *p*, thus *p* cannot be cancelled.

- (3) #Ame-ga futta kedo jitsu-wa futtenai.
rain-NOM fell but fact-TOP fall-NEG
'#It rained but in fact it didn't.'
- (4) #Ame-ga futta darou kedo jitsu-wa futtenai.
rain-NOM fell probably but fact-TOP fall-NEG
'#Probably, it rained but in fact it didn't.'

1.2 What Is Indirect Evidence?

McCready and Ogata (2007) treat evidentials as modals and offer a Bayesian semantics for evidentials, including *youda*. McCready and Ogata's account has the following two components: *p-youda*, relativized to agent *a*, indicates that 1. some information *q* has led *a* to raise the subjective probability of *p*. 2. *a*

¹ A similar argument is made for reportative evidentials by Faller (2002); Murray (2010); AnderBois (2014).

takes p to be probably but not certainly true ($.5 < P_a(p) < 1$) after learning q . According to McCready and Ogata, thus, what counts as evidence is some information q that has led a to raise the subjective probability of p . In (5), a learns that the street is wet, which has led a to raise her subjective probability of p , hence the use of *youda* is acceptable. McCready and Ogata's theory provides a concrete way to define evidence and a reasonable analysis for (5).

- (5) a. (Looking at wet streets)
 b. Ame-ga futta youda.
 rain-NOM fell EVID
 'It seems that it rained.'

However, D&H show that it makes wrong predictions if we switch p and q , as in (6). Learning that it is raining should also raise the agent's subjective probability of "the streets are wet", thus McCready and Ogata wrongly predict that *youda* is acceptable in (6).

- (6) a. (Looking at falling raindrops)
 b. #Michi-ga nureteiru youda.
 street-NOM wet EVID
 '#It seems that the streets are wet.'

1.3 Evidentiality via Causality

From the observations discussed so far, D&H make the following two claims:

1. The prejacent p in *p-youda* is not an epistemic commitment but a cancellable implicature and the evidentiality component (M2 of (1)) is the at-issue commitment of *p-youda*.
2. The notion of evidentiality needs to encode *asymmetric* causal dependencies, e.g., rain causes wet streets but not *vice versa*. In other words, what counts as evidence is an effect state of a cause-effect dependency (see also Sawada 2006; Takubo 2009).

Simply put, D&H define the interpretation of *p-youda* as follows:

- (7) Evid(p) is true at w iff $\exists q$ such that the speaker perceives a state q at w and p causes q .

2 Proposal

Although D&H overcome the problems of the previous analyses, the notion of causality is left unanalyzed. This paper formally implements the notion of causality in Kaufmann's (2013) causal premise semantics. Kaufmann introduces causal networks to Kratzer's (2005, a.o.) premise semantics to interpret counterfactuals. We show that the same apparatus can predict interpretations of evidentials.

2.1 Premise Structures

Kaufmann’s framework extends Kratzerian premise semantics by deriving and ranking premise sets. Let f be a Kratzerian conversational background, which is a function from possible worlds to sets of propositions. Then, \mathbf{f} is a *premise background* which is a function from possible worlds to sets of sets of propositions defined in (8). Note that \mathbf{f} alone contains no more or less information than f .

- (8) A *premise background* \mathbf{f} structures a Kratzerian conversational background f iff at all worlds v , $\mathbf{f}(v)$ is a set of subsets of $f(v)$. (Kaufmann 2013, 1144)

To introduce the ranking of premise sets, a set of *sequence structures* is recursively defined as in (9). “ $\leq_1 \times \leq_2$ ” signifies the lexicographic order on the Cartesian product.

- (9) a. If Φ is a set of sets of propositions, then $\langle \Phi, \leq \rangle$ is a (basic) sequence structure.
 b. If $\langle \Phi_1, \leq_1 \rangle$ and $\langle \Phi_2, \leq_2 \rangle$ are sequence structures, then so is $\langle \Phi_1, \leq_1 \rangle * \langle \Phi_2, \leq_2 \rangle$, defined as $\langle \Phi_1 \times \Phi_2, \leq_1 \times \leq_2 \rangle$. (Kaufmann 2013, 1146)

A premise structure that is used for interpreting modal sentences is the set of *consistent* sequence structures (10).

- (10) *Premise structure:*
 Prem($\langle \Phi, \leq \rangle$) is the pair $\langle \Phi', \leq' \rangle$, where Φ' is the set of consistent sequences in Φ and \leq' is the restriction of \leq to Φ' . (Kaufmann 2013, 1147)

2.2 Causal Premise Semantics

Now we are ready to introduce causal structures (11) to capture causal asymmetries. A causal network has two components. The first part is a *directed acyclic graph* (DAG) in which vertices represent variables/partitions (e.g., R , H , and D in Fig. 1 representing “whether it is raining”, “whether water is hose-sprayed” and “whether streets are dry”, respectively) and edges represent causal influence. The second component is that only the values of its immediate parents influence each variable.

- (11) A *causal structure* for non-empty W is a pair $\mathcal{C} = \langle U, < \rangle$, where U is a set of finite partitions on W and $<$ is a directed acyclic graph over U . (Kaufmann, 2013, 1151)

As with Kaufmann, we assume that causal dependency is not deterministic: the values of its parents determine whether the value of each variable is a necessity or a possibility (12). Furthermore, the premise background constrained by the ordering source determines the value of parents.

- (12) a. $\text{Must}(q)$ is true at $\mathbf{f}, \mathbf{g}, w$ iff q is a necessity relative to $\text{Prem}((\mathbf{f} * \mathbf{g})(w))$.
- b. $\text{May}(q)$ is true at $\mathbf{f}, \mathbf{g}, w$ iff q is a possibility relative to $\text{Prem}((\mathbf{f} * \mathbf{g})(w))$. (Kaufmann 2013, 1148)

In order to interpret evidentials, we follow Kaufmann (2013) and postulate a causal premise background \mathbf{f}_c . \mathbf{f}_c consists of causally relevant truths (13b).

- (13) a. The set Π^U of *causally relevant propositions* is the set of all cells of all partitions in U .
- b. The set of *causally relevant truths* at w : $\Pi_w^U = \{p \in \Pi^U \mid p \text{ is true at } w\}$ (U is omitted hereafter.) (Kaufmann 2013, 1152)

Furthermore, \mathbf{f}_c is constrained by the closure under ancestors, which ensures the *asymmetric* relation between variables X and Y . We need to introduce two notions to define the closure under ancestors, *setting* and *descendant*:

- (14) a. A variable X is *set* in a set of propositions P iff exactly one of X 's cells is in P .
- b. X is a *descendant* of Y iff there is a path from Y to X of zero or more steps along the direction of causal influence. (Kaufmann 2013, 1153)

Closure under ancestors is defined as follows:

- (15) A subset P' of P is *closed under ancestors* in P iff [for all $X, Y \in U$ such that X is a descendant of Y and both are set in P], [if X is set in P' , then Y is also set in P']. (Kaufmann 2013, 1153)

Taken together, \mathbf{f}_c is postulated as in (16).

- (16) $\mathbf{f}_c(w) := \{X \subseteq \Pi_w \mid X \text{ is closed under ancestors in } \Pi_w\}$ (Kaufmann 2013, 1153)

Also, the other premise background, i.e., the ordering source \mathbf{g} satisfies the *Causal Markov condition* relative to a causal structure \mathcal{C} . To define Causal Markov condition, a brief introduction to Conditional Independence is in order. The idea is the following: Consider a partition $X \in U$ and sets of partitions $\mathbf{Y}, \mathbf{Z} \subseteq U$. X is conditionally independent of \mathbf{Y} given \mathbf{Z} under $\mathbf{g}(w)$ if and only if learning the setting of \mathbf{Y} in \mathbf{Z} does not alter the value of any cells in X :

- (17) *Conditional independence*
 Let \mathbf{g} be a premise background, w a possible world, and U a set of partitions. For any $X \in U$ and disjoint sets $\mathbf{Y}, \mathbf{Z} \subseteq U$ not containing X : X is *conditionally independent* of \mathbf{Y} given \mathbf{Z} under $\mathbf{g}(w)$ iff for all cells $x \in X$, partial settings \mathbf{y} of \mathbf{Y} and settings \mathbf{z} of \mathbf{Z} such that $\mathbf{y} \cup \mathbf{z}$ is consistent, x is a necessity (possibility) relative to $\text{Prem}(\{\mathbf{z}\} * \mathbf{g}(w))$ iff x is a necessity (possibility) relative to $\text{Prem}(\{\mathbf{z} \cup \mathbf{y}\} * \mathbf{g}(w))$ (Kaufmann 2013, 1155)

The idea behind the Causal Markov condition is that any partition X in the causal structure is independent of any of X 's ancestors except for X 's immediate parents. Let $pa(X)$ be the set of X 's parents and $de(X)$ be the set of X 's descendants. Causal Markov condition is defined as follows:

- (18) *Causal Markov condition*
 Let $\mathcal{C} = \langle U, < \rangle$ be a causal structure and \mathbf{g} a premise background. \mathbf{g} satisfies the **Markov condition** relative to \mathcal{C} if and only if for all $w \in W$ and $X \in U$, X is conditionally independent of $U \setminus (de(X) \cup pa(X))$, given $pa(X)$, under $\mathbf{g}(w)$. (Kaufmann 2013, 1156)

2.3 Deriving Evidentiality from Causality

Our interpretation of evidentials is built on the general interpretation of conditionals. In the current framework, we obtain a premise background $\mathbf{f}[p]$ by hypothetically updating a premise background \mathbf{f} with the antecedent proposition p :

- (19) *Hypothetical update*
 For all w : $\mathbf{f}[p](w) := \{\{p\}\} * \mathbf{f}(w)$. (Kaufmann 2013, 1148)

Finally, we define the interpretation of evidentials. $\text{Evid}(p)$ is true at $\mathbf{f}, \mathbf{g}, w$ when there is some state q such that the speaker perceives q at w and q is a necessity relative to $\text{Prem}((\mathbf{f}_c[p] * \mathbf{g})(w))$:

- (20) *Interpretation of evidentials*
 $\text{Evid}(p)$ is true at $\mathbf{f}, \mathbf{g}, w$ iff $\exists q$ such that the speaker perceives q at w and $\text{Must}_p(q)$ is true at $\mathbf{f}, \mathbf{g}, w$.

Note that the preajcent p of p -youda contributes to the antecedent rather than the consequent. In other words, $\text{Evid}(p)$ only commits the speaker to $\text{Must}_p(q)$ and not to $\text{Must}(p)$.

Let us illustrate the working of (20). Consider now the three-variable network in Fig. 1 (the variable H represents “whether water is hose-sprayed”) with the causally relevant propositions $\Pi = \{r, \bar{r}, h, \bar{h}, d, \bar{d}\}$.

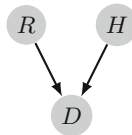


Fig. 1. Rain causes non-dry (wet) streets and hose-spraying water causes non-dry streets.

Let us take example (2) repeated here as (21) first. (21) shows that the speaker can felicitously deny the preajcent proposition of the evidential statement.

- (21) Ame-ga futta youda kedo, jitsu-wa futte-nai.
 rain-NOM fell EVID but fact-TOP fall-NEG
 ‘It seems that it rained, but in fact it didn’t.’

Suppose that at w , it is not raining (\bar{r}), water is hose-sprayed (h), and streets are wet (\bar{d}) as in (22a). By (16), we obtain (22b). Note that $\{\bar{d}\} \notin \mathbf{f}_c(w)$, that is, $\{\bar{d}\}$ is not closed under ancestors in Π_w since D is set but R and H are not set in $\{\bar{d}\}$. Next, we obtain the premise structure (22c) by hypothetically updating $\mathbf{f}_c(w)$ with r and removing any inconsistent sequences. As for the ordring source, let us assume that \mathbf{g} prescribes that normally, water is not hose-sprayed, rain implies wet streets and so does hose-spraying (22d). From (22c) and (22d), we obtain (22e). Since \bar{d} is a necessity relative to $\text{Prem}((\mathbf{f}_c[r]^*\mathbf{g})(w))$, $\text{Must}_r(\bar{d})$ is true at $\mathbf{f}, \mathbf{g}, w$.

- (22) a. $\Pi_w = \{\bar{r}, h, \bar{d}\}$
 b. $\mathbf{f}_c(w) = \{\emptyset, \{\bar{r}\}, \{h\}, \{\bar{r}, h\}, \{\bar{r}, h, \bar{d}\}\}$
 c. $\text{Prem}(\mathbf{f}_c[r](w)) = \{r., r.h\}$
 d. $\mathbf{g}(w) = \{\emptyset, \{\bar{h}\}, \{r \rightarrow \bar{d}\}, \{h \rightarrow \bar{d}\}\}$
 e. $\max \text{Prem}((\mathbf{f}_c[r]^*\mathbf{g})(w)) = \{r.h.(r \rightarrow \bar{d}), r.h.(h \rightarrow \bar{d})\}$

As a result, given that the speaker perceives \bar{d} at w , $\text{Evid}(r)$ is true at $\mathbf{f}_c, \mathbf{g}, w$ because $\text{Must}_r(\bar{d})$ is true at $\mathbf{f}_c, \mathbf{g}, w$, even though r is not true at w . Thus, (21) can be uttered felicitously without the speaker’s commitment to r .

Next, let us see if (20) can explain the evidential asymmetry. Let us derive the felicitous interpretation of (5) repeated here as (23) first.

- (23) a. (Looking at wet streets)
 b. Ame-ga futta youda.
 rain-NOM fell EVID
 ‘It seems that it rained.’

Suppose that the causally relevant truths at v are as in (24a). Note that $\mathbf{g}(w) = \mathbf{g}(v)$.

- (24) a. $\Pi_v = \{r, \bar{h}, \bar{d}\}$
 b. $\mathbf{f}_c(v) = \{\emptyset, \{r\}, \{\bar{h}\}, \{r, \bar{h}\}, \{r, \bar{h}, \bar{d}\}\}$
 c. $\text{Prem}(\mathbf{f}_c[r](v)) = \{r., r.r, r.\bar{h}, r.r\bar{h}, r.r\bar{h}\bar{d}\}$
 d. $\mathbf{g}(v) = \{\emptyset, \{\bar{h}\}, \{r \rightarrow \bar{d}\}, \{h \rightarrow \bar{d}\}\}$
 e. $\max \text{Prem}((\mathbf{f}_c[r]^*\mathbf{g})(v)) = \{r.\bar{h}.(r \rightarrow \bar{d}), r.\bar{h}.(h \rightarrow \bar{d}), r.r\bar{h}\bar{d}.(r \rightarrow \bar{d}), r.r\bar{h}\bar{d}.(h \rightarrow \bar{d})\}$

Assuming that the speaker perceives \bar{d} at v , $\text{Evid}(r)$ is true at $\mathbf{f}_c, \mathbf{g}, v$ because \bar{d} is a necessity relative to $\text{Prem}((\mathbf{f}_c[r]^*\mathbf{g})(v))$, hence $\text{Must}_r(\bar{d})$ is true at $\mathbf{f}_c, \mathbf{g}, v$. The speaker of (23) is asserting that she perceived wet streets and if it rains, streets must be wet, i.e., rain causes wet streets.

Finally, we derive the infelicity of (6), repeated here as (25).

- (25) a. (Looking at falling raindrops)
 b. #Michi-ga nureteiru youda.
 street-NOM wet EVID
 ‘#It seems that the streets are wet.’

Take the same world v as above, thus we have the same causally relevant truths Π_v and causal modal base $\mathbf{f}_c(v)$ as the ones in (24). Now (25) translates to $\text{Evid}(\bar{d})$, so the modal base is altered by hypothetically updating $\mathbf{f}_c(v)$ with \bar{d} as in (26c). Together with the ordering source $\mathbf{g}(v)$, $\text{Evid}(\bar{d})$ is interpreted relative to the set of premise structures shown in (26e).

- (26) .
 a. $\Pi_v = \{r, \bar{h}, \bar{d}\}$
 b. $\mathbf{f}_c(v) = \{\emptyset, \{r\}, \{\bar{h}\}, \{r, \bar{h}\}, \{r, \bar{h}, \bar{d}\}\}$
 c. $\text{Prem}(\mathbf{f}_c[\bar{d}](v)) = \{\bar{d}, \bar{d}.r, \bar{d}.\bar{h}, \bar{d}.r\bar{h}\}$
 d. $\mathbf{g}(v) = \{\emptyset, \{\bar{h}\}, \{r \rightarrow \bar{d}\}, \{h \rightarrow \bar{d}\}\}$
 e. $\max \text{Prem}((\mathbf{f}_c[\bar{d}] * \mathbf{g})(v)) \supseteq \{\bar{d}.\bar{h}.(r \rightarrow \bar{d}), \bar{d}.\bar{h}.(h \rightarrow \bar{d})\}$

$\text{Evid}(\bar{d})$ is *not* true at $\mathbf{f}_c, \mathbf{g}, w$ because r is not a necessity relative to $\text{Prem}((\mathbf{f}_c[\bar{d}] * \mathbf{g})(v))$, i.e., $\text{Must}_{\bar{d}}(r)$ is false at $\mathbf{f}_c, \mathbf{g}, v$. Put another way, (25) is infelicitous since the speaker is making a false claim, ‘wet streets cause rain’ (even if the speaker does observe rain).

2.4 Summary

We offer a formal implementation of the causal component of evidentiality which correctly predicts the lack of commitment to the prejacent proposition and derive the asymmetry of the evidential dependency between the prejacent and the evidence source from the asymmetry between the ancestor and the descendent in a causal network.

3 Naturalness Rating Experiment

Davis and Hara (2014) conducted a series of rating studies to affirm that their observation that the interpretation of evidentials is dependent on causal relations. However, there are at least two problems in their experimental design. First, in D&H’s design, there were no explicit conversational agents specified in the contexts and target sentences, thus it was not clear who holds the knowledge and utters a sentence to convey the knowledge. Second, D&H did not test the subjects’ background knowledge about the plausibility of causal relations, thus it was unclear that the subjects were actually using causal relations to judge the naturalness of the target sentences. This section presents our experiment that overcomes these problems.

3.1 Method

The stimuli had two fully-crossed factors—Context (Witness/(Cause-)Effect) and Verbal morphology (Null/Youda), which resulted in four conditions—Witness-Null, Witness-Youda, Effect-Null, Effect-Youda. (*Youdesu* is a polite form of *youda*.) Each condition had 80 items. 80 fillers were included. As can be seen in (27), the context of each stimulus specify the holder of the relevant knowledge and the speaker of the target sentence.

- (27) Context:
- a. Witness context: A looked outside through the window. C asked A what happened on the phone. A answered:
 - b. (Cause-)Effect context: A's shoes got wet by the paddle on the street. C asked A what happened on the phone. A answered:
- (28) Target sentences (Verbal morphology):
- a. Ame-ga furi masita.
rain-NOM fall POL.PAST
'It rained.' (Null)
 - b. Ame-ga futta youdesu.
rain-NOM fall.PAST EVID.POL
'It seems it rained.' (Youda)

The stimuli were presented via a web-based online survey system, Qualtrics.² The experiment was counterbalanced so that one participant will not see the same context twice.

36 native speakers of Japanese participated in the rating experiment. The participants rated the naturalness of the target sentences as in (28) on a 1-to-7 scale. To analyze the results, a general linear mixed model was run in which context and verbal morphology were the fixed factors and speakers and items were the random factors. The current analysis gives rise to the predictions in (29). Furthermore, the evidential form *p-youda* *semantically* requires that the speaker observes the effect state q of the Cause-Effect relation $\text{Must}_p(q)$, while the null form $p-\emptyset$ *pragmatically* requires that the speaker witnesses the event p via Gricean Quality Maxim. In other words, the Effect requirement for *p-youda* is a semantic commitment while the Witness requirement for $p-\emptyset$ is a cancellable implicature. Thus, we predict (30).

- (29) a. In a Witness context, the null form $p-\emptyset$ is preferred over *p-youda*.
b. In a (Cause-)Effect context, the evidential form *p-youda* is preferred over $p-\emptyset$.
- (30) Witness-Youda is considered as a more serious violation than Effect-Null.

² The output for this paper was generated using Qualtrics software, Version 022018 of the Qualtrics Research Suite. ©2017 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. <http://www.qualtrics.com>.

As a preliminary experiment, we have also tested the plausibility of Witness and Cause-Effect contexts by asking questions as exemplified in (31). We expect the plausibility of Witness/Cause-Effect contexts correlate naturalness of fit conditions as in (32).

- (31) a. Witness check: “A looked outside through the window.”
Do you think A saw the rain?
b. Cause-Effect check: “Because it rained, A’s shoes got wet.”
How natural do you think this sentence is?
- (32) a. The ratings of Witness conditions in witness-checking test correlate the ratings of Witness-Null conditions in the main experiment.
b. The ratings of Cause-Effect conditions in causal checking test correlate the ratings of Effect-Youda conditions in the main experiment.

To analyze the results, a general linear mixed model (Baayen 2008; Baayen et al. 2008; Bates 2005) was run using the lme4 package (Bates et al. 2011) implemented in R (R Core Team 2017). Contexts and sentence-endings were the fixed factors. Speakers and items were the random factors. The p -values were calculated by the Markov chain Monte Carlo method using the LanguageR package (Baayen 2009).

3.2 Result

Figure 2 shows the average and median naturalness ratings of the main experiment. In Witness Contexts, the null ending is preferred over *youda* ($t = 16.995; p < 0.001$). In Effect Contexts, *youda* is preferred over the null ending ($t = 11.448; p < 0.001$). Witness-Youda is considered as a more serious violation than Effect-Null ($t = -6.979; p < 0.001$).

Figures 3(a) and (b) show the ratings of the preliminary experiment. There were weak correlations between the plausibility of contexts and the ratings of the main experiment in Spearman’s rho: between the ratings of witness contexts and the ratings of Witness-Null ($rs = 0.21; p = 0$) and between the ratings of effect contexts and the ratings of Effect-Youda ($rs = 0.26; p = 0$).

3.3 Discussion

The result shows that context affects the choice of the evidential form but does so in a complex way: p - \emptyset *pragmatically* requires Witness context, while p -*youda* *semantically* requires Cause-Effect context. This result supports the current analysis: The semantic content of p -*youda* is that the speaker observes some state q and p causes q .

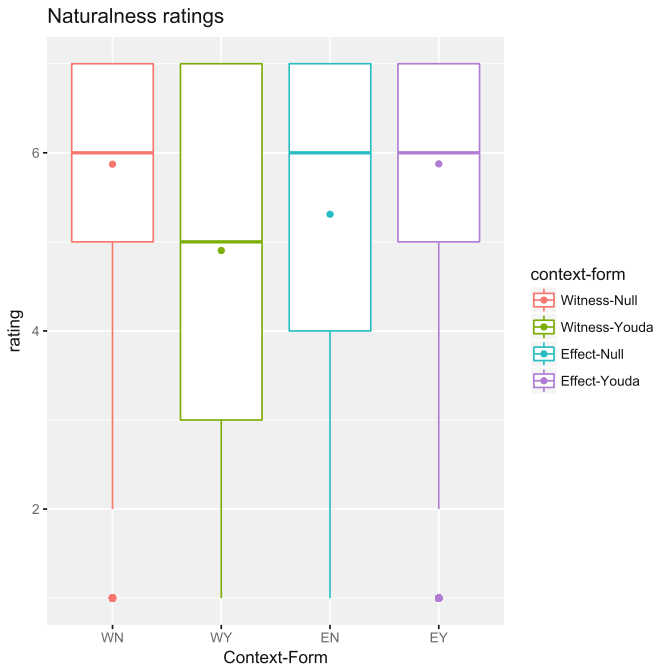


Fig. 2. Naturalness ratings of context-form combinations

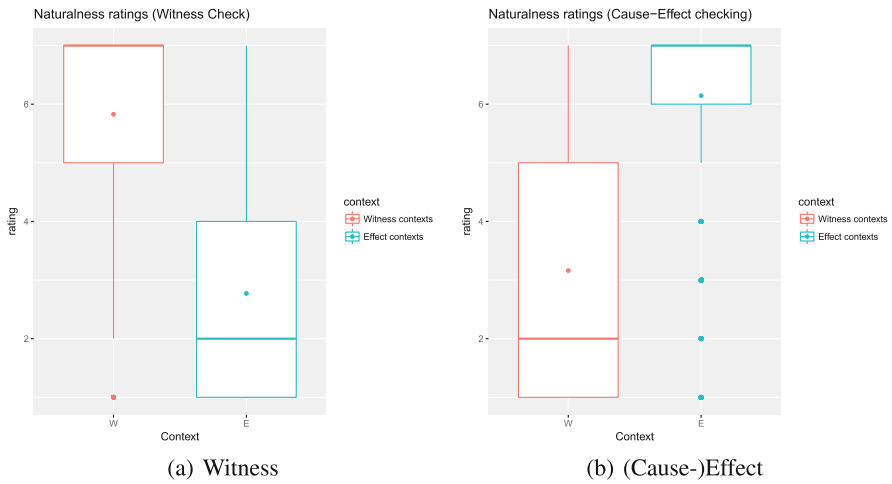


Fig. 3. Naturalness ratings

4 Corpus Study

The theoretical discussion in Sect. 2 shows that the distribution of *youda* is dependent on the causal relation. In particular, *youda* is attached to a proposition that denote a causal event of cause-effect relation. The result of the rating study in Sect. 3 also affirms that the causal factor in the context affects the choice of the evidential morphology. The rating study, however, has its own limitation, since it only shows that causality is a sufficient condition for the choice of the evidential morphology but it does not tell us whether causality is a necessary condition. In other words, there might be factors other than causality that determine the choice of the evidential morphology. To overcome this problem, we conducted a corpus study. If causality is a necessary condition for the choice of *youda*, it is predicted that a significant number of occurrences of *youda* in corpus are attributed to the cause-effect dependency. More specifically, we predict that the proportion of the predicates that denote cause events preceding *youda* is significantly large compared to those of other predicates. Our result shows that the prediction is indeed attested.

4.1 Predictions

In this corpus study, we looked at *youda* and its dual *darou*. As discussed in Sects. 2 and 3, *youda* is attached to a proposition that denotes a causal event of cause-effect relation. Given a causal relation ‘If it rains, the streets are wet.’, *youda* can be attached to the cause event but not to the effect state:

- (33) Cause: rain \rightarrow Effect: wet streets
- a. Michi-ga nureteiru. Ame-ga futta youda.
street-NOM wet. rain-NOM fell EVID
‘The streets are wet. It seems that it rained.’
 - b. #Ame-ga futta. Michi-ga nureteiru youda.
rain-NOM fell. street-NOM wet EVID
‘#It rained. It seems that the streets are wet.’

In contrast, the modal auxiliary *darou* has exactly the opposite distribution. *Darou* cannot be attached to the cause event, but it can be attached to the effect state:

- (34) Cause: rain \rightarrow Effect: wet streets
- a. #Michi-ga nureteiru. Ame-ga futta darou.
street-NOM wet. rain-NOM fell probably
‘The streets are wet. Probably, it rained.’
 - b. Ame-ga futta. Michi-ga nureteiru darou.
rain-NOM fell. street-NOM wet probably
‘It rained. Probably, the streets are wet.’

Takubo (2009) claims that *darou* is a marker of deductive reasoning while *youda* is a maker of abductive reasoning. Thus, according to Takubo (2009),

darou is attached to a consequent of a conditional statement. In our terms, *darou* is attached to a proposition that denotes an effect state (see also Sawada 2006). Our hypotheses are summarized in the following:

- (35) Hypotheses
- a. *Youda* is attached to a causal event of cause-effect dependency.
 - b. *Darou* is attached to an effect state of cause-effect dependency.

Before testing our hypotheses against the corpus data, we need to make several assumptions regarding causality. First, we assume that causality describes a relation between the event that happens first (e.g., rain) and the state that results from the event (e.g., wet streets). Thus, predicates that denote causes of cause-effect dependency tend to denote *events* while predicates that denote effects of cause-effect dependency tend to denote *states*. Second, the cause event temporally precedes the effect state. Our Assumptions are summarized in the following:

- (36) Assumptions
- a. Causes are events while effects are states.
 - b. The cause event temporally precedes the effect state.

Taken together, we have the following predictions:

- (37) Predictions
- a. *Youda* tends to be attached to past-tensed (*-ta* form) and eventive predicates.
 - b. *Darou* tends to be attached to non-past (*-ru* form) and stative predicates.

4.2 Method

To test the predictions in (37), we use Balanced Corpus of Contemporary Written Japanese (Maekawa et al. 2014) containing approximately 100 million words collected from various kinds of Japanese texts. The corpus includes annotations of morphological information. We extract sentences ending with *darou* and *-youda*, resulting in 20110 *youda* sentences and 36439 *-darou* sentences. We exclude nominal predicates preceded by the target auxiliaries, namely, *youda* preceded by a noun-*no* sequence (e.g., *sensei-no-youda*), and *darou* preceded by a noun (e.g., *sensei darou*) in order to focus on tensed predicates, i.e., verbs and adjectives. Also, we excluded *darou* preceded by a predicate-*no* sequence because *no darou* patterns like *youda* as can be seen in (38).

- (38) Cause: rain → Effect: wet streets
- a. Michi-ga nureteiru. Ame-ga futta no darou.
 street-NOM wet. rain-NOM fell PRT probably
 ‘The streets are wet. Probably because it rained.’

b. #Ame-ga futta. Michi-ga nureteiru no darou.
 rain-NOM fell. street-NOM wet PRT probably
 ‘It rained. Probably, the streets are wet.’

The particle *no* is known to be multifunction. It functions as a nominalizer, a question particle, an evidential marker, a genitive case, a pronoun, etc. To make our study simple, we abstract away from these *no-darou* sequences.

4.3 Results

Figure 4 shows the distribution of parts of speech preceding *darou* and *youda*. Figure 5 shows the distribution of auxiliaries preceding *darou* and *youda*. The χ^2 test shows a significant difference between *youda* and *darou* in these two distributions ($p < 0.0001$). Tables 1 and 2 show the frequency of each item.

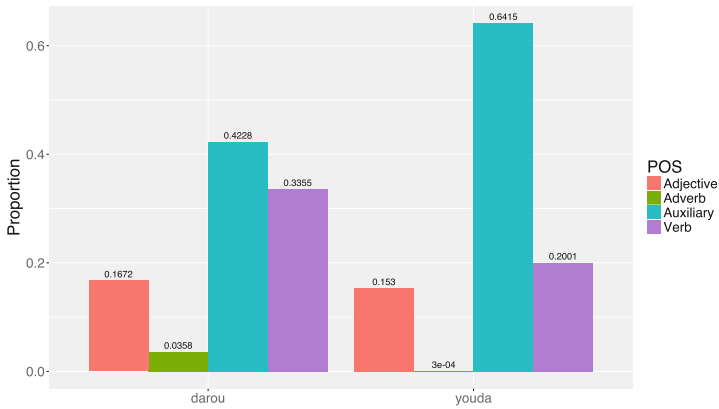


Fig. 4. Distributions of parts of speech preceding *darou* and *youda*

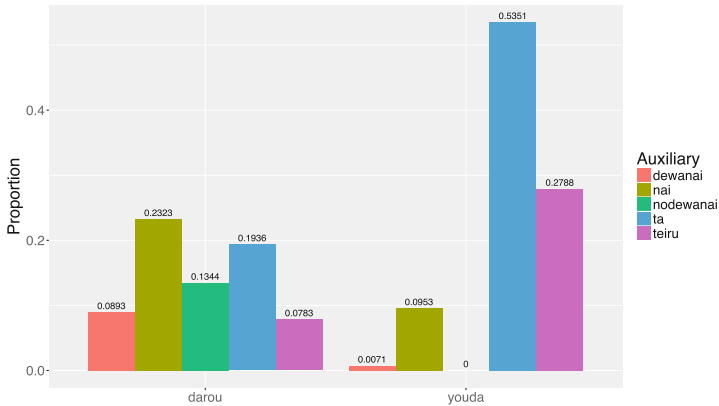


Fig. 5. Distributions of auxiliaries preceding *darou* and *youda*

Table 1. Frequencies of parts of speech

	<i>darou</i>	<i>youda</i>
Adjective	6092	3077
Adverb	1303	7
Auxiliary	15407	12900
Verb	12227	4024
Else	1410	102

Table 2. Frequencies of auxiliaries

	<i>darou</i>	<i>youda</i>
<i>dewanai</i>	1144	89
<i>nai</i>	3579	1230
<i>nodewanai</i>	1722	0
<i>ta</i>	2983	6903
<i>teiru</i>	1206	3597
Else	2181	720

Note in particular that as for parts of speech (see Fig. 4 and Table 1), adjectives and bare (i.e., non-past) verbs precede *darou* more frequently than *youda*. As for auxiliaries (see Fig. 5 and Table 2), the past tense *-ta* precedes *youda* more frequently than *darou* and the negations, *-dewanai*, *-nai* and *-nodewanai*, precede *darou* more frequently than *youda*.

4.4 Discussion

The results show that the predictions summarized in (37) (repeated here as (39)) are indeed borne out.

(39) Predictions

- a. *Youda* tends to be attached to past-tensed (*-ta* form) and eventive predicates.
- b. *Darou* tends to be attached to non-past (*-ru* form) and stative predicates.

The past tense *-ta* precedes *youda* more frequently than *darou* (see Fig. 5), while the dictionary forms of verbs (*-ru* form) precede *darou* more frequently than *youda* (see Fig. 4). Furthermore, the negation auxiliaries, *-dewanai*, *-nai* and *-nodewanai*, precede *darou* more often than *youda* (see Fig. 5). This is compatible with the fact that a negation of an event predicate refers to a fusion-state (see Krifka 1990). Finally, adjectives which tend to denote states rather than events precede *darou* more frequently than *youda* (see Fig. 4).

Put another way, the proportion of the past tense *-ta* preceding *youda* (0.343262059) is bigger than any other categories. Also, the proportion of the non-past tense items (i.e., bare verbs, bare adjectives and negation morphemes) preceding *darou* (0.679601526) is bigger than any other categories.

Thus, together with the assumptions given in (36), our hypotheses make correct predictions. The evidential morpheme *youda* tends to be attached to a cause event of a cause-effect relation.

5 Conclusion

We formalized the causal component of D&H's analysis of Japanese evidentiality, which defines "indirect evidence" as the *effect* state of the cause-effect dependency, correctly predicts that uttering *p-youda* only commits the speaker to $\text{Must}_p(q)$ but not to the prejacent *p*, and successfully derives the asymmetry between the prejacent *p* and the evidence source *q*. Also, the results of the rating study and corpus study show that evidentiality is dependent on causality.

Acknowledgement. This research was supported by the project "Cognitive Neuroscience of Linguistic Variation in Pragmatic Inference" at the National Institute of Japanese Language and Linguistics (PI: Hiromu Sakai, Waseda University).


References

- AnderBois, S.: On the exceptional status of reportative evidentials. *Proc. SALT* **24**, 234–254 (2014)
- Baayen, H.R.: *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge (2008)
- Baayen, H.R.: *LanguageR*. R package (2009)
- Baayen, H.R., Davidson, D.J., Bates, D.M.: Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390–412 (2008)
- Bates, D.: Fitting linear mixed models in R. *R News* **5**, 27–30 (2005)
- Bates, D., Maechler, M., Bolker, B.: *lme4: Linear mixed-effects models using Eigen and Eigen++*. R package (2011)
- Davis, C., Hara, Y.: Evidentiality as a causal relation: a case study from Japanese *youda*. In: Piñón, C.P. (ed.) *Empirical Issues in Syntax and Semantics*, vol. 10 (2014)
- Faller, M.: *Semantics and Pragmatics of Evidentials in Cuzco Quechua*. Ph.D. thesis, Stanford University (2002)
- Izvorski, R.: The present perfect as an epistemic modal. *Proc. SALT* **7**, 222–239 (1997)
- Kaufmann, S.: Causal premise semantics. *Cognit. Sci.* **37**, 1136–1170 (2013)
- Kratzer, A.: Constraining premise sets for counterfactuals. *J. Semant.* **22**, 153–158 (2005)
- Krifka, M.: Boolean and non-boolean 'and'. In: Kálmán, L., Pólos, L. (eds.) *Papers from the Second Symposium on Logic and Language*, pp. 161–188. Akadémiai Kiadó, Budapest (1990)
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y.: Balanced corpus of contemporary written Japanese. *Lang. Resour. Eval.* **48**(2), 345–371 (2014)
- Matthewson, L., Rullmann, H., Davis, H.: Evidentials are epistemic modals in St'át'imcets. In: Kiyota, M., Thompson, J.L., Yamane-Tanaka, N. (eds.) *Papers for the 41st International Conference on Salish and Neighbouring Languages*, vol. 18, pp. 221–263. University of British Columbia Working Papers in Linguistics (2006)
- McCready, E., Ogata, N.: Evidentiality, modality and probability. *Linguist. Philos.* **30**(2), 35–63 (2007)
- Murray, S.E.: *Evidentiality and the Structure of Speech Acts*. Ph.D. thesis, Rutgers (2010). <http://www.semanticsarchive.net/Archive/WViOGQxY/>

- R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). <https://www.R-project.org/>
- Sawada, H.: Modaritii. Kaitakusha (2006)
- Takubo, Y.: Conditional modality: two types of modal auxiliaries in Japanese. In: Pizziconi, B., Kizu, M. (eds.) Japanese Modality: Exploring its Scope and Interpretation. Palgrave Macmillan (2009)



Annotating Syntax and Lexical Semantics With(out) Indexing

Alastair Butler¹ and Stephen Wright Horn²

¹ Faculty of Humanities, Hirosaki University,
Bunkyo-cho 1, Hirosaki-shi 036-8560, Japan
ajb129@hotmail.com

² National Institute for Japanese Language and Linguistics, 10-2 Midori-cho,
Tachikawa City, Tokyo 190-8561, Japan
horn.s.w@ninjal.ac.jp

Abstract. A new method of annotation is introduced that is able to establish a rich range of dependencies without recourse to indexing. With well defined annotation practices for assigning structure, the index-less annotation can be subsequently processed by means of a mechanism of semantic calculation that identifies relationships (dependencies) by reference to structure. These relationships can be subsequently expressed through a derived indexing, but don't rely on indexing in order to be established. The technique is illustrated capturing dependencies seen with PropBank/FrameNet annotations.

Keywords: Lexical semantics · Syntax · Annotation · PropBank
FrameNet

1 Introduction

This paper offers a new approach to the challenge of annotating language data for encoding syntactic and lexical semantic information, useful e.g., to provide searchable resources for theoretical linguistic research, or training data for Natural Language Processing applications. Such an enterprise rarely gets far without indexing being employed to establish dependencies between constituents. An index might be a shared mark (e.g., a numeral), or might exist as a value that references the position of a component of the annotation within a structure or sequence.

This paper benefited from the comments of two anonymous reviewers and the participants of LENLS 14, who we gratefully acknowledge. We are particularly indebted to Noritugu Hayashi, Yumiko Kinjo, Ken Kishiyama, Iku Nagasaki, and Kei Yoshimoto, for stimulating discussions concerning annotation. This research was supported by the NINJAL Parsed Corpus of Modern Japanese (NPCMJ) project funded by the National Institute for Japanese Language and Linguistics (NINJAL), and the Japan Society for the Promotion of Science (JSPS), Research Project Number: 15K02469.

We consider reliance on indexing in an annotation format to be troublesome. In the first place, using indexing to establish dependencies is frequently costly, since it requires human annotator effort to specify the source of a dependency so it can be marked with an index value. Furthermore, to the extent that the values for indices are associated with target positions either in a sequence or in a structure, the dependencies they are used to establish are brittle: Even small modifications (e.g., changes in segmentation, additions of null elements such as zero pronouns, etc.) can break dependencies established through indexing. Setting aside corrections and additions to an annotated tree, postprocessing or converting annotation with indexing into alternative formats can also be prohibitively difficult. Without a holistic handle on the annotation format (for example, a set of algorithms that can recalculate the values of indices based on the consequences of indefinitely many manipulations of the string or tree), preserving dependencies through such processes can be impossible.

These points are illustrated with examples of existing annotation methods that are reliant on indexing for capturing lexical semantic information (PropBank in Sect. 2, and FrameNet in Sect. 3). These are projects that re-express collocations between arguments and heads as manifestations of tuples of role-types peculiar to a particular lexical item (in the case of PropBank) or to a particular semantic frame (in the case of FrameNet). A generalised description that covers a variety of manifestations can be thought of as a lexical profile. Recovering the elements involved in a natural language instantiation of a particular predicate or frame obviously requires sorted mappings between potentially complex elements in linear order. For a given language, building a descriptively adequate inventory of lexical profiles is often pursued through large scale analyses of natural language corpora. The challenge is to annotate the corpus in a way that makes the relationships reliably recoverable and enables the data to be enriched, corrected, or modified, all while not placing undue burdens on a human annotator. A detailed look at the strategies employed by PropBank and FrameNet reveals some weaknesses in the use of indexing to achieve these ends.

As an alternative, Sect. 4 introduces a novel method of annotation that is able to establish a rich range of dependencies without recourse to indexing. With well defined annotation practices for assigning structure, the index-less annotation can be subsequently processed by means of a mechanism of semantic calculation that identifies relationships (dependencies) by reference to structure. These relationships can be subsequently expressed through a derived indexing, but don't rely on indexing in order to be established. The technique is illustrated capturing the dependencies seen with the earlier PropBank/FrameNet annotations.

Examples in Sects. 2, 3, and 4 will involve annotations for the following sentence:

- (1) People John asked smiled while laughing at him.

Example (1) has a participle clause controlled by the subject noun phrase, which in turn contains a relative clause, so that *people* has an argument role to play for all the verbs: *asked*, *smiled*, and *laughing*. The subject relation obtaining between

John and *asked*, the subject relation between *people John asked* and *smiled*, and the object relation between *at him* and *laughing*, are all local dependencies. There are also two non-local dependencies to capture: Control by upstairs argument *people John asked* into the participle *laughing*, and the unbounded dependency between the indirect object position for *asked* in the relative clause and the head of that relative clause *people*. Also, *him* can be anaphorically linked to *John*. Setting aside the possible existence of an elided direct object of *asked* and the mediation of a null relative pronoun in the complex noun phrase, the sentence in (1) then exhibits at least six basic dependencies that are relevant for generating lexical profiles.

2 PropBank Annotation

This section considers PropBank (Bonial et al. 2010) annotation for (1) as a first example of annotation reliant on indexing. PropBank annotation is carried out on top of the phrase structure trees of Penn Treebank analysed data (Bies et al. 1995). Given such trees, PropBank labels every predicate occurrence with a sense ID (e.g., `ask.01`) and labels every argument of the predicate with a semantic role in relation to the sense ID. The sense ID relates to a frame file that provides predicate-specific descriptions of the semantic roles.

A PropBank entry is specific to a particular word-sense (or roleset), distinguished from other senses of the same word-form by the set of roles performed by co-occurring elements. For example, the first verb in (1) is an instance of `ask.01` “to ask a question” with the following associated roles (which include links to richer VerbNet role descriptions):

- ARG0-PAG: asker (vnrole: 37.1.2-agent)
- ARG1-PPT: question (vnrole: 37.1.2-topic)
- ARG2-GOL: hearer (vnrole: 37.1.2-recipient)
- ARG3-PRD: attributive

To extrapolate a full set of roles for a given word-sense, it is often necessary to look at more than one instance of its use. (We have already seen in (1) that it is possible to drop the direct object of a ditransitive verb in some contexts in English, so an instantiation of a ditransitive verb might not exhibit all the roles relevant to its word-sense.) This fact alone underscores the utility of large-scale natural language corpus analysis. PropBank annotation contains PropBank instances with the following format (here abridged, with paths to reference files and trees removed):

```

<instance>      ::= <framefile> <lemma>.<roleset_id> ----- (<argument>)+
<argument>     ::= <terminal_id>:<height>-<label>
<framefile>    ::= name of the frame file to consult for this roleset
<roleset_id>   ::= sense ID of the predicate
<terminal_id>  ::= ID of the 1st terminal node in this argument
<height>       ::= height of this argument from its 1st terminal node
<label>        ::= PropBank label

```

The following is an example of PropBank annotation for (1) based on the tree structure that follows:

```
ask ask.01 ----- 2:1-ARGO-PAG 3:0-re1 0:1*1:1*4:1-ARG2-GOL
smile smile.01 ----- 0:2-ARGO-PAG 5:0-re1
laugh laugh.01 ----- 0:2*7:1-ARGO-PAG 8:0-re1 9:1-ARGM-ADV
```

The first PropBank instance indicates a verb-predicate “ask” whose `<roleset_id>` (sense ID) is `ask.01`, which is the 4th terminal node (`<terminal_id> = 3`) of the tree presented below. Note that the roles appearing in this instance are `ARGO-PAG asker` and `ARG2-GOL hearer`, a subset of the full roleset. The PropBank project uses data in the Penn Treebank corpus format: Each sentence forms a tree which is a connected graph of bracketed nodes where dominance is expressed by embedding. Thus nodes B and C are immediate constituents of the root node A in the expression (A (B leaf_1) (C leaf_2)). The strings that form the text (here, `leaf_1`, `leaf_2`) are terminals. In the tree below the overall root is the sentence node (S ...) (or “S”) and the first constituent under S is a subject noun phrase with index 1. The index corresponds to a null subject pronoun in the participle clause *while laughing at him*. This NP-SBJ-1 is in turn complex, consisting of a noun phrase modified by a relative clause construction SBAR. This contains an empty relative pronoun (WHNP-2 (-NONE- 0)) with index 2 paired with a sentence S. This sentence contains a subject noun phrase and a verb phrase. The verb phrase contains a past-tense verb VBD and a noun phrase with a null element trace *T*-2 which shares an index with the preceding relative pronoun.

```
0: (S (NP-SBJ-1 (NP (NNS People))
1:         (SBAR (WHNP-2 (-NONE- 0))
2:             (S (NP-SBJ (NNP John))
3:               (VP (VBD asked)
4:                 (NP (-NONE- *T*-2))))))
5:   (VP (VBD smiled)
6:     (SBAR-TMP (IN while)
7:       (S (NP-SBJ (-NONE- *PRO*-1))
8:         (VP (VBG laughing)
9:           (PP (IN at)
10:            (NP (PRP him)))))))
11:  (. .))
```

The arguments of the verb “ask” are identified by position in the tree by the following coordinates: `<terminal_id>:<height>-<label>`, where the values for the serial order of a terminal node and its height are counted from 0. `2:1-ARGO-PAG` identifies that the constituent beginning with the third terminal node `John` and subsumed under the second node label above the terminal NP-SBJ corresponds to `ARGO-PAG` (which is “asker” in the PropBank entry for “ask”). `3:0-re1` identifies that the constituent beginning from the fourth terminal node and subsumed under the label for the terminal is the word-form instantiating the lemma “ask”.

0:1*1:1*4:1-ARG2-GOL indicates that the chain (NNP *people*) -> (WHNP-2 (-NONE-0)) -> (NP (-NONE- *T*-2)) corresponds to ARG2-GOL (which is “hearer” in the PropBank entry).

Human annotators (using the Jubilee graphical tool) associate each predicate with a sense ID, as well as mark the arguments of the predicate that are local to the predicate, together with a selection of the appropriate semantic role for each argument. In the case of *people John asked*, automatic post-processing establishes the chain of links between the trace, the zero relative pronoun and the modified head, relying on the index 2 of the sourced syntactic tree. Thus, the method for describing dependencies in the PropBank annotation for (1) requires that the source tree be in the correct Penn Treebank format, which is itself established with indexing. This example illustrates how PropBank annotation requires phrase structure trees that will not change in any structural aspect, since changes create domino effects that knock off the alignments of the indexing.

3 FrameNet Annotation

In this section we present another example of an annotation project which characterises dependencies between elements in a text. FrameNet (Ruppenhofer et al. 2016), like PropBank, is a semantic role labeling annotation project. But while in PropBank a given sense ID and associated roles relate to a frame file providing predicate-specific descriptions of the semantic roles, a FrameNet role relates to a frame that subsumes multiple predicates with various manifestations for frame-specific roles.

Like PropBank annotation, FrameNet annotation is mediated by an annotation tool. However there is no linking to any external parsed annotation. Instead, predicates and their arguments are anchored to the character string of the source data, essentially by having a human paint in character spans with the annotation tool. The following demonstrates (abridged) resulting FrameNet XML annotation for (1):

```
<text>People John asked smiled while laughing at him .</text>
<annotationSet luID="8421" luName="ask.v" frameID="40"
  frameName="Questioning">
  <layer rank="1" name="Target">
    <label end="16" start="12" name="Target"/>
  </layer>
  <layer rank="1" name="FE">
    <label feID="1" end="10" start="7" name="Speaker"/>
    <label feID="2" end="5" start="0" name="Addressee"/>
  </layer>
</annotationSet>
<annotationSet luID="9738" luName="smile.v" frameID="794"
  frameName="Making_faces">
  <layer rank="1" name="Target">
    <label end="23" start="18" name="Target"/>
  </layer>
</annotationSet>
```



```

</layer>
<layer rank="1" name="FE">
  <label feID="1" end="16" start="0" name="Agent"/>
  <label feID="2" end="45" start="25" name="Time"/>
</layer>
</annotationSet>
<annotationSet luID="8720" luName="laugh.v" frameID="69"
  frameName="Make_noise">
  <layer rank="1" name="Target">
    <label end="38" start="31" name="Target"/>
  </layer>
  <layer rank="1" name="FE">
    <label feID="1" end="16" start="0" name="Sound_source"/>
    <label feID="1" end="45" start="40" name="Addressee"/>
  </layer>
</annotationSet>

```

The annotation consists of source character data as the `<text>` content, followed by annotations for the predicates as `<annotationSet>` content. Predicates are picked out with `start` and `end` attributes for a `Target`. For example, the `Target` for the `annotationSet` with `luName="ask.v"` is the 13th (`start="12"`) to 17th (`end="16"`) characters of the text content, namely, `"asked"`. Arguments with roles designated through the `name` attribute are similarly established as spans of characters of the source string.

This style of annotation is robust insofar as the string remains constant, but it is still susceptible to problems if word segmentations were to change, since word spaces are counted as single characters. This is a notable problem for a language like Japanese, where word boundaries can be a point of contention. This system precludes the addition of null pronouns as well, another problematic point for pervasive pro-drop languages such as Japanese.

4 Annotation with Subsequent Interpretation

This section sketches an alternative method for obtaining the kinds of information on dependencies that are stipulated in the sort of index based annotation seen with PropBank and FrameNet, only this new method derives output from annotations that need no recourse to indexing. In this system all structural relations are related to rules that interpret those relations as dependencies. With such a system in place, (1) can be annotated as follows:

```
(IP-MAT (NP-SBJ (NS People)
              (IP-REL (NP-SBJ (NPR John))
                       (VBD asked)
                       (NP-OB1 *T*)))
         (VBD smiled)
         (PP-TMP (P-ROLE while)
                 (IP-ADV (VAG laughing)
                          (PP (P-ROLE at)
                               (NP (PRO him))))))
         (PU .))
```

As a phrase structure analysis, this is much akin to the Penn Treebank annotation seen in Sect. 2. Yet while there is construction particular clause typing (“IP-REL”) and a trace (“(NP-OB1 *T*)” indicating an object argument for *asked*) to form the relative clause encoding, there is no indexing, as was the case with the Penn Treebank variant, where trace (“(NP (-NONE- *T*-2))”) is indexed to a null relative pronoun (“(WHNP-2 (-NONE- 0))”). Furthermore, there is no marking for the control dependency, while the Penn Treebank variant employs “(NP-SBJ (-NONE- *PRO*-1))” coindexed with *People John asked*. The simpler annotation above is possible because annotation practices are specifically designed to take advantage of a systematic conversion into a format that can be processed to render a semantic expression representing all dependencies. Grammatical processes such as control and relativisation are defined over generalised structural configurations, allowing the automatic establishment of non-local dependencies as long as the right structural conditions are met. For example, the subject noun phrase of the verb *smiled* is identified as the subject in the subordinate clause headed by *laughing* by virtue of the definition of the relationship of control as obtaining between an argument (subject, direct object, indirect object, etc.) and an empty subject position in an immediately subordinate clause. An accessibility hierarchy resolves cases where more than one candidate argument is present. Likewise, in a relativisation, “identity” between trace and modified head is established by satisfying the conditions that IP-REL dominate the trace and be sister to the head. The conditions are flexible enough to allow for variation in word order and depth of embedding.

When human annotators of the parsed corpus assign structures to meanings according to such definitions, a conversion pipeline can recast the dependencies so established into very different forms. At the core is the production of a well-formed expression in the Scope Control Theory (SCT) system of Butler (2015). This decomposes lexical items into combined predicates/sorted bindings (SCT expressions) and provides a system of calculation that manages sequences of values assigned to the binding names by taking into account syntactic information over the entire span of a discourse. With the calculation, an SCT expression is transformed into a predicate logic expression in which the original syntactic dependencies from the source parse are recast as, for example, a given variable appearing as a bound argument for multiple predicates, or as identity between different variables. Dependencies thus established can be subsequently re-expressed as derived indices that can be introduced into the original tree structure.

The conversion presented here takes the form of a number of transformation steps, the first of which involves normalising the tree structure. Language-specific algorithms are applied to produce structures capturing grammatical relations in a form which is readable by a language independent script. This step regularises structure and reduces the inventory of tag labels. More specifically, “NP-SBJ” and “NP-OB1” tags are changed into NPs tagged “NP” which are placed under a “PP” projection that gives role information (“ARG0” for “-SBJ” in a non-passive clause, and “ARG1” for “-OB1”). As a consequence, all arguments are “PP” projections with role information, so, e.g., “(PP (P-ROLE at) (NP (PRO him)))” warrants no change. Other changes include regularising “NS”, marking a plural noun, to “N”, a noun without number information, with the plural information retained as “SORT” information. Similarly, all verbs are regularised to the “VB” tag, e.g., “VBD” (past tense verb) leads to the creation of an “ACT” tag to retain the tense information. Thus a normalised tree is reached:

```
(IP-MAT (ACT past)
  (PP (P-ROLE ARG0)
    (NP (SORT *GROUP*)
      (N People)
      (CP-REL (IP-SUB (ACT past)
        (PP (P-ROLE ARG0)
          (NP (NPR John)))
        (VB asked)
        (PP (P-ROLE ARG1)
          (NP *T*))))))
    (VB smiled)
    (PP (SORT TMP_while)
      (P-ROLE TMP_while)
      (IP-ADV (VB laughing)
        (PP (P-ROLE at)
          (NP (PRO him))))))
  (PU .))
```

The second step is to convert the normalised tree into an expression that can serve as input to the semantic calculation system of Scope Control Theory (SCT). A transformation turns the information from the normalised tree into a complex expression in an intermediate language (constructed through declarations in functional programming). This expression is built exploiting the phrase structure by locating any complement for the phrase head to scope over, adding modifiers as elements that scope above the head, and keeping track of the binding names (sorted variables) introduced in the course of the transformation. For example, the following input and output pairs illustrate the intermediate results of the conversion arising from building up the representation of the relative clause for (1).

```

NP-in:
(NPR John)
NP-out:
(NP npr "ENTITY" "John"__LOCAL__)

PP-in:
(P-ROLE ARG0)@(NP npr "ENTITY" "John"__LOCAL__)
PP-out:
(PP-NP npr "ENTITY" "John" "ARG0"__LOCAL__"ARG0")

NP-in:
*T*
NP-out:
(NP arg "T"__LOCAL__)

PP-in:
(P-ROLE ARG1)@(NP arg "T"__LOCAL__)
PP-out:
(PP-NP arg "T" "ARG1"__LOCAL__"ARG1")

IP-SUB-in:
(ACT past)@(PP-NP npr "ENTITY" "John" "ARG0"__LOCAL__"ARG0")@(VB asked)@
(PP-NP arg "T" "ARG1"__LOCAL__"ARG1")
IP-SUB-out:
(IP-SUB-FACT npr "ENTITY" "John" "ARG0" (arg "T" "ARG1" (past ".event"
(verb lc ".event" ["ARG1", "ARGO"] "asked" (gen "EVENT"))))__LOCAL__"ARGO"
@NAME@"ARG1")

CP-REL-in:
(IP-SUB-FACT subord lc nil ((fn lc => (npr "ENTITY" "John" "ARGO" (arg
"T" "ARG1" (past ".event" (verb lc ".event" ["ARG1", "ARGO"] "asked"
(gen "EVENT")))))) [ "ARGO", "ARG1"]__LOCAL__)
CP-REL-out:
(CP-REL-FACT Lam ("h", "T", subord lc nil ((fn lc => (npr "ENTITY" "John"
"ARGO" (arg "T" "ARG1" (past ".event" (verb lc ".event" ["ARG1", "ARGO"]
"asked" (gen "EVENT")))))) [ "ARGO", "ARG1"]__LOCAL__)

```

This conversion transforms the part of speech tags given by the nodes immediately dominating the terminals of the input constituent tree into operations for creating SCT expression content (“npr” (proper name), “arg T” (bound “T” name), “verb”, etc.). Conversion also adds construction information from the constituent nodes (“subord” (subordinate clause), “Lam (“h”, “T”, ...)” (an instruction to make the open “h” binding (the head binding internal to a noun phrase) into a “T” binding (the trace binding internal to a relative clause)), etc.). Conversion also adds information about binding names (e.g., “ARGO” (logical subject role), “ARG1” (logical object role), “at” (“at” role), “h” (nominal binding role), “.event” (event binding)). Conversion also adds instructions (e.g., “gen EVENT”) to generate what will become bound variables of a resulting semantic calculation.

The individual operations called by parts of the normalised tree combine into the complex expression below.

```
( fn fh =>
  ( fn lc =>
    ( some lc fh ".e" ( gen "GROUP")
      ( ( fn lc =>
          ( scon fh "&"
            ( Lam
              ( "h", "T",subord lc nil
                ( ( fn lc =>
                    ( npr "ENTITY" "John" "ARGO"
                      ( arg "T" "ARG1"
                        ( past ".event"
                          ( verb lc ".event" ["ARG1", "ARGO"] "asked"
                            ( gen "EVENT"))))))))
                    [ "ARGO", "ARG1"])))
                ( nn lc "People")) [ "h"]) "ARGO"
          ( ( someFact fh ".e" "FACT"
              ( gen "TMP_WHILE")
              ( control lc
                ( ( fn lc =>
                    ( pro ["*"] [ "ENTITY"] ".e" "him" ( gen "ENTITY") "at"
                      ( verb lc ".event" ["at"] "laughing"
                        ( gen "EVENT"))))
                    [ "ARGO", "at"]))) "TMP_while"
              ( past ".event"
                ( verb lc ".event" ["TMP_while", "ARGO"] "smiled"
                  ( gen "EVENT")))))
                [ "ARGO", "TMP_while"]
          [ "@e", ".event", ".e"]
```

The created complex set of operations will subsequently reduce to primitives of the SCT language as demonstrated by the following abridged output:

```
Head ("exists",["@e", ".event", ".e"],
Body ("exists",["@e", ".event", ".e"],
Clean (0, ["ARGO"], "*",
Namely (X (1, "GROUP"),".e",
Lam (".e", "ARGO",
Rel (["@e", ".event", ".e"], ["*", "*", "*"], "", [
BodyClimb (".e",
.....
Namely (X (2, "EVENT"),".event",
Rel ([], [], "asked", [At (T ("ARG1", 0), "ARG1"),
At (T ("ARGO", 0), "ARGO"),
At (T (".event", 0), "EVENT"))]),
.....
```

```

    Rel ([], [], "People", [At (T ("h", 0), "h")]])))]))))) ,
Namely (X (3, "TMP_WHILE"), ".e",
Lam (".e", "TMP_while",
    Rel (["@e", ".event", ".e"], ["*", "*", "*"], "", [
    BodyClimb (".e",
    Rel ([], [], "FACT", [At (T ("TMP_while", 0), "FACT"),
    At (
    .....
        Rel ([], [], "laughing", [At (T ("at", 0), "at"),
        At (T ("ARGO", 0), "ARGO"),
        At (T (".event", 0), "EVENT")]]))]])))]))))) ,
If (fn,
    Namely (X (6, "EVENT"), ".event",
    Rel ([], [], "smiled", [At (T ("TMP_while", 0), "TMP_while"),
    At (T ("ARGO", 0), "ARGO"), At (T (".event", 0), "EVENT")]])),
    .....
    BodyClimb (".event",
    Rel ([], [], "", [At (T (".event", 0), "h"),
    At (T ("*event", 0), "before")]]))]]))]]))]]))]])))))

```

At this point the expression carries enough information to specify the assignment function for binding names. SCT language primitives (some of which are set out below) access the content of a sequence based information state (cf. Vermeulen 1993; Dekker 2012). Some primitives refer to the content (e.g., quantifying over the initial value in a sequence). Other primitives can potentially alter the content (e.g., changing the order in a sequence, or adding or subtracting a value from a sequence). The primitives include:

- “Head” to bring about quantificational closure, like the top of a DRS (Kamp and Reyle 1993),
- “Body” to collect conditions, like the bottom of a DRS,
- “Clean” to remove bindings, like an adbmal operation (Hendriks and van Oostrom 2003),
- “Namely” like a static dynamic existential quantifier (Cresswell 2002) to place fresh bindings into the assignment,
- “Lam” an operation to transfer values between bindings,
- “BodyClimb” a structure relocating operation,
- “If” to guide the evaluation to different choice points, based on tests of the assignment state,
- “Rel” for predicate relations, and
- “At” to preserve role information (Nakashima et al. 1996).

A resulting SCT calculation checks the state of the assignment function for binding names against the positions in the discourse in which they are invoked, identifying appropriate antecedents for pronouns, long distance relationships such as control and Across the Board Extraction, etc. The binding names are sourced from the conversion of the normalised tree annotation

("ARG0", "ARG1", "at", etc.). The values assigned to these binding names appear as the sorted variables (entity, event, group, etc.) in the resulting predicate logic expressions that the calculation serves to produce. For our running examples, the following is the returned predicate logic expression:

```
exists ENTITY[4] TMP_WHILE[3] GROUP[1] EVENT[2] EVENT[5] EVENT[6].(
  past(EVENT[2])
  & past(EVENT[6])
  & asked(EVENT[2], ENTITY[John], GROUP[1])
  & People(GROUP[1])
  & ENTITY[4] = ENTITY[John]
  & is_FACT_THT(TMP_WHILE[3], (laughing(EVENT[5], GROUP[1])
    & at(EVENT[5]) = ENTITY[4]))
  & smiled(EVENT[6], GROUP[1])
  & TMP_while(EVENT[6]) = TMP_WHILE[3])
```

Such a derived analysis (with relationships expressed by sorted variables appearing in multiple contexts, or related to other variables) can be re-expressed as indices shared between nodes in a tree structure. Such indices, being derived from the source annotation in the first place, are in principle a redundant notational variant rather than an addition of information. Nevertheless, they are convenient as loci for adding information about dependencies between constituents in tree structures. The indices can be embedded back into the source phrase structure tree annotation to yield:

```
(IP-MAT
  (NP-SBJ;<GROUP[1]>
    (NS;<,GROUP[1]@h,> People)
    (IP-REL
      (NP-SBJ;<ENTITY[John]> (NPR;<,ENTITY[John]@h,> John))
      (VBD;<,GROUP[1]@ARG1,ENTITY[John]@ARG0,EVENT[2]@EVENT,> asked)
      (NP-OB1 *T*;<{GROUP[1]}>))
    (VBD;<,TMP_WHILE[3]@TMP_while,GROUP[1]@ARGO,EVENT[6]@EVENT,> smiled)
    (PP-TMP-while;<TMP_WHILE[3]>
      (P-ROLE while)
      (IP-ADV
        (NP-SBJ *PRO*;<{GROUP[1]}>
          (VAG;<,ENTITY[4]@at,GROUP[1]@ARGO,EVENT[5]@EVENT,> laughing)
          (PP-at;<ENTITY[4]>
            (P-ROLE at)
            (NP (PRO him;<{,ENTITY[John],})))))
        (PU .))
```

This "indexed" view gives a view of the tree structure with indexing information that specifies argument relationships and antecedence relationships (including big PRO, as the terminal *PRO*;<{GROUP[1]}>). This gives explicit indexing of grammatical dependencies that the original annotation had left implicit. The indexing makes the following kinds of contributions:

- Indexing given the form “<variable>” marks a node that serves as an argument for a predicate.
- The arguments that a predicate takes are marked on the pre-terminal node for the predicate with a “<, ..., variable@role, ...>” format, with “variable” providing information to locate the argument and “role” stating the argument role.
- Control and trace information is presented with the format “{variable}”, that is, specifying a single obligatory antecedent.
- Pronominal information is presented with the format “{,variable,...,}”, that is, specifying potentially multiple antecedents.

With the targets for dependencies spelled out in the predicate nodes, this is now a basis for deriving as output the kinds of formatted annotation seen in Sects. 2 and 3. This will especially be the case, e.g., for obtaining the FrameNet style of annotation, given that the unindexed source annotation is supplemented by inserting lexical unit (“LU”) and role information, here, offset from the predicate node in a “FRAME” node, as follows:

```
(IP-MAT (NP-SBJ (NS People)
  (IP-REL (NP-SBJ (PRO I))
    (VBD asked)
    (FRAME (LU *8421*)
      (ARG1 *Addressee*)
      (ARG0 *Speaker*))
    (NP-OB1 *T*)))
  (VBD smiled)
  (FRAME (LU *9738*)
    (TMP_while *Time*)
    (ARGO *Agent*))
  (PP-TMP (P-ROLE while)
    (IP-ADV (VAG laughing)
      (FRAME (LU *8720*)
        (ARGO *Sound_source*)
        (at *Addressee*))
      (PP (P-ROLE at)
        (NP (PRO him))))))
  (PU .))
```

The offset notation contains hooks that link up with the predicate argument information derived from the semantic representation, to give the same content as is seen, for example, with the FrameNet analysis in Sect. 3.

5 Summary

To sum up, this paper introduced annotation of the PropBank and FrameNet projects, noting how each is dependent on human supplemented indexing. The proposal was a technique of annotation that shifts the role of indexing onto a

principled assignment of structural positions in a syntactic tree. These principles are meshed with an interpretive process that creates the specifications of dependencies. As many of the dependencies in natural language can be robustly established by imposing some simple universal constraints on the assignment of grammatical structure, source annotation is greatly simplified. Annotators assign structure in predefined ways on the basis of intuitions about argumenthood, modification, and identity between elements. The semantic calculation uses the same definitions to confirm the analysis from a very different perspective, where predicates introduce events, nominal expressions introduce entities, and grammaticality is recast as the unambiguous management of information about all the elements in a given discourse. Language specific characteristics reveal themselves in the source annotation to the extent that a faithful representation of meaning requires divergence from the default constraints used in the semantic calculation. Normalisation adjusts the information encoded in a specific language to the forms that feed the calculation. These defaults have been developed in the process of evolving a system that accounts for the dependencies in English¹, Contemporary Japanese², and Old Japanese³.

Enriching syntactic trees with lexical semantic annotation in the manner of PropBank and FrameNet is one way to bring coherence to structure. Structural analyses of texts are useless as representations of natural language if they are not based on interpretations of meaning. In this sense, while meanings are in a sense independent of structure, they are determinative of structure depending on the lexical resources available. One of the challenges of formal linguistics is to articulate this interplay of structure and meaning. Building the necessary components into a model for corpus annotation is a natural extension of the trajectory where descriptive and theoretical linguistics converge. What we have demonstrated in this paper is an application of theory to facilitate this kind of data-driven research.

References

- Bies, A., Ferguson, M., Katz, K., MacIntyre, R.: Bracketing guidelines for Treebank II style Penn Treebank project. Technical report MS-CIS-95-06, LINC LAB 281, University of Pennsylvania Computer and Information Science Department (1995)
- Bonial, C., Babko-Malaya, O., Choi, J.D., Hwang, J., Palmer, M.: PropBank Annotation Guidelines, 3rd edn. Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder (2010)
- Butler, A.: *Linguistic Expressions and Semantic Processing: A Practical Approach*. Springer, Heidelberg (2015). <https://doi.org/10.1007/978-3-319-18830-0>
- Cresswell, M.J.: Static semantics for dynamic discourse. *Linguist. Philos.* **25**, 545–571 (2002)
- Dekker, P.: *Dynamic Semantics*. *Studies in Linguistics and Philosophy*, vol. 91. Springer, Dordrecht (2012). <https://doi.org/10.1007/978-94-007-4869-9>

¹ <http://www.compling.jp/ajb129/tspc.html>.

² http://npcmj.ninjal.ac.jp/interfaces/index_en.html.

³ <http://www.compling.jp/m97pc>.

- Hendriks, D., van Oostrom, V.: *Adbmal-calculus*. Utrecht University, Department of Philosophy (2003)
- Kamp, H., Reyle, U.: *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht (1993)
- Nakashima, H., Noda, I., Handa, K.: *Organic programming language GAEA for multi-agents*. In: *Proceedings of the Second International Conference on Multiagent Systems (ICMAS-96)*. AAAI (1996)
- Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Baker, C.F., Schefczyk, J.: *FrameNet II: extended theory and practice*, Berkeley (2016)
- Vermeulen, C.F.M.: *Sequence semantics for dynamic predicate logic*. *J. Log. Lang. Inf.* **2**, 217–254 (1993)



Discontinuity in Potential Sentences in Japanese

Hiroaki Nakamura^(✉)

Institute for the Promotion of Higher Education,
Hirosaki University, Bunkyo-cho 1, Hirosaki-shi 036-8560, Japan
hnakamura@juno.ocn.ne.jp
<http://www.hirosaki-u.ac.jp/>

Abstract. Japanese potential sentences have long been receiving considerable attention, especially with respect to the case alternation phenomena of object noun phrases. We suggest that the fundamental claims that have been assumed in Japanese linguistics are wrong or insufficient both from empirical and theoretical points of view. The points we argue in this paper are: (1) the Japanese linguists have paid too much attention on the case alternation with the object arguments of base verbs in potential sentences, trying to explain potential sentences in comparison with passive sentences formed with the same suffix *-rare*. In this paper, we will see some important counterexamples, and suggest that the two uses of the suffix must be distinguished syntactically and semantically at least in the present-day Japanese. (2) Then, we will present an array of new data and propose a completely new analysis, adopting a version of categorial/type-logical grammar framework to deal with discontinuity frequently found in potential sentences. Finally, we will see some consequences of our analysis, focusing on the relationship between case alternations and scope alternations between quantified NPs and the suffix *rare*. We wrap up this paper with considering some solution to remaining problems of our analysis.

Keywords: Japanese potential constructions
Discontinuous constituency · Type-logical grammar · Focus particle
Generic operator

1 Introduction

Japanese traditional and generative linguistics have long paid much attention on the case alternation phenomena found in potential constructions. See Inoue [11]

My work on this paper has benefited tremendously from our lively discussions over the past several years with Kei Yoshimoto and Yoshiki Mori. I am grateful to the participants in LENLS 14, who have responded critically and fruitfully to my original ideas, in particular, Daisuke Bekki, Chungmin Lee, Alastair Butler and Koji Mineshima. Remaining errors are my sole responsibility.

for a summary of case assignment in Japanese generative grammar. In this paper we will argue that there is good reason to doubt that their approaches correctly have explained the syntactic and semantic properties of these constructions. We will analyze fuller linguistic coverage and capture important syntactic and semantic consequences of the case alternation in these constructions from a new point of view.

Let us observe a typical example of the case alternation the linguists have been interested in so far.

- (1) Taroo-ga eigo-o/-ga hanas-(rar)e-ru.
 Taroo-Nom English-Acc/-Nom speak-CAN-Pres
 ‘Taroo can speak English.’

Japanese generative grammar has called the object marked with nominative in (1) ‘nominative object’ and tried to explain the *ga/o*-alternation in (1) mainly in terms of the Case-checking/licensing mechanism.¹ Japanese traditional grammar has examined the same phenomena mainly in terms of ‘voice’ because it has assumed that the potential suffix and passive suffix originates from the same auxiliary verb *-rare* (namely, they have been treated as two different uses of the one and the same suffix).

In the next section, we will review some of the past literature on potential constructions both in the traditional and generative linguistics, pointing out major drawbacks in their analyses. Then, we will argue that the passive and potential uses of the suffix *-rare* should be analyzed separately at least in modern Japanese. The passive suffix combines with base verbs in the lexicon and causes the changes in their argument structures (the demotion of agent arguments and promotion/externalization of theme arguments). On the other hand, we show that no such argument structure change occurs in the complex verbs comprising base verbs and the potential suffix, adducing evidence from subject honorification examples (see Hasegawa [9] for an overview of honorification in syntax).

Here, let us attend to the fact that a wide variety of arguments of base verbs can be marked nominative (actually, subjectivized) in potential sentences, as illustrated in (2).

- (2) a. Kono beddo-de/-ga yoku nemur-e-ru.
 This bed -On/Nom well sleep-CAN-Pres.
 ‘I can sleep well on this bed.’
 b. Kono fude-de/ga kireina ji-o/-ga kak-e-ru.
 this ink-blush-With/Nom beautiful letter-Acc/-Nom write-CAN-Pres
 ‘I can write beautiful characters with this ink blush.’

¹ The voices in Japanese are sometimes classified as active, passive, potential and spontaneous voices (Teramura [22]), and some grammarians include causative as a kind of voice.

Potential sentences with the locative and instrumental arguments of the base verbs marked with nominative in (2) are quite common in ordinary Japanese.² It should be noted that ‘-ga/-no’ conversion is also possible when they are nominalized with the nominalization suffix *-koto*. In the sentences without *-rare* ‘can’, these arguments can never be marked with genitive even when followed by *-koto*.

- (3) a. kono beddo-de/no yoku nemur-e-ru-koto
 this bed -On/Gen well sleep-CAN-Pres-Fact
 cf. *kono-beddo-no yoku nemur-u-koto’
 b. kono fude-de/no kireina ji-o/-ga kak-e-ru.
 this ink-blush-With/Gen beautiful letter-Acc/-Nom write-CAN-Pres
 cf. *kono-fude-no kireina ji-no/-o kak-u-koto

Clearly the case alternations between locative/instrumental and nominative/genitive cases cannot be accounted for in terms of the notion of voice. We will show that a wide variety of case alternations can be elegantly explained in our categorial/type-logical grammar formalism in which the syntax and semantics work in tandem, providing a well-formed expression and its model-theoretic interpretation at each step of the process of structure building.

Finally, we will consider the difference in scope between noun phrases marked with nominative and other cases and suggest that the logical approach adopted here to derive a variety of discontinuous constituents can provide all potential sentences with proper semantic interpretations in an explicit manner.

2 A Brief Review of the Literature on Potential Sentences

Let us start our discussions with reviewing some descriptive analyses of the potential sentences. It may look strange that the same suffix has the four apparently different usages that have long assumed in Japanese traditional grammar, such as passive, potential, honorific, and spontaneous uses. Cho [6] adopts the view that the original use of the suffix *-rare* was spontaneous, citing a lot of interesting examples. Though we do not discuss the historical development of the different uses of *rare*, we agree with the descriptive linguistics view that potential sentences mean the abilities of agents of base verbs, or possibilities of the states that such actions are realized. The typical meanings of the potential sentences are sometimes grouped into the following three types (Aoki [1] from Cho [6]). They denote

- (4) a. inherent or enduring capabilities of the agents to do the actions denoted
 by base verbs
 b. possibilities or values of the targets of the actions denoted by base verbs.
 c. temporal possibilities of the agents to do the actions denoted by base verbs

² On the other hand, any argument other than the theme argument may never be marked nominative in passive sentences containing the etymologically same suffix *-rare*.

The potential sentences, however, can mean the possibilities/capabilities of arguments other than agents or themes of base verbs, as we saw in sentences (2), which should not enter into any of these types. What we do not accept are simple analyses of potential sentences like Teramura [22] based on just the two kinds of case arrays, as in (5):

- (5) a. X-ga Y-o V-(rar)e-ru (N-Nom Y-Acc)
 b. X-ni/-ga Y-ga V-(rar)e-ru (X-Dat/-Nom Y-Nom).

In both (5a) and (5b), X has an ability or (it is possible for X) to do the action denoted by the V, whereas Y will be affected by the action. Teramura suggests that potential sentences mean that it is possible for X to do some action V to affect Y, or that X carries the potential to do V to affect Y. Based on the semantic roles of X and Y in (5), Teramura divided potential sentences into the *active* and *passive potentials* (Teramura [22], 259), giving the examples illustrated in (6):

- (6) a. Kono sakana-wa ki-ni nobor-e-ru
 This fish-Top tree-To climb-CAN-Pres
 ‘This fish can climb up a tree.’
 b. Kono sakana-wa taber-rare-ru .
 This fish-Top eat-CAN-Pres
 ‘This fish is edible.’

His view has boosted the analyses of potential sentences in terms of voice thereafter, but we will argue that this voice-oriented view of case alternations in these sentences is simply wrong because sentences like (2) with oblique arguments marked nominative are quite common in Japanese and cannot be explained by the notion of voice. We find that even the possessor argument of any argument of a base verb can easily be subjectivized, as shown in (7).

- (7) Kono naifu-ga t sentan-de enpitu-ga kezur-(rar)e-ru.
 this knife-Nom pencil-Nom sharpen-CAN-Pres
 ‘This knife enables you to sharpen a pencil with its tip.’

Can sentences like (2) or (7) be classified as an active or passive potential? The subjects in (2) and (7) are not agents or targets of the actions meant by the base verbs, but have the properties of enabling someone to realize the actions of base verbs. (7) denotes an inherent property of this knife which makes it possible for someone to sharpen a pencil with its tip in some imaginary situation(s). It is obvious that any analysis from the viewpoint of voice misses important properties these constructions have. We adopt the traditional grammarians’ view that the base verbs should be action verbs with (often implicit) agents (which are usually suppressed in our analysis) (also see Bekki [2] and Iida [10] for formal-semantic analyses of Japanese complex potential verbs), and reject their passive-active dichotomy of these sentences.

Incidentally, considering a completely different piece of evidence to differentiate the passive and potential uses of the suffix *-rare* is in order here. Subject honorification in Japanese is characterized as targeting subjects, referring to the

individuals to be worthy of respect, and indicated by wrapping base verbs with the discontinuous honorific morpheme *o/-ninar*, as illustrated in (8).

- (8) a. Sensei-ga gakusei-o o-sikari-ni-natta.
 teacher-Nom student-Acc Hon-scold-Hon-Past
 ‘The teacher scolded a student’.
- b. Sensei-ga gakusei-ni o-sikar-are-ni-natta.
 teacher-Nom student-By Hon-scold-Pass-Hon-Past
 ‘The teacher was blamed by a student.’

In passive (8b), the derived subject *sensei-ga* is referred to as a person who the speaker should show respect to. On the other hand, regardless of the semantic roles of the subjects for base verbs, the base verbs are first wrapped by the discontinuous honorific form *o/-ninar* and, the derived complex verbs are followed by the potential suffix *-rare*, as in (8a). Subject honorification can always occur in potential constructions and marks the agents of base verbs as the person to show respect to, as illustrated in (9).

- (9) a. Sensei-ni/ga gakusei-o/-ga o-sikari-ni-nar-e-ru.
 teacher-Dat/-Nom student-Acc/-Nom Hon-scold-Hon-CAN-Pres
- b. Kono beddo-de/-ga yoku o-nemur-ni-nar-e-ru.
 This bed-On/-Nom well Hon-sleep-Hon-CAN-Pres
 ‘You can sleep on this bed.’
- c. Kono fude-de/-ga kireina ji-o/ji-ga
 This ink-blush-With/-Nom beautiful letter-Acc/-Nom
 o-kaki-ni-nar-e-ru.
 Hon-write-Hon-CAN-Pres
 ‘You can write beautiful letters with this ink-blush.’

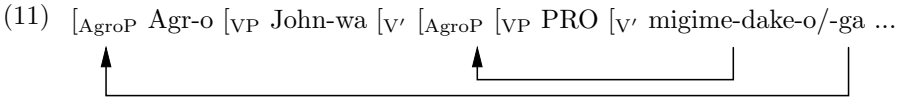
The difference in the order of morphemes between the passive and potential complex predicates is clear. In passives, the honorific morpheme must wrap the derived passive complex verb, whereas, in potentials, it must wrap the base verbs first, and then, the potential suffix is attached to the honorific complex form. We may say, incidentally, that subject honorification actually marks the external arguments of base verbs which it wraps, not the sentential subjects. This is another piece of evidence which suggests we should distinguish the two uses of *-rare* at least in current Japanese.

Japanese generative grammar has also approached to potential constructions in rather different ways. They do not admit the active-passive dichotomy, but have studied the interesting correlations between case and scope alternations. Among others, Tada [20] noticed the fact that there is an astonishing difference in scope interpretations between accusative and nominative objects. Observe the sentences in (10)³:

³ Tada suggests that the accusative object does not have a wide scope reading, while several linguists admit scope ambiguity with it. I agree with the latter’s judgement, but we do not pursue the scope alternation with the accusative object here.

- (10) a. John-wa migime-dake-o tsumur-e-ru.
 John-Top right-eye-only-Acc close-CAN-Pres
 (i) can < only (John can wink his right eye.)
 (ii)?* only > can (It is only his right eye that he can close.)
- b. John-wa migime-dake-ga tsumur-e-ru.
 John-Top right-eye-only-Nom close-CAN-Pres
 (i)* can < only
 (ii) only > can

Tada, assuming a biclausal structure with the suffix *-rare* and the base verb projecting the matrix and embedded clauses respectively, suggests that the case feature of the accusative object is checked by the lower Agr-o, whereas the nominative object moves out of the scope domain of *-rare* and its case is checked by the higher Agr-o. The different Case checking for the two kinds of objects seems to account for the difference in scope.



This account was criticized by Takano [21] and other generative grammarians because it cannot explain why the nominative *dake*-NP does not have a narrow scope reading. Case-checking movement usually induces reconstruction effects regarding scope interpretation, so “can > only” interpretation should be possible in (8b). To explain the lack of reconstruction effects with the nominative objects in potential sentences, Saito and Hoshi [18] and Takano [21], among others, posit different (LF) structures for the sentences with the accusative and nominative objects in (8). The former proposes a mono-clausal structure derived by HEAD-HEAD MERGER, and the latter suggests that the nominative object is base-generated in the matrix clause. We do not examine these solutions further here because any account in generative grammar cannot explain the case and scope alternations we saw in (2) and (7). When these oblique arguments are followed by *dake* ‘only’ and marked nominative, they must scope over the potential suffix *-rare*, whereas those marked with their original cases can/must take a narrow scope in a similar way.

- (12) a. Kono beddo-dake-de/-ga yoku nemur-e-ru.
 This bed-Only-On/-Nom well sleep-CAN-Pres
 Locative: can < only, ?only > can
 Nominative: only > can, *can > only
- b. Kono fude-dake-de/-ga kireina ji-o/-ga kak-e-ru.
 this ink-blush-Only-With/-Nom beautiful letter-Acc/-Nom write-CAN-Pres
 Instrumental: can < only, ?only > can
 Nominative: only > can, *can > only

The difference in the accounts for the lack of narrow scope reading with nominative NPs is highly theoretical and stipulative, and cannot give a proper explanation to the same effects of the case alternations between oblique and nominative NPs, as illustrated in (12). From the beginning, the oblique NPs cannot move up or be licensed in the same way as the nominative objects.

3 Discontinuity in Type-Logical/Categorial Grammar

In this paper, we adopt the multi-modal approach developed by Morrill [14–16] to analyze the derivations and interpretations of potential constructions. We assume here that the readers are familiar with the basic notions of type-logical grammar, and refer to Moortgat [13], Carpenter [4], and Morrill [15,16].

Our framework includes two extra operators in addition to normal directional implications (divisions) ‘\’ and ‘/’ to deal with discontinuous constituents in potential constructions: infixing and wrapping operators. $B\uparrow A$ stands for a functor that wraps around an A argument to form an expression of category B . $A\downarrow B$ stands for a functor that infix itself in its A argument to form an expression of category B . These operators enables us to deal with discontinuous constituents we find in potential sentences. The idea of wrapping/infixing operations has been extensively studied in categorial grammars, and shares basic ideas with the notion of continuation in Barker and Shan [3]. Morrill shows that his treatment of discontinuity can be applied to a wide variety of phenomena including quantification, anaphora binding, relativization and pied-piping. Incidentally, discontinuous constituency has also been much studied in Combinatory Categorial Grammar framework using function composition as a primitive (see Steedman [19] among others), but the effects of function composition can be obtained with combinations of elimination and introduction rules in Type-logical grammar. We will show a linguistic expression as a triplet of the form $\langle \text{prosodic form, syntactic category, meaning} \rangle$. We show the split-wrap interaction as in (13):

$$(13) s1 + s2 + s3 = (s1, s3) W s2$$

To manipulate discontinuity, we posit the following “structural rules” (Morrill [15], 195).

$$(14) \frac{\begin{array}{c} \vdots \\ \delta((\alpha, \beta) W \gamma) \end{array}}{\delta(\alpha + \gamma + \beta)} \text{WN} \qquad \frac{\begin{array}{c} \vdots \\ \delta(\alpha + \gamma + \beta) \end{array}}{\delta((\alpha, \beta) W \gamma)} \text{WN}$$

$$(15) \frac{\begin{array}{c} \vdots \\ B(\alpha) \end{array}}{B(\alpha + \varepsilon)} \qquad \frac{\begin{array}{c} \vdots \\ B(\alpha) \end{array}}{B(\alpha + \varepsilon)} \qquad \frac{\begin{array}{c} \vdots \\ B(\alpha + \varepsilon) \end{array}}{B(\alpha)} \qquad \frac{\begin{array}{c} \vdots \\ B(\alpha + \varepsilon) \end{array}}{B(\alpha)}$$

Here ε stands for an empty string. The logical rules involving the wrapping and infixing operators are shown in Prawitz-style, as in (16) and (17):

$$(16) \begin{array}{ccc} \downarrow \text{Elimination rule (Infixation)} & & \downarrow \text{Introduction rule} \\ \alpha - A : \phi \quad \gamma - A \downarrow B : \chi & & [\alpha - A : x]^{\downarrow} \\ \hline & \downarrow E & \vdots \end{array}$$

$$\begin{array}{c}
 \alpha W\gamma - \chi\phi \\
 \hline
 \gamma - C \uparrow B : \chi \quad \beta - B : \psi \\
 \hline
 \gamma W\beta - C : \chi\psi \quad \uparrow E
 \end{array}
 \qquad
 \begin{array}{c}
 \alpha W\gamma - B : \chi \\
 \hline
 \gamma - A \downarrow B : \lambda x.\chi \quad \downarrow I^i \\
 \vdots \\
 \alpha W\gamma - B : \chi \\
 \hline
 \gamma - C \uparrow B : \lambda y.\chi \quad \uparrow I^j
 \end{array}$$

(17) \uparrow Elimination rule \uparrow Introduction rule (Extraction)

In Morrill’s work, Fitch-style deduction is often used, so we introduce the logical rules in this style, as in (18) and (19).

$$\begin{array}{c}
 (18) \quad \begin{array}{l}
 \text{n. } \alpha - A : \phi \\
 \text{m. } \gamma - A \downarrow B : \chi \\
 \hline
 (\alpha W\gamma) - B : (\chi\phi) \quad \text{E}\downarrow\text{n, m}
 \end{array} \\
 \\
 \begin{array}{l}
 \text{n. } \alpha - A : \phi \quad \text{H} \\
 \hline
 \text{m. } \begin{array}{l}
 (\alpha W\gamma) - B : (\chi\phi) \quad \text{unique } a \text{ as indicated} \\
 \gamma - A \downarrow B : \lambda x\phi \quad \text{I}\downarrow\text{n, m}
 \end{array}
 \end{array}
 \end{array}$$

$$\begin{array}{c}
 (19) \quad \begin{array}{l}
 \text{n. } \alpha - A : \phi \\
 \text{m. } \gamma - B \uparrow A : \chi \\
 \hline
 (\gamma W\alpha) - (\chi\phi) : B \quad \text{E}\uparrow\text{n, m}
 \end{array} \\
 \\
 \begin{array}{l}
 \text{n. } \alpha - A : \phi \quad \text{H} \\
 \hline
 \text{m. } \begin{array}{l}
 (\alpha W\gamma) - B : \phi \quad \text{unique } a \text{ as indicated} \\
 \gamma - B \uparrow A : \lambda x\phi \quad \text{I}\uparrow\text{n, m}
 \end{array}
 \end{array}
 \end{array}$$

We will frequently use these functors and rules later to explain scope interactions between quantified NPs and the modal suffix in potential sentences. The infix operator can be used, for example, to derive the complex form of a base verb wrapped by the discontinuous honorific expression, which is then followed by the potential suffix, as shown in (20).

- (20) 1. (o, ninar) - V \uparrow V: Hon
 2. hanas - (V \uparrow V) \downarrow V: speak
 3. (o,-ninar) W hanasi - V: speak(\underline{x} , ...) 1, 2, E \downarrow
 4. o-hanasi-ni-nar - V: speak(\underline{x} , y) = 3

The \downarrow Elimination rule enables the base verb *hanasi* to infix itself in the discontinuous honorific form (*o, -ninar*) of category V \uparrow V, which marks the referent of its external argument as a person worthy of respect from the speaker’s

viewpoint. In this case, the base verb is wrapped but is still a higher functor. The discontinuous honorific form does not do anything but to mark the external argument of a base verb as the target of respect.

4 Derivations of Potential Sentences

In this section, let us consider the derivations and interpretations of potential constructions with a wide variety of arguments of base verbs marked with nominative case. Before showing that these oblique arguments are in fact subjectivized in these sentences, let me point out some properties of the potential suffix *-rare*. As in Japanese generative grammar (see Takano [21] and references cited there), we assume that the complex potential verbs projects a biclausal structure, which is a kind of control structure and the argument structure of a base verb will not be changed by this suffix (case absorption and promotion and/or demotion of arguments are not induced by the potential *-rare*). We do not consider the interpretation of the modal operator in detail, which is assumed here to be simply an operator that introduce quantification over the possible worlds. *Rare* XP is interpreted as $[[XP - rare]] = \exists w' \in W. [[XP]]_{M,g,w}$. Here the set W is defined to be circumstantial and/or dispositional in the potential sentences (see Portner [17] for the detailed discussions about modality).

$$(21) \textit{-rare} - (N \uparrow S) \setminus S : \lambda P. \lambda x. \diamond Px$$

If there is no overt experiencer in the matrix clause, the agent argument of a base verb is interpreted to denote an arbitrary individual (*pro* in generative grammar). As I suggested before, any argument of a base verb can show up as a subject in a potential sentence. We also saw that all potential sentences have one and the same structure (no active/passive distinction). We follow Iida [10] in that the subject of potential sentences has a kind of modality (generic) inherently, and we express this property as an inherent generalized quantifier which scope over the whole predicate comprising a base verb and the suffix *-rare*, and it undergoes “quantifying-in” according to the lexical category defined by the wrapping and/or infixing operators (see Krifka et al. [12] for an overview of genericity). This category of the subject makes it possible to take the whole remaining predicate as scope, and no scope ambiguity (namely, no reconstruction effect) occurs with the nominative noun phrases in potential sentences.

Let us take a simplified version of (7) as an example in which the oblique argument of the base verb *kezur* ‘sharpen’ can be subjectivized. Compare.

- (22) a. Kono naifu-de enpitsu-o kezur-(rar)e-ru.
 This knife-With pencil-Acc sharpen-CAN-Pres
 ‘You can sharpen pencils with this knife.’
 b. Kono naifu-ga enpitsu-o kezur-(rar)e-ru.
 this ink-blush-Nom pencils-Acc sharpen-CAN-Pres
 ‘This pencil enables you to sharpen pencils with it.’

(22a) sounds a bit strange because at least one subject (nominative NP) is needed in a Japanese sentence. Potential sentences state that the referents of subject NPs have more or less permanent properties (inherent/enduring abilities or possibilities). On the other hand, (22b) sounds quite natural, predicating the inherent property of the subject *naifu* ‘knife’, which enables an arbitrary person to sharpen pencils with it. In addition to the instrumental argument, the theme argument *enpitsu* ‘pencil’ of the base verb can also be subjectivized (in Japanese, nominative subjects can multiply appear in a sentence in which open propositions comprising the remaining parts of the sentence can be predicated of the corresponding subjects, forming layers of predication). The derivation of (22a) and (22b) can be shown as in (23a) and (23b) respectively:

- (23) a. Kono naifu-de enpitsu-o kezur (rar)e-ru.

$$\begin{array}{ccccccc} N_{Inst}: & N_{Th}: & N_{Th} \setminus (N_{Inst} \setminus (N_{pro} \setminus V)): & V \setminus (N) \setminus S: & & & \\ \text{with-this-knife}' & \text{pencil}' & \text{Sharpen}' & \diamond P(\text{pro}) & & & \\ \hline & & & & & & /E \\ & & & & & & N \setminus V: \text{Sharpen}'(\text{pro}, \text{pencil}') \\ \hline & & & & & & /E \\ & & & & & & V: \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-knife}') \\ \hline & & & & & & /E \\ & & & & & & S: \diamond \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-knife}') \end{array}$$
- b. Kono naifu-ga [a]¹ enpitsu-o kezur (rar)e-ru.

$$\begin{array}{ccccccc} N_{Nom}: & N_{Inst}:x & N_{Th} & N_{Th} \setminus (N_{Inst} \setminus (N_{pro} \setminus V)): & V \setminus (N) \setminus S: & & \\ \text{this-knife}' & & \text{pencil}' & \text{Sharpen}' & \diamond P(\text{pro}) & & \\ \hline & & & & & & /E \\ & & & & & & V: \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-a}') \\ \hline & & & & & & /E \\ & & & & & & S: \diamond \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-a}') \\ \hline & & & & & & \uparrow I \\ (S \uparrow N) \downarrow S: & & & & & & \\ \text{Gen } y(\text{Knife}'(y) \rightarrow P(y)) & & & S \uparrow N: \lambda x(\diamond \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-x})) & & & \\ \hline & & & & & & \\ & & & & & & S: \text{Gen } x(\text{Knife}'(x) \rightarrow \diamond \text{Sharpen}'(\text{pro}, \text{pencil}', \text{with-x})) \end{array}$$

The nominative instrumental argument must scope over the corresponding predicates due to the definition of their syntactic categories and semantic types. We will explore the use of the generic (Gen) operator and its quasi-universal property further in the final section. The derivation shows that we cannot expect ‘reconstruction effects’ to occur with respect to any argument of the base verb if it is marked nominative. We can multiply apply the \uparrow introduction rule if we have made more than hypotheses in advance and discharge them in the derivation. In the next section, let us make use of the lack of reconstruction effects to explain scope interactions we saw in (10) and (12).

5 Lack of the Reconstruction Effects and Scope Interactions

As we have seen in Sect. 2, Tada's [20] explanation of scope interactions between the nominative and accusative objects and the potential suffix *-rare* fails to account for the lack of reconstruction effects with the nominative objects. Other generative grammarians (see Takano [21] and references cited there) try to explain it by positing the completely different (surface/LF) structures for potential sentences with accusative objects and those with nominative objects. All of these explanations miss a very important generalization about subjectivization in these constructions. As we have seen so far, any argument when it is marked nominative is in sharp contrast in scope interpretation to the argument marked with its original. Observe the following dialogues in (24) and (25):

- (24) a. 101-kyositu-dake-de 200-nin-ga hair-e-ru.
 101-classroom-Only-In 200-people-Nom enter-CAN-Pres
 '200 people can enter only in Classroom 101.'
- b. 101-kyositu-dake-ga 200-nin-ga hair-e-ru.
 101-classroom-Only-Nom 200-people-Nom enter-CAN-Pres
 'Only Classroom 101 can accommodate 200 people.'
- (25) Iiya, 201-kyositu-mo hair-(rar)e-ru.
 No Classroom-201-Also enter-CAN-Pres
 'No. Classroom 201 can also accommodate (200 people).'

(24a) just says that Room 101 can accommodate 200 people at a time and does not mention anything about other rooms. If someone suggests (25), there is no contradiction between the two statements. (24b) states, however, that only Room 101 can accommodate 200 people, and no other room can, so (24b) is inconsistent with (25b). We can easily account for the (non-)contradiction between (24a, b) and (25) by using our definition of the subjects and predicates in potential sentences.

Our treatment of *only*-quantifiers is quite usual. We simply define the meaning of *101-kyositu-dake* as in (26):

- (26) $\lambda P. \exists x(x = \text{Room}101 \wedge \neg \exists y(y \neq x \wedge Py))$

The noun phrase followed by the postposition *-de* 'in' cannot combine with the matrix suffix *-rare*, so *101-kyositu-dake-de* 'only in Classroom 101' must take narrower scope with respect to CAN in (24a), whereas the same NP marked nominative must scope over the matrix *-rare* according to our definition of the subjects in potential sentences (and probably, in stative sentences in general)

with the infixing and wrapping operators in (24b).⁴ The derivations and interpretations can be shown in parallel, as in (27a) and (27b):

- (27) a. 101-kyositu-dake-de 200-nin(-ga) [a]¹ hair -(rar)e-ru.
 N_{Loc} : N_{Sb} N_{Loc} $N_{Loc} \setminus (N_{Sb} \setminus S_{Inf})$: $S_{Inf} \setminus S$:
 Only-In-Room101 200-people(-Nom) In-a Enter'- $\diamond P$
 $\frac{\quad}{\quad} \setminus E$
 $\frac{S_{Inf} / (N_{Loc} \setminus S_{Inf}) : \lambda P(P(In - R100) \wedge \exists y(y \neq R101 \wedge P(In-y))) \quad S_{Inf} : (Enter'(200, In-a))}{S_{Inf} \setminus N_{Loc} : \lambda x(Enter'(200, In-x))} \setminus I^1$
 $\frac{S_{Inf} : Enter'(200, In-R101) \wedge \exists y(y \neq 101 \wedge Enter'(200, In-y))}{S : \diamond(Enter'(200, In-R101) \wedge \exists y(y \neq 101 \wedge Enter'(200, In-y)))}$
- (27) b. 101-kyositu-dake-ge 200-nin(-ga) [a]¹ hair -(rar)e-ru.
 N_{Nom} : N_{Sb} N_{Loc} $N_{Loc} \setminus (N_{Sb} \setminus S_{Inf})$: $S_{Inf} \setminus S$:
 Only-In-Room101 200-people(-Nom) In-a Enter'- $\diamond P$
 $\frac{\quad}{\quad} \setminus E$
 $\frac{(S \uparrow N) \downarrow S : \lambda P(P(R100) \wedge \exists y(y \neq R101 \wedge P(y))) \quad S_{Inf} : (Enter'(200, In-a))}{S : \diamond Enter'(200, In-a)} \setminus E$
 $\frac{S : \diamond Enter'(200, In-a)}{S \uparrow N : \lambda x(\diamond Enter'(200, In-x))} \uparrow I^1$
 $\frac{S : (\diamond Enter'(200, In-R101) \wedge \exists y(y \neq 101 \wedge \diamond Enter'(200, In-y)))}{S : (\diamond Enter'(200, In-R101) \wedge \exists y(y \neq 101 \wedge \diamond Enter'(200, In-y)))}$

Though we should include more information on genericity inherent within the subject NPs, we focus on the scope interpretation of *dake*-NP here for the sake of simplicity. It is easy to see that the interpretation derived in (27b) contradicts the meaning of (25), whereas there are assignments which makes both of (27a) and (25) true. Based on our definition of the category for the subjects, nominative NPs have no choice but to take scope over *-rare* or whole predicates comprising base verbs, the potential suffix, negation and tense. This is the best advantage of our analysis of subjectivization as cases of quantifying-in, which do not differentiate gaps corresponding to arguments/adjuncts of base verbs.

⁴ Not all stative complex predicates in Japanese do not allow the case alternations exemplified here. For instance, complex predicates comprising base verbs and the desiderative suffix *-tai* 'want' allows objects to be marked nominative or accusative, but does never allow adjunct arguments to be marked nominative.

- (i) a. Gaikoku-he-iku-to daremo-ga nihon-syoku-o/-ga tabe-tai.
 foreign-country-to-go-when anybody-Nom Japanese-dishes-Acc/-Nom eat-WANT
 'When they go abroad, all Japanese want to eat Japanese dishes.'
 b. Gaikoku-he-iku-to daremo-ga tatami-no-ue-de/*-ga ne-tai.
 foreign-country-to-go-when anybody-Nom straw-mat-On/-Nom sleep-WANT
 'When they go abroad, all Japanese want to sleep on the straw mat.'

We do not pursue the differences in case alternation among complex stative predicates here.

6 Remaining Problems and Alternative Analysis

The motivation of this research was twofold: first, we wanted to show the accounts given so far for the potential sentences, whether in descriptive or generative linguistics, have focused too much on the case alternations with theme arguments (so-called nominative objects) and missed an important generalization with respect to the subjectivization phenomena in these constructions. Actually, any argument of base verbs or even its possessor argument can become the subject of the complex predicates. Second, using a version of type-logical framework developed by Morrill, among others, we hoped to explain the fact that the subjectivized arguments or adjuncts always take a wide scope with respect to the possibility operator corresponding to the suffix *-rare*. Namely, there is no reconstruction effect with nominative arguments/adjuncts, which has been discussed to reject the movement analysis of nominative objects in Japanese generative grammar (see Takano [21]). We have successfully given a proper account for the absence of reconstruction effects in terms of the special syntactic category for the subjects defined in terms of the discontinuity operators.

On the other hand, there are some problems which remain to be explained. One of them is the existence of island phenomena which will be induced by the relatively free use of the introduction rule with the discontinuity operators. Some restriction should probably be placed on uses of these operators. Another serious problem is also pointed out by a reviewer and an audience member at LENLS 14. The problem of my analysis can be shown graphically, as follows:

(28) Summary of Predication Process

Subject	Predicate	
$\alpha-(S \uparrow N) \downarrow S : \gamma$	$\dots [a-N]^i \dots N \setminus \dots S : \chi$	
	$\frac{\quad}{S : \beta(a)} \setminus E$	[1]
	Elimination using syntactic information	
	$\frac{\quad}{S \uparrow N : \lambda x. \beta(x)} \uparrow I^i$	[2]
	Introduction of the lambda-operator	
	$\frac{\quad}{\alpha + \beta-S : \chi(\gamma)} \downarrow E$	[3]

First, we use the elimination syntactically in [1] to form a proposition, which is then converted to a propositional function by the introduction of \uparrow operator in [2]. Finally, the elimination is applied semantically to let the nominative NP (a major subject) of the special higher-order functor category with \uparrow and \downarrow operators apply to the derived predicate in [3]. This complex derivation is merely to overcome ostensible case conflict between the filler and the gap in (28). (27b), for instance, is derived with the meaning ‘Room 101 can accommodate 100 people at a time, and no other room other than Room 101 can accommodate 200 people’ (27a), meaning that it is possible that 200 people are accommodated only in Room 101 at a time. Though the difference in interpretation can be

accounted for in our approach, the syntactic derivation for (27b) does not say anything about how the nominative subject can be identified with the gap of the (optional) adjunct. In other words, semantic predication in [3] cannot help to solve the case conflict in syntax.

A more serious problem lies in our description of the meanings of potential sentences. Potential sentence (27b), for instance, does not mean that it can be true if and only if Room 101 accommodates 200 people at least in a world (and time). This sentence should be taken to be a characteristic sentence describing a permanent or tendentially stable property of the entity Room 101. The proper meaning of Sentence (27b) is roughly that Room 101 accommodate 200 people at a time in general. Potential sentences in general are ‘generic’ and the potential suffix *-rare* converts episodic action verbs into stative predicates with habitual readings of the actions denoted by base verbs. They also refer to possibilities which are evaluated in worlds of utterances. We can express the meaning of (27b) using adverbs of quantification, as in ‘it is always or usually possible for Room 101 to accommodate 200 people at a time, and no other room cannot accommodate such a big number of people in the same (similar) situations.’ The property which enables 200 people to enter at a time is predicated of Room 101 as one of its tendentially stable properties. This generic interpretation is brought out when we consider that potential sentences always can cooccur with Q-adverbs (Chierchia [5]) like *usually*, *always*, *generally*, etc. Potential sentences quite easily co-occur with Q-adverbs, as shown in (29):

- (29) 101-kyositsu-ga futsuu/ itsudemo 200-nin-ga hair-(ar)e-ru.
 Room101 usually/ always Enter-CAN-Press

Chierchia [5] summarizes the properties of Q(uantificational)-adverbs, the three of which we are especially interested in here, namely Property A, D and E:

Property A: Q-adverbs can bind eventualities.

Property B: Q-adverbs can bind variables provided by indefinites.

Property C: Q-adverbs can bind variables provided by kind-denoting definites.

Property D: Q-adverbs can bind more than one variable.

Property E: Q-adverbs can (by and large) freely select the arguments they bind.

To solve our two problems, the case conflict and the semantic interpretation, let us introduce the notion of unselective binding in our analysis of potential sentences, which yields the meaning of potential sentences as in (30).

- (30) Gen x, y, s [Subject(x) \wedge C(x, s)] [$\diamond s$ P(x)]

The variable s is a situation variable over eventualities in the sense of Davidson [7]. We use the notation $\diamond s$ to denote a possibility in a situation s , which is, in turn, bound by the generic operator Gen that is introduced by adverbs of quantification or phonologically null Q-adverbs. When a Q-adverb occurs explicitly, we use its meaning to specify the interpretation of Gen (*always*, *usually*, *sometimes*, etc.). ‘C’ stands for a set of contextually specified occasions (Chierchia [5], 189). For example, we may be able to write ‘Most x, \dots, s [Subject(x) \wedge

$C(x, s) [\diamond_s P(x)]$ for the logical forms of potential sentences. We do not discuss whether we need to posit situation variables for eventualities denoted by base verbs here. It suffices at this point to assume the situation variable only for the possibility operator to mean that ‘it is always/usually/sometimes possible to do something in some world accessible from each situation,’ which expresses the meaning of potential sentences more accurately. We do not want to say that the nominative NPs (of individual-level predicates) are automatically mapped into a restrictor while other arguments into a nuclear scope, which, *prima facie*, seems correct as the partitioning of generic sentences, as suggested by Diesing [8], among others. We have already shown how the focus-ground partitioning is implemented in potential sentences with the discontinuity operators. Here we would just like to suggest that the potential sentences are not simply derived by functional application, as shown in (28)), but by unselective binding where the Gen-operator projected from adverbs of quantification or phonologically null Q-adverbs bind an arbitrary number of variables occurring both in a restrictor and a nuclear scope which are already structured by our derivations. We agree with Chierchia [5] that the splitting of generic sentences including Japanese potential sentences should be more flexible, not strictly correlational with syntactic structures assumed as in generative grammar.

We cannot pursue this problem further here due to space limitations. Let us just modify (28) by assuming the category $(S/[S \uparrow N])/[(S \uparrow N) \downarrow S]$ for Q-adverbs, which combines with a nominative NP first, and then with an open proposition including the potential suffix, as shown in (30):

(31) Summary of Predication Process

Q-adverb	Subject	Predicate	
$(S/[S \uparrow N])/[(S \uparrow N) \downarrow S]$	$\alpha - (S \uparrow N) \downarrow S : \gamma$	$\dots [a-N]^i \dots N \setminus \dots S : \chi$	$\xrightarrow{\quad} \chi \text{E} \quad [1]$
		$S : \beta(a)$	$\xrightarrow{\quad} I^i \text{E} \quad [2]$
		$S \uparrow N : \lambda x. \beta(x)$	$\xrightarrow{\quad} /E \quad [3]$
<div style="display: flex; justify-content: space-between; align-items: center;"> Q-adverb + α + β - S: Gen $x, s [x = \gamma \wedge C(x, s)] [\diamond_s \beta(x)]$ </div>			

Though we cannot show its detailed derivation, the interpretation of (27b), for instance, should be like (32):

(32) S: Gen $x, s [\diamond_s \text{Enter}'(200, \text{In-R101})] \wedge [\exists y (y \neq 101 \wedge \diamond \text{Enter}'(200, \text{In} - y))]$

Let us just abandon the last part of the derivations in (27) implemented by functional application using the *downarrow* operator. At this stage, the dyadic structure is already given. What we need is let the Gen-operator bind the situation and other variables in the both clauses, assuming the properties of Q-adverbs summarized by Chierchia, according to which Q-adverbs can be defined to take two open clauses. If they do not appear explicitly, we posit phonologically null Q-adverbs in the same position though their positions are not important because we can deal with any discontinuous constituency quite easily, as illustrated in

Sects. 4 and 5.⁵ Q-adverbs syntactically takes two constituents with different discontinuity operators (resulting in a dyadic structure comprising potentially discontinuous constituents), one with the ↓ operator and the other with ↑ operator, which introduce variables corresponding to the filler (nominative NP) and its gap to be unselectively bound by the Gen-operator.

References

1. Aoki, R.: *Kanoo-Hyogen*. In: Kokugo-Gakkai (ed.) *Kokugo-gaku Daijiten*. Tokyo-Dō, Tokyo (1980)
2. Bekki, D.: *Nihongo-Bunpo-no Keishiki-Riron: Katsuyo-Taikei-Tougo-Kouzo-Imi-Gousei*. Kuroshio Publishers, Tokyo (2010)
3. Barker, C., Shan, C.: *Continuations and Natural Language*. Oxford University Press, New York (2015)
4. Carpenter, B.: *Type-Logical Semantics*. MIT Press, New York (1997)
5. Chierchia, G.: Individual-level predicates as inherent generics. In: Carlson, G., Pelletier, F.J. (eds.) *The Generic Book*, pp. 1–123. The University of Chicago Press, Chicago (1995)
6. Cho, I.: *Kekka-Kanou-Hyogen-no Kenkyu-Nihongo/Chugoku-go Taisyo-Kenkyu-no Tachiba-kara*. Kuroshio Publishers, Tokyo (1998)
7. Davidson, D.: The logical form of action sentences. In: Rescher, N. (ed.) *The Logic of Decision and Action*, pp. 81–95. University of Pittsburgh Press, Pittsburgh (1967)
8. Diesing, M.: *Indefinites*. MIT Press, Cambridge (1992)
9. Hasegawa, N.: Honorifics. In: Everaert, M., van Riemsdijk, H. (eds.) *The Blackwell Companion to Syntax*, pp. 493–543. Blackwell, Boston (2006)
10. Iida, T.: Semantics of possibility suffix “(Rar)e”. In: Nakakoji, K., Murakami, Y., McCready, E. (eds.) *JSAI-isAI 2009. LNCS (LNAI)*, vol. 6284, pp. 217–234. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14888-0_19
11. Inoue, K.: Case (with special reference to Japanese). In: Everaert, M., van Riemsdijk, H. (eds.) *The Blackwell Companion to Syntax*, vol. 1, pp. 295–373. Blackwell, Boston (2006)
12. Krifka, M., Pelletier, J., Carlson, G.: Genericity: an introduction. In: Carlson, G., Pelletier, F.J. (eds.) *The Generic Book*, pp. 1–123. The University of Chicago Press, Chicago (1995)
13. Moortgat, M.: Categorical type logics. In: van Benthem, J., ter Meulen, A. (eds.) *Handbook of Logic and Language*, 2nd edn, pp. 95–179. Elsevier, Amsterdam (2011)
14. Morrill, G.: *Type-Logical Grammar*. Springer, Dordrecht (1994). <https://doi.org/10.1007/978-94-011-1042-6>
15. Morrill, G.: Discontinuity in Categorical Grammar. *Linguist. Philos.* **18**, 175–219 (1995)

⁵ Chierchia [5] suggests they are generated in the Spec, AspectPhrase position, which agrees with its head (Asp) when the latter has the feature [+Q], standing for ‘quantificational.’ The head of ASP, Hab, with [+Q], allows episodic, stage-level predicates to receive habitual readings, yielding individual-level (inherently generic) verb phrases.

16. Morrill, G.: *Categorial Grammar - Logical Syntax, Semantics, and Processing*. Oxford University Press, Oxford (2011)
17. Portner, P.: *Modality*. Oxford University Press, New York (2009)
18. Saito, M., Hoshi, H.: Control in complex predicates. In: Report of the Special Research Project for the Typological Investigation of Languages and Cultures of the East and West, pp. 15–46. University of Tsukuba, Tsukuba (1998)
19. Steedman, M.: *Surface Structure and Interpretation*. MIT Press, New York (1996)
20. Tada, H.: Nominative objects in Japanese. *J. Jpn Linguist.* **14**, 91–108 (1992)
21. Takano, Y.: Nominative objects in Japanese and complex predicate constructions: a prolepsis analysis. *Nat. Lang. Linguist. Theory* **21**, 279–834 (2003)
22. Teremura, H.: *Nihongo-no Sinakusu-to Imi*. Kuroshio Publishers, Tokyo (1982)

AAA 2017

3rd Workshop on Argument for Agreement and Assurance (AAA 2017)

Kazuko Takahashi¹, Yoshiki Kinoshita², Tim Kelly³, and Hiroyuki Kido⁴

¹ Kwansai Gakuin University

² Kanagawa University

³ University of York

⁴ Sun Yat-sen University

Argumentation has now become an interdisciplinary research subject receiving much attention from diverse communities including formal logic, informal logic and artificial intelligence. It aims at analysing, evaluating and systematising various aspects of human argument appeared in television, newspapers, WWW, etc. and also artificial arguments constructed from structured knowledge with logical language and inference rules. Their research achievements are widely applicable to various domains such as safety, political, medical and legal domains.

In particular, safety engineering appreciates Toulmin's argument model, starting from his critical opinion on formal logic, and recent intensive studies of formal argumentation. There is a growing interest in the use of an evidence-based argument often called a safety case, assurance case or dependability case. Nowadays, it is becoming necessary for developing and operating bodies to comply international standards, for system stakeholders to make agreement, for system administrators to achieve accountability.

The international workshop on argument for agreement and assurance started on 2013, that aims at deepening mutual understanding to explore a new research field among researchers/practitioners in formal and informal logic, artificial intelligence, and safety engineering working on agreement and assurance through argument.

To that end, we called for submissions of the work in the following topics.

- Abstract and structured argumentation systems, e.g., frameworks, proof-theories, semantics and complexity.
- Dialogue systems, e.g., persuasion, negotiation, deliberation, eristic, and information-seeking dialogue systems.
- Formal underpinnings on assurance cases, e.g., frameworks, proof-theories, semantics and complexity.
- Confidence evaluation for assurance cases based on new metrics.
- Studies on patterns for assurance cases including new syntax, formal semantics, evaluation on their effectiveness.
- Argument-based agreement and assurance technologies for safety cases, assurance cases and dependability cases.
- Applications of argumentation and dialogue systems to agreement technologies, systems assurance, safety engineering, systems resilience, practical reasoning, belief revision, multi-agent systems, learning, and semantic web.

- Tools for argumentation systems, dialogue systems, safety case construction system, argument-based stakeholders' agreement, argument-based accountability achievement, argument-based open systems dependability, and argument-based verification and validation.

We have three contributions and two invited talks. One invited talk is by Professor Robin Bloomfield from the viewpoint of assurance case and the other is by Professor Anthony Hunter from the viewpoint of argumentation in artificial intelligence. In addition, we have five position papers at the session.

We thank all of the reviewers for their valuable comments, all the participants for fruitful discussions, and JSAI for giving us the opportunity to hold this international workshop.

Kazuko Takahashi, Yoshiki Kinoshita, Tim Kelly and Hiroyuki Kido (AAA 2017 organizers)



Invited Talk: Structured Engineering Argumentation

Robin E. Bloomfield^(✉)

Adelard LLP and City, University of London, London, UK
reb@adelard.com

Abstract. The decision to trust an engineering system—whether to fly in an aircraft, to drive a car - can have real world, societal, environmental and economic consequences. Engineering arguments are multidisciplinary and have a number of characteristics. A significant component of these decisions is science-based and may deploy sophisticated engineering calculations, mathematical models, simulations of the world and the engineered systems. However, this does not mean the judgments are purely deductive or logical. The framing of the problems, the validation of the assumptions, the application of “stopping rules” to decide when there is sufficient confidence is often an exercise in expert judgement. The overall process is socio-technical with challenge necessary to build confidence, and seeking dissent and counter-evidence important. One contribution to achieving confidence in engineering decisions is assurance cases: “a documented body of evidence that provides a convincing and valid argument that a system is adequately dependable for a given application in a given environment”. Our approach is based on the key concepts of claims, arguments and evidence (CAE): Claims—statements about a property of the system, Evidence that is used as the basis of the justification of the claim, Arguments link the evidence to the claim. Engineering justifications are too complex to express in terms of a simple CAE triple. If we are developing a top down justification, the claims need to be expanded into subclaims until we can identify evidence that can directly support the subclaims. Engineering assurance arguments tend to be some 10 s to 100 s of nodes and have considerable supporting narrative. We have developed an approach to structuring such arguments based on a set of archetypal CAE fragments that we have termed CAE building blocks. The identification of the blocks was supported by an empirical analysis of the types of engineering arguments that are made about safety and dependability from defence, finance and medical applications. Our approach factors out the argument into parts that can be addressed deductively and the side-warrant, which highlights the properties assumed of the world and have an inductive component. In this way we hope to get the benefits of deductive reasoning without losing the important argument that justifies why, in the real world, such deduction is appropriate and valid. These two aspects: the use of CAE fragments and the factorisation of deductive and inductive allow us to speculate how we can best exploit a variety of automated reasoning approaches.



Invited Talk: Computational Persuasion with Applications in Behaviour Change

Anthony Hunter^(✉)

Department of Computer Science, University College London, London, UK
anthony.hunter@ucl.ac.uk

Abstract. Persuasion is an activity that involves one party trying to induce another party to believe something or to do something. It is an important and multifaceted human facility. Obviously, sales and marketing is heavily dependent on persuasion. But many other activities involve persuasion such as a doctor persuading a patient to drink less alcohol, a road safety expert persuading drivers to not text while driving, or an online safety expert persuading users of social media sites to not reveal too much personal information online. As computing becomes involved in every sphere of life, so too is persuasion a target for applying computer-based solutions. An automated persuasion system (APS) is a system that can engage in a dialogue with a user (the persuadee) in order to persuade the persuadee to do (or not do) some action or to believe (or not believe) something. To do this, an APS aims to use convincing arguments and counterarguments in order to persuade the persuadee. Computational persuasion is a new field for the study of formal models of dialogues involving arguments and counterarguments, of user models, and strategies, for APSs. A promising application area for computational persuasion is in behaviour change. In this talk, I will review ongoing funded project (For more information, see www.computationalpersuasion.com) being undertaken in the UCL Intelligent Systems Group on developing a framework for computational persuasion for behaviour change technology [1, 2].

References

1. Hunter, A.: Computational persuasion with applications in behaviour change. In: Computational Models of Argument (COMMA 2016), pp. 5–18. IOS Press (2016)
2. Hunter, A.: Towards a framework for computational persuasion with applications in behaviour change. *Argument and Computation* (2018, in press)

SCIDOCA 17

Second International Workshop on Scientific Document Analysis (SCIDOCA 2017)

Yuji Matsumoto and Hiroshi Noji

Nara Institute of Science and Technology
{matsu,noji}@is.naist.jp

The Second International Workshop on SCIENTific DOCument Analysis (SCIDOCA 2017) associated with JSAI International Symposia on AI 2017 (IsAI-2017) was held at Bunkyo School Building on Tokyo Campus, University of Tsukuba Tokyo, on November 14 and 15, 2017.

SCIDOCA is an annual international workshop focusing on various aspects and perspectives of scientific document analysis for their efficient use and exploration. Recent proliferation of scientific papers and expert documents has become an obstacle to efficient acquisition of new information and findings described in documents in various fields. It is almost impossible for individual researchers to investigate and read all related documents. Even retrieving relevant documents is becoming harder and harder. This workshop attempted to gather researchers and experts who aim at scientific document analysis from various perspectives, and invited technical papers that cover any aspects of scientific document analysis.

This year, the workshop featured an invited talk by Simone Teufel titled “Do Future Work sections have a purpose? (and other global scientometric questions),” and 13 oral presentations (5 long papers and 8 short papers) selected by the program committee. The topics cover named-entity, relation and entailment recognition, document summarization, semantic analysis, citation analysis and other applications.

This volume includes the following two papers selected through peer review from the revised and resubmitted papers after the workshop.

- A Hierarchical Neural Extractive Summarizer for Academic Papers
by Kazutaka Kinugawa and Yoshimasa Tsuruoka
- Leveraging Document-specific Information for Classifying Relations in Scientific Articles
by Qin Daiy, Naoya Inouey, Paul Reiserz and Kentaro Inui

Finally, we would like to thank all the people who helped to organize this workshop. We would expecially thank the Program Committee members of the workshop: Takeshi Abekawa, Akiko Aizawa, Naoya Inoue, Kentaro Inui, Yoshinobu Kano, Yusuke Miyao, Junichiro Mori, Hidetsugu Nanba, Shoshin Nomura, Ken Satoh, Hiroyuki Shindo, Yoshimasa Tsuruoka, Minh Le Nguyen, and Pontus Stenetorp.



A Hierarchical Neural Extractive Summarizer for Academic Papers

Kazutaka Kinugawa^(✉) and Yoshimasa Tsuruoka

Department of Electrical Engineering and Information Systems,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{kinugawa,tsuruoka}@logos.t.u-tokyo.ac.jp

Abstract. Recent neural network-based models have proven successful in summarization tasks. However, previous studies mostly focus on comparatively short texts and it is still challenging for neural models to summarize long documents such as academic papers. Because of their large size, summarization for academic papers has two obstacles: it is hard for a recurrent neural network (RNN) to squash all the information on the source document into a latent vector, and it is simply difficult to pinpoint a few correct sentences among a large number of sentences. In this paper, we present an extractive summarizer for academic papers. The idea is converting a paper into a tree structure composed of nodes corresponding to sections, paragraphs, and sentences. First, we build a hierarchical encoder-decoder model based on the tree. This design eases the load on the RNNs and enables us to effectively obtain vectors that represent paragraphs and sections. Second, we propose a tree structure-based scoring method to steer our model toward correct sentences, which also helps the model to avoid selecting irrelevant sentences. We collect academic papers available from PubMed Central, and build the training data suited for supervised machine learning-based extractive summarization. Our experimental results show that the proposed model outperforms several baselines and reduces high-impact errors.

Keywords: Extractive summarization · Supervised machine learning
Recurrent neural network

1 Introduction

Automatic text summarization is the task of generating a shorter version of one or more documents while preserving essential information. Most of the existing summarization techniques are classified into two types: extractive summarization or abstractive summarization. Extractive summarization aims to select a set of important sentences from the original text and connect them coherently. In contrast, abstractive summarization, as humans do, aims to generate a summary which may contain words or phrases that do not exist in the source document. While neural abstractive summarization techniques are more actively studied

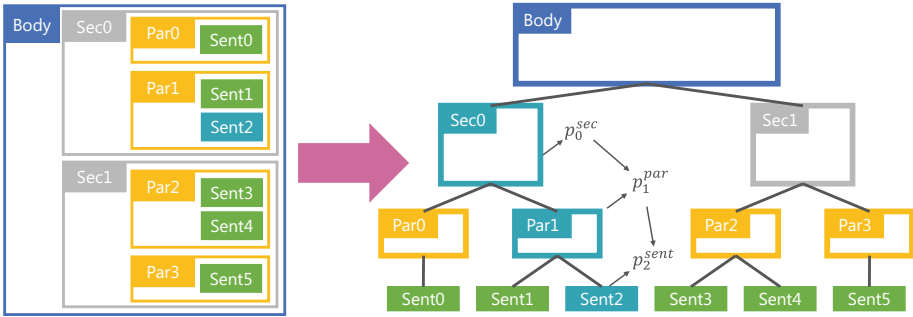


Fig. 1. An academic paper converted into a tree. Blocks in emerald green represent positive nodes. (Color figure online)

recently [1–3], extractive summarization is still attractive because it requires lower calculation costs and its outputs are guaranteed to be grammatically correct [4].

In recent summarization research, comparatively short texts such as news articles are often used [4–6], but extractive summarization for academic papers has also been studied [7, 8]. Academic papers are suited for building a corpus for single document summarization because abstracts can be used as reliable and high-quality reference summaries. Additionally, a large number of papers are downloadable from several databases and can be used to create training data for supervised learning. One of the main purposes of extractive summarization of academic papers is to generate structured abstracts. In general, the authors of a paper write an abstract to briefly explain the overview of their study, but many abstracts are not structured. By extracting salient sentences from the main body of a paper and annotating them with the names of the sections where respective sentences are extracted, the communicative function of every sentence in the summary becomes clear. Compared to directly predicting the particular role for every sentence in an abstract, summaries generated by this approach can be a more content-rich summary. On the other hand, this task is very challenging because the size of a paper is much larger than that of other corpora for summarization. More specifically, there are two problems due to their large sizes. First, it is difficult for a recurrent neural network (RNN) to obtain latent vector representations from long texts. RNNs are widely used for encoding sequences of words or sentences. When applying RNN-based approaches to this task however, RNNs are required to compress all the necessary information on many sentences included in a paper into fixed-size vectors. Second, there are simply too many sentence candidates to be extracted in a paper. The summarization model needs to select a few correct sentences among several hundreds of sentences.

In this paper, we present a novel neural extractive summarizer for academic papers to address the above two problems. The core of our proposal is the document structure of papers as shown in Fig. 1. For the first of the aforementioned problems, we have incorporated the tree structure into a hierarchical encoder-

decoder model, inspired by previous work [6,9]. For the second problem, we propose a tree structure-based scoring method. In the standard supervised learning-based approaches, every sentence in the training data is labeled as positive or negative beforehand, and a model is trained to predict sentence labels accurately. While most of those approaches focus only on sentences, we define positive and negative labels for sections and paragraphs as well, and train the model to predict their labels in parallel with those of sentences. In our method, the paragraphs and sections that have at least one positive sentence are defined as positive. Instead of directly scoring sentence nodes, our model calculates scores in order from the top nodes of the tree, and utilizes the scores of the parent nodes to calculate the scores of the lower nodes as shown in Fig. 1. This leads the model to predict sentence labels more accurately. Moreover, this method is expected to have another positive effect: the model may decrease errors that are particularly harmful. In preliminary experiments, we found that the sentences under positive nodes tend to be more relevant to its abstract, and sentences under negative nodes tend to be less relevant. Our tree-based selection is thus expected to prevent the model from extracting irrelevant sentences because the model is trained to prefer positive nodes to negative nodes.

We conduct experiments on academic papers in the biomedical domain available from PubMed Central. The results show that our model achieves better scores than several baseline models and improves the quality of errors. In addition, our approach could be applied to other types of long texts because we only used information about document structure for our model and it is not a specific feature to academic papers.

2 Document Structure and Summarization

In supervised machine learning-based approaches to extractive summarization, every sentence in the training data is labeled as positive or negative beforehand. Some studies automatically conducted this preparation by using some algorithm [4–6]. In our case, the sentences included in the set that best represent the abstract are labeled as positive, and the rest are labeled as negative. Basically, the model is trained to extract the positive sentences. However, because of its large size, some negative sentences are also expected to contain essential information. Here, we assume that the sentences located near a positive sentence are related to the abstract. In contrast, sentences located far from a positive sentence are expected to be irrelevant. In order to verify our assumption, we categorized all negative sentences into the following three groups in terms of paragraphs that the sentence belongs to.

Group A

belonging to the paragraphs that have at least one positive sentence.

Group B

belonging to the paragraphs that do not have any positive sentence but are included in the same section as group A.

Table 1. ROUGE-1 F_1 score and ROUGE-2 F_1 score of the three groups.

Group	ROUGE-1	ROUGE-2
Group A	25.4	6.25
Group B	25.1	5.39
Group C	20.9	3.90

Group C

belonging to the paragraphs whose parent section does not have any positive sentence.

In the case of Fig. 1, Sentence 1 is in group A, Sentence 0 is in group B, and Sentence 3, 4, and 5 are in group C. The three groups are closer to the positive sentences in order from group A to group C. We conducted a preliminary experiment to check that the closer a paragraph is to positive sentences, the more important that paragraph is. We calculated the ROUGE scores [10] of every independent paragraph against its corresponding abstract¹, and took the average on each group. The scores are shown in Table 1. The results quantitatively show that sentences located near a positive sentence tend to be relevant to the abstract. Therefore, it is also important to avoid selecting sentences in group C.

3 Problem Formulation

First, we assume that an academic paper is hierarchically comprised of four types of segments: section, paragraph, sentence and word. The inclusion relation between a section and its subsection(s) is not considered, and subsections are regarded as independent sections. Let x , c , p , s and w denote a paper, a section, a paragraph, a sentence and a word, respectively. Here, x is comprised of a sequence of sections, and similarly, c , p and s are comprised of a sequence of corresponding lower-level segments. If a segment is located in the top/end of the sequence, its suffix is represented as *top/end*. And the suffix of a parent node is represented as *parent*. Besides, a section, a paragraph, a sentence and a word are associated with a vector \mathbf{v}^{sec} , \mathbf{v}^{par} , \mathbf{v}^{sent} and \mathbf{w} , respectively.

Sentential extractive summarization aims to output a summary from x by selecting a subset of the original sentences. In our case, we regard abstracts as reference summaries, and do not use any information on an abstract for extracting salient sentences from the main body of the paper. We address this task as the problem of binary classification of each sentence. Specifically, by supervised training, we build a classifier that scores each sentence in x and predicts a label $y_t^{sent} \in \{0, 1\}$ which indicates whether the t -th sentence should be included in the summary or not. For simplicity, we denote $p(y_t^{sent} = 1|x) \in [0, 1]$ as p_t^{sent} , which is the score of the t -th sentence calculated by the summarizer. In addition, as mentioned before, we train the model to predict labels of paragraphs

¹ When we calculated ROUGE scores, positive sentences in group A were removed.

and sections as well. Labels and scores of paragraphs and sections are defined analogously to sentences. The objective function to be maximized is as follows:

$$\begin{aligned}
 E = & - \sum_i \{y_i^{sent} \log p_i^{sent} + (1 - y_i^{sent}) \log (1 - p_i^{sent})\} \\
 & - \sum_j \{y_j^{par} \log p_j^{par} + (1 - y_j^{par}) \log (1 - p_j^{par})\} \\
 & - \sum_k \{y_k^{sec} \log p_k^{sec} + (1 - y_k^{sec}) \log (1 - p_k^{sec})\}
 \end{aligned} \tag{1}$$

4 Summarization Model

This section introduces our encoder-decoder summarization model, which is inspired by Cheng and Lapata [6]. We describe four sub-modules of the entire model: the convolutional sentence encoder, the document encoder, the document decoder and the tree structure-based classifier. The process proceeds as follows. First, every sentence included in the main body of a paper is vectorized by the convolutional sentence encoder. Second, the sequence of sentence vectors is fed into the document encoder hierarchically. Third, the sequence of salience scores of every sentence is obtained using the document decoder and the classifier. Finally, we use sentences with the highest scores as the summary subject to the word limit.

4.1 Convolutional Sentence Encoder

We use a Convolutional Neural Network (CNN) to obtain a sentence vector \mathbf{v}^{sent} from the corresponding word-vector sequence $[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{end}]$. CNNs have been shown to extract word-sequence patterns helpful to classification tasks [11]. In addition, a single-layer CNN can be calculated faster than an RNN. A CNN is suited for our task because the size of an input sequence is much larger than that in widely used summarization corpora. Let $\mathbf{w}_{j:j+a-1}$ denote the concatenated vector of successive a words from the j -th word in the sentence. First, a convolution operation involves the i -th filter with width a , which is applied to the input $\mathbf{w}_{j:j+a-1}$ to produce a new feature $f_{a,j}^{(i)}$ as

$$f_{a,j}^{(i)} = \tanh(\mathbf{K}_a^{(i)} \cdot \mathbf{w}_{j:j+a-1} + b_a^{(i)}) \tag{2}$$

where $\mathbf{K}_a^{(i)}$ and $b_a^{(i)}$ are both learnable parameters of the CNN. This filter is applied to each possible window of words in the sentence to produce a feature map $\mathbf{f}_a^{(i)} = [f_{a,1}^{(i)}, f_{a,2}^{(i)}, \dots, f_{a,end-a+1}^{(i)}]$. Then a max pooling operation is applied over the feature map, and the maximum value $\hat{f}_a^{(i)} = \max\{f_a^{(i)}\}$ is taken as the feature corresponding to this particular filter. The series of calculations is applied for each a and i . Finally, the obtained $\hat{f}_a^{(i)}$ are all concatenated to produce a sentence vector \mathbf{v}^{sent} .

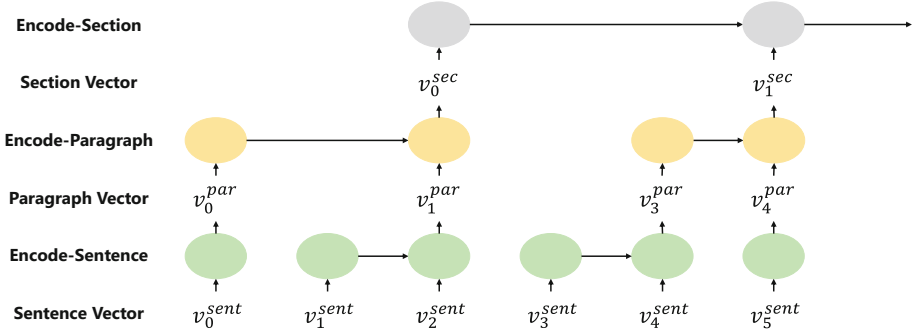


Fig. 2. Neural hierarchical encoder. Green, yellow and grey blocks represent LSTM_{enc}^{sent} , LSTM_{enc}^{par} and LSTM_{enc}^{sec} , respectively. (Color figure online)

4.2 Document Encoder

We use a Long Short-Term Memory (LSTM) unit [12] as the basic building block of our document encoder and decoder. For simplicity, we define $\mathbf{h}_t = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1})$ to be the LSTM operation on an input vector \mathbf{x}_t and the previous hidden state \mathbf{h}_{t-1} to compute the current hidden state \mathbf{h}_t . Our encoder-decoder model is comprised of three layers: the sentential layer, the paragraphical layer and the sectional layer. This hierarchical design is inspired by Li et al. [9]. The LSTM of each layer is denoted as LSTM_{enc}^{sent} .

First, all sentence vectors of a paragraph are input into sentential encoders.

$$\mathbf{h}_t^{sent} = \text{LSTM}_{enc}^{sent}(\mathbf{v}_t^{sent}, \mathbf{h}_{t-1}^{sent}) \quad (3)$$

If a sentence is located in the top of the paragraph, the previous hidden states are set to a zero vector.

$$\mathbf{h}_{top}^{sent} = \text{LSTM}_{enc}^{sent}(\mathbf{v}_{top}^{sent}, \mathbf{0}) \quad (4)$$

The hidden state vector at the end of the paragraph is used to represent the entire paragraph as

$$\mathbf{v}_{parent}^{par} = \mathbf{h}_{end}^{sent} \quad (5)$$

This is a simple way to obtain a paragraph vector, and there are several possible approaches in this step such as calculating the weighted sum of the hidden state vectors of all sentences in that paragraph using the self-attention mechanism [13]. Similarly, paragraphs and sections are encoded hierarchically, receiving embeddings of the lower layers as input.

$$\mathbf{h}_{top}^{par} = \text{LSTM}_{enc}^{par}(\mathbf{v}_{top}^{par}, \mathbf{0}) \quad (6)$$

$$\mathbf{h}_t^{par} = \text{LSTM}_{enc}^{par}(\mathbf{v}_t^{par}, \mathbf{h}_{t-1}^{par}) \quad (7)$$

$$\mathbf{v}_{parent}^{sec} = \mathbf{h}_{end}^{par} \quad (8)$$

$$\mathbf{h}_{top}^{sec} = \text{LSTM}_{enc}^{sec}(\mathbf{v}_{top}^{sec}, \mathbf{0}) \quad (9)$$

$$\mathbf{h}_t^{sec} = \text{LSTM}_{enc}^{sec}(\mathbf{v}_t^{sec}, \mathbf{h}_{t-1}^{sec}) \quad (10)$$

The encoding steps are shown in Fig. 2.

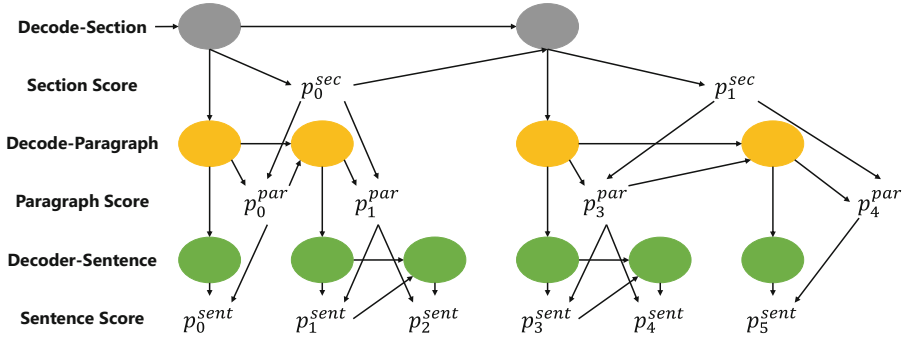


Fig. 3. Neural hierarchical decoder. Green, yellow and grey blocks represent $LSTM_{dec}^{sent}$, $LSTM_{dec}^{par}$ and $LSTM_{dec}^{sec}$, respectively. (Color figure online)

4.3 Document Decoder

On the decoder side, vectors are propagated through the three layers in the reverse order of the encoder side. On each layer, the $(t-1)$ -th predicted score is used at the next time step t for the decoder to memorize how much previous segments are extracted.

First, the sectional decoder receives the end-of-the document token (EOD) as input, and the output of the sectional encoder \mathbf{h}_{end}^{sec} as a hidden state.

$$\bar{\mathbf{h}}_{top}^{sec} = LSTM_{dec}^{sec}(\mathbf{v}_{eod}, \mathbf{h}_{end}^{sec}) \tag{11}$$

where \mathbf{v}_{eod} is a vector representing the EOD token, and this is also a model parameter. The sectional decoder outputs section-level representations, receiving the vector representing the previous section \mathbf{v}_{t-1}^{sec} , which is obtained on the encoder side, and the previous score p_{t-1}^{sec} , which is calculated on the classification layer, as the input.

$$\bar{\mathbf{h}}_t^{sec} = LSTM_{dec}^{sec}(p_{t-1}^{sec} \mathbf{v}_{t-1}^{sec}, \bar{\mathbf{h}}_{t-1}^{sec}) \tag{12}$$

Next, $\bar{\mathbf{h}}_t^{sec}$ is used as the initial input into the paragraphical decoder corresponding to the top of the section to output paragraph-level representations.

$$\bar{\mathbf{h}}_{top}^{par} = LSTM_{dec}^{par}(\bar{\mathbf{h}}_{parent}^{sec}, \mathbf{0}) \tag{13}$$

Similarly, paragraph vectors and sentence vectors are processed sequentially on respective layers.

$$\bar{\mathbf{h}}_t^{par} = LSTM_{dec}^{par}(p_{t-1}^{par} \mathbf{v}_{t-1}^{par}, \bar{\mathbf{h}}_{t-1}^{par}) \tag{14}$$

$$\bar{\mathbf{h}}_{top}^{sent} = LSTM_{dec}^{sent}(\bar{\mathbf{h}}_{parent}^{par}, \mathbf{0}) \tag{15}$$

$$\bar{\mathbf{h}}_t^{sent} = LSTM_{dec}^{sent}(p_{t-1}^{sent} \mathbf{v}_{t-1}^{sent}, \bar{\mathbf{h}}_{t-1}^{sent}) \tag{16}$$

The decoding steps are shown in Fig. 3.

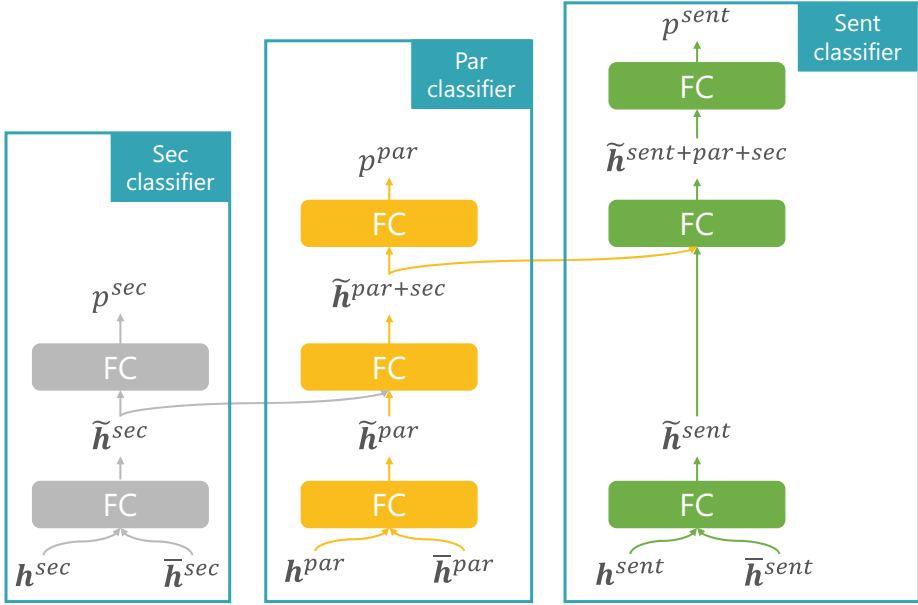


Fig. 4. Multi-layer perceptrons considering the tree structure. FC represents a fully connected layer, and \tilde{h} represents a hidden state vector of MLPs.

4.4 Classifiers

We use multi-layer perceptrons (MLPs) as the classifiers for sections, paragraphs and sentences. Each MLP receives a pair of hidden state vectors, h_t and \bar{h}_t , from the encoder and the decoder.

First, the score of a parent section is calculated as

$$\tilde{h}_t^{sec} = \tanh(W_1^{sec}[h_t^{sec}; \bar{h}_t^{sec}] + b_1^{sec}) \tag{17}$$

$$p_t^{sec} = \sigma(W_2^{sec}\tilde{h}_t^{sec} + b_2^{sec}) \tag{18}$$

where σ indicates the sigmoid function, and the semicolon represents the concatenation operation. Here, the hidden state vector \tilde{h}_t^{sec} obtains the semantic information on its section from h_t^{sec} and \bar{h}_t^{sec} , and also obtains the latent features representing how important its section is through the backpropagation from the loss between p_t^{sec} and y_t^{sec} .

Next, the scores of the child paragraphs are calculated by the MLP receiving the hidden state vector of the parent section as

$$\tilde{h}_t^{par} = \tanh(W_1^{par}[h_t^{par}; \bar{h}_t^{par}] + b_1^{par}) \tag{19}$$

$$\tilde{h}_t^{par+sec} = \tanh(W_2^{par}[\tilde{h}_{parent}^{sec}; \tilde{h}_t^{par}] + b_2^{par}) \tag{20}$$

$$p_t^{par} = \sigma(W_3^{par}\tilde{h}_t^{par+sec} + b_3^{par}) \tag{21}$$

where $\mathbf{h}_{parent}^{sec}$ indicates the hidden state vector of the MLP for the parent section. Like $\tilde{\mathbf{h}}_{parent}^{sec}$, $\tilde{\mathbf{h}}_t^{par+sec}$ contains not only the semantic information on its paragraph, but also the latent features representing how important its paragraph and its parent section are.

Finally, the scores of the child sentences are calculated as

$$\tilde{\mathbf{h}}^{sent} = \tanh(W_1^{sent}[\mathbf{h}^{sent}; \bar{\mathbf{h}}^{sent}] + \mathbf{b}_1^{sent}) \quad (22)$$

$$\tilde{\mathbf{h}}^{sent+par+sec} = \tanh(W_2^{sent}[\tilde{\mathbf{h}}_{parent}^{par+sec}; \tilde{\mathbf{h}}^{sent}] + \mathbf{b}_2^{sent}) \quad (23)$$

$$p^{sent} = \sigma(W_3^{sent}\tilde{\mathbf{h}}^{sent+par+sec} + \mathbf{b}_3^{sent}) \quad (24)$$

where $\mathbf{h}_{parent}^{par+sec}$ indicates the hidden state vector of the MLP for the parent paragraph, and $\tilde{\mathbf{h}}^{sent+par+sec}$ indicates the hidden state vector of this MLP, which contains information on its sentence, its parent paragraph and its parent section and is used for calculation of the sentence score. The classification steps are shown in Fig. 4.

5 Experimental Setup

Dataset. A subset of the papers available from PubMed Central² is used in the experiments. The papers are downloadable as XML files. We eliminate unnecessary tags, and perform tokenization and sentence splitting using the Stanford coreNLP tool³. In our experiments, the total number of papers is 30,000. The papers are split into approximately 90% for training, 5% for validation and 5% for testing. The number of words in the extracted sentences was restricted to be smaller than 6% of the total number of words in the main body of the corresponding paper.

As mentioned before, we address the summarization task as the problem of binary classification of each sentence in the main body of a paper. One problem with this approach is how to create training data because human-written summaries cannot be readily used as the gold-standard labels for this task. To tackle this problem, in the extractive summarization domain, simple unsupervised approaches such as greedy selection are often used to convert the human-written reference summaries to extractive labels [4, 5]. In our case, we need to compute a mapping between the sentences in each main article and its abstract. To identify the set of sentences that best represents the abstract, we use a dynamic programming method [14] which allows us to obtain the set of sentences that (approximately) maximizes the ROUGE-2 F-score [10] against the reference abstract. The sentences included in the resulting set are labeled as positive (i.e. the ones that should be extracted), and the rest are labeled as negative.

² ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/.

³ <http://stanfordnlp.github.io/CoreNLP/>.

Implementation Details. The word embeddings were pre-trained using the Skip-gram model [15] on our training data. We replaced the words which appear less than 5 times in the training data with the token *UNK*. The dimensions of word and sentence embeddings were set to 300 and 600, respectively. The sizes of all LSTM hidden states were set to 600. We used a list of kernel sizes $\{1, 2, 3, 4, 5, 6\}$ for the CNN. We performed mini-batch training with a batch size of 30 articles. Adam [16] was used for updating parameters. The parameters were set as $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We use dropout [17] with probability $p = 0.5$. We implement the code in C++⁴ using the Eigen library⁵, a template library for linear algebra, and run all of the experiments on multicore CPUs.

Baselines. As a simple baseline, we adopted LEAD, which returns the first sentences of a paper. Next, we built a classifier with the logistic regression model, LREG. For LREG, the following hand-crafted features were used: the words included in a sentence, the position of the sentence in the paper, that in the paragraph, that in the section, the position of the paragraph in the paper, that in the section and the position of the section in the paper. In addition, we implemented NN-SE [6]. NN-SE is a single-layer encoder-decoder model, and the difference between NN-SE and our model is the introduction of the hierarchical structure for the encoder-decoder part and the tree structure-based classifier. For comparison, we also implemented THREE-LAYER NN-SE, which is a third-layer encoder-decoder model focusing only on sentences. The hyperparameters were set to the same values as those of our model.

6 Results

First, we evaluated the generated summaries using the ROUGE metric [10]. We report unigram and bigram overlap (ROUGE-1 and ROUGE-2), and the longest common sequence (ROUGE-L)⁶. Table 2 shows our results, where GOLD shows the scores of the sentence sets extracted by the aforementioned dynamic programming method. The scores of GOLD therefore indicate the approximate upper bounds of the performance that can be achieved by the sentence extraction approach. Our proposed model outperforms all baseline models. However, the fact that GOLD achieves much better scores than the others means that extractive summarization has plenty of room for improvement.

Next, we evaluated the percentage of positive sentences successfully extracted by models to the corresponding gold-standard set. This metric measures how much each model learns patterns of sentence extraction provided by GOLD. The results are shown in Table 3. Our proposed model achieves better performance than the other two models, and this fact indicates that our tree-based scoring method effectively helps the model to extract correct sentences.

⁴ Our code is available at <https://github.com/kazu-kinugawa/HNES>.

⁵ <http://eigen.tuxfamily.org/index.php>.

⁶ The ROUGE evaluation option is, -m -n 2 -w 1.2.

Table 2. ROUGE F_1 scores (%) on the test set.

Model	ROUGE-1	ROUGE-2	ROUGE-L
GOLD	62.4	39.0	59.5
LEAD	37.1	10.2	34.3
LREG	49.7	20.6	46.4
NN-SE	50.0	21.0	47.9
THREE-LAYER NN-SE	50.1	21.3	48.1
PROPOSED MODEL	50.7	22.1	48.6

Table 3. Coverage of sentences (%) on the test set.

Model	Coverage
NN-SE	28.9
THREE-LAYER NN-SE	30.0
PROPOSED MODEL	31.2

Table 4. The average number of wrong sentences in each group on the test set.

Model	Group A	Group B	Group C	Total
NN-SE	2.68	2.39	2.59	7.65
THREE-LAYER NN-SE	2.78	2.35	2.46	7.59
PROPOSED MODEL	2.92	2.06	2.34	7.33

Table 5. Error distribution (%) on the test set.

Model	Group A	Group B	Group C
NN-SE	34.9	29.9	35.3
THREE-LAYER NN-SE	36.7	29.7	33.5
PROPOSED MODEL	40.3	26.7	33.1

In addition, we investigated the breakdowns of negative sentences extracted by respective models in terms of the aforementioned three groups. Tables 4 and 5 show the average number of sentences in each group and its distribution, respectively. Our proposed model achieves the least number of errors among the three models, and succeeds in decreasing the number of sentences in group B and C, and increasing the number of sentences in group A. This fact indicates that our proposed method has the potential to improve the quality of errors.

Figure 5 shows an example of extracted sentences by respective models. In this case, although the proposed model extracts just one more correct sentence than THREE-LAYER NN-SE, the proposed model outperforms THREE-LAYER NN-SE by approximately 9.0 ROUGE-2 scores. This result is supported

by the fact that our model successfully collects more sentences in group A. The detailed contents of these sentences are shown in Fig. 6. In this case, the target paper describes the process of developing a measure of teamwork for community-based health teams in rural Zambia. Sentence 76, which is a positive sentence, says that the authors focus on processes that comprise the teamwork construct among selected factors. Sentence 72 and 79, which are extracted by the proposed model, explain the analysis of several factors for measuring teamwork and they are closely relevant to Sentence 76. In contrast, Sentence 38 and 47, which are extracted by the existing model, report a group discussion in the experiment, and they are irrelevant to this context. Moreover, we visualize the section scores of this paper predicted by the proposed model, as shown in Fig. 7. In this figure, the model succeed in focusing on important sections like “background”, “determinants of teamwork”, “discussion” and “conclusion”. However, the model wrongly pays attention to several irrelevant sections like “author’s contribution”, even though the model does not extract sentences from this section.

<p>GOLD: 10, 13, 19, 25, 76, 118, 120, 122, 164</p> <hr/> <p>THREE-LAYER NN-SE: <u>23</u>, <u>25</u>, <u>26</u>, <u>38</u>, <u>47</u>, <u>106</u>, <u>108</u>, <u>110</u>, 118, <u>162</u>, <u>163</u>, 164, <u>169</u>, <u>170</u></p> <hr/> <p>PROPOSED MODEL: 25, <u>26</u>, <u>72</u>, <u>79</u>, <u>106</u>, 118, <u>119</u>, 120, <u>121</u>, <u>161</u>, <u>162</u>, <u>163</u>, 164</p>
--

Fig. 5. IDs of sentences extracted by GOLD, THREE-LAYER NN-SE and PROPOSED MODEL. The blue numbers indicate that they are correct, the green numbers underscored with an underline indicate that they are wrong but included in the group A, and the red numbers underscored with a wavy line indicate that they are wrong and included in group C. (Color figure online)

Finally, we report a characteristic error in Fig. 8. We found that the model sometimes selects negative sentences successively. A possible reason is that this pattern is included in some gold-standard sets of sentences. As mentioned before, the gold-standard sets of sentences are automatically made with the dynamic programming method in this experiment, and thus such an unnatural pattern sometimes appear. Therefore it is not necessarily wrong for the model to learn to give higher scores successively, but if the model starts this action from a wrong sentence, it will cause a fatal error. One approach to alleviate this problem is to consider coverage and redundancy. For example, Chen et al. [18] proposed a distraction mechanism to force a model to pay attention to content of a document all around. Another possible solution is to take measures against the data imbalance. In our dataset, the number of negative examples is much larger than that of positive ones and the trained model outputs very low scores overall. It is needed to adjust the loss function in Eq. 1 to impose different penalties on

<p>Abstract: (...) Measuring teamwork requires identifying dimensions of teamwork or processes that comprise the teamwork construct, while taskwork requires identifying specific team functions. (...)</p>
<p>Sentence #76 (GOLD): We further categorized the selected factors into dimensions of teamwork, or processes that comprise the teamwork construct.</p>
<p>Sentence #72 (PROPOSED MODEL): We used a weighting system to select factors for measuring teamwork from those identified and sorted by the participants.</p>
<p>Sentence #79 (PROPOSED MODEL): We categorized these factors (determinants) into three groups: personal, community-related and service-related.</p>
<p>Sentence #38 (THREE-LAYER NN-SE): A total of 36 individuals were involved.</p>
<p>Sentence #47 (THREE-LAYER NN-SE): The timeline activity initiated dialogue on teamwork.</p>

Fig. 6. Contents of sentences extracted by respective models. The blue words indicate keywords in this context. (Color figure online)

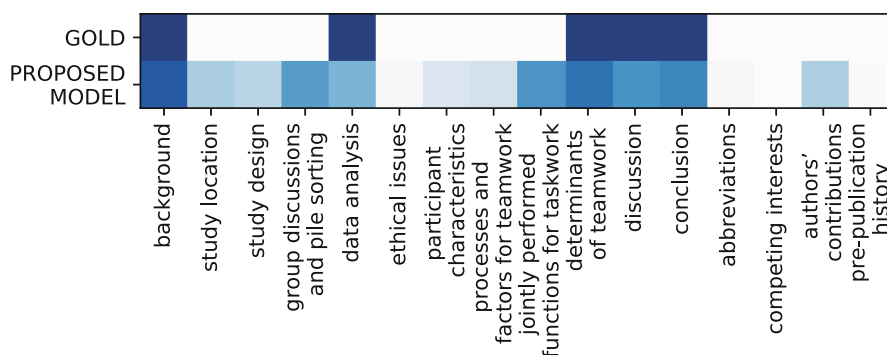


Fig. 7. Visualization of respective section scores. Dark blue cells indicate that their corresponding scores are close to 1. Especially, dark blue cells in the first line indicate that their corresponding sections include positive sentences. (Color figure online)

GOLD: 0, 4, 19, 22, 64, 82, 148, 153, 192, 193
<hr/> PROPOSED MODEL: 0, 11, 14, 22, 44, 148, 150, 169, 184, <u>233</u> , <u>234</u> , <u>235</u> , <u>236</u>

Fig. 8. IDs of sentences extracted by GOLD and PROPOSED MODEL. The blue numbers indicate that they are correct, and the magenta numbers underscored with an double underline indicate that they are successively wrong. (Color figure online)

positive sentences and negative ones, and then an appropriate threshold should be introduced for the model to discard unnecessary sentences. We leave them to future work.

7 Related Work

Machine learning-based approaches to extractive summarization traditionally rely on hand-crafted features such as sentence position and length. Recently, however, several powerful neural models have drawn much attention. Cheng and Lapata [6] proposed a novel data-driven model based on the sequence-to-sequence framework as in machine translation. In their model, an encoder receives a sequence of sentences represented as vectors, and a decoder outputs a sequence of respective salience scores. Based on Cheng and Lapata’s model [6], Isonuma et al. [5] advocated the idea of combing extractive summarization and document classification, which enables sentence extraction for datasets which lack reference summaries. Narayan et al. [19] also expanded Cheng and Lapata’s model [6] to utilize image captions included in news articles. Nallapati et al. [4] proposed a two-layer bidirectional RNN model, which encodes a text hierarchically and classifies each sentence on the basis of its hidden state vector and several human engineered features.

There have been several studies on extractive summarization for single academic papers. Contractor et al. [7] proposed an approach to using Argumentative Zoning for extractive summarization. Collins et al. [8] introduced a new feature, AbstractROUGE, which is the ROUGE-L score of a sentence against the corresponding abstract. While these studies are based on feature engineering specific to academic papers, our approach can be applied to any types of text which have some document structure.

Obtaining latent vector representations from a long document is a challenging task. Several studies tackle this problem by designing hierarchical RNN models. Li et al. [9] built a two-layer autoencoder composed of a word-level LSTM and a sentence-level LSTM, which constructs embeddings in a bottom-up fashion. Yang et al. [13] introduced a hierarchical bidirectional RNN model equipped with an internal attention mechanism for document classification. Tai et al. [20] proposed the Tree-LSTM, which is suited for tree structures such as dependency trees and constituency trees, although this is not a study for documents. Gur et

al. [21] presented an approach to dividing a long document into small pieces and then encoding all windows in parallel with an RNN.

8 Conclusion

In this paper, we present a hierarchical encoder-decoder summarizer for academic papers. Our focus is the document structure of academic papers. First, we design an encoder-decoder model according to the document structure. Second, we target paragraphs and sections for extraction as well and train the model to predict their labels in parallel with sentences. Our model calculates the salience score of each sentence considering the salience scores of its parent paragraph and section based on the tree structure. This helps the model to extract correct sentences accurately, and discourages the model from making high-impact mistakes at the same time. Paragraph and section vectors obtained with the above hierarchical design are utilized in the scoring stage. Experimental results show that our summarizer achieves higher ROUGE scores than baseline models. Additionally, the number of fatal errors is reduced.

References

1. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for sentence summarization. In: Proceedings of EMNLP, pp. 17–21 (2015)
2. Zhou, Q., Yang, N., Wei, F., Zhou, M.: Selective encoding for abstractive sentence summarization. In: Proceedings of ACL, pp. 1095–1104 (2017)
3. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. In: Proceedings of ACL, pp. 1073–1083 (2017)
4. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: an interpretable recurrent neural network model for extractive summarization. In: Proceedings of AAAI, pp. 3075–3081 (2017)
5. Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., Sakata, I.: Extractive summarization using multi-task learning with document classification. In: Proceedings of EMNLP, pp. 2101–2110 (2017)
6. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. In: Proceedings of ACL, pp. 484–494 (2016)
7. Contractor, D., Guo, Y., Korhonen, A.: Using argumentative zones for extractive summarization of scientific articles. In: Proceedings of COLING, pp. 663–678 (2012)
8. Collins, E., Augenstein, I., Riedel, S.: A supervised approach to extractive summarisation of scientific papers. In: Proceedings of CoNLL, pp. 195–205 (2017)
9. Li, J., Luong, M.T., Jurafsky, D.: A hierarchical neural autoencoder for paragraphs and documents. In: Proceedings of ACL, pp. 1106–1115 (2015)
10. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out, pp. 74–81 (2004)
11. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of EMNLP, pp. 1746–1751 (2014)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)

13. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of NAACL, pp. 1480–1489 (2016)
14. Nakasuka, K., Tsuruoka, Y.: Auto summarization for academic papers based on discourse structure. In: Proceedings of ANLP 2015, pp. 569–572 (2015). (in Japanese)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in NIPS, pp. 3111–3119 (2013)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (2015)
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
18. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H.: Distraction-based neural networks for modeling documents. In: Proceedings of IJCAI-2016, pp. 2754–2760 (2016)
19. Narayan, S., Papasrantopoulos, N., Lapata, M., Cohen, S.B.: Neural extractive summarization with side information. [arXiv:1704.04530](https://arxiv.org/abs/1704.04530) (2017)
20. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of ACL, pp. 1556–1566 (2015)
21. Gur, I., Hewlett, D., Lacoste, A., Jones, L.: Accurate supervised and semi-supervised machine reading for long documents. In: Proceedings of EMNLP, pp. 2011–2020 (2017)



Leveraging Document-Specific Information for Classifying Relations in Scientific Articles

Qin Dai¹(✉), Naoya Inoue¹, Paul Reisert², and Kentaro Inui^{1,2}

¹ Tohoku University, Sendai, Japan

{daiqin,naoya-i,inui}@ecei.tohoku.ac.jp

² RIKEN Center for Advanced Intelligence Project, Wako, Japan

preisert@ecei.tohoku.ac.jp

Abstract. Tremendous amount of knowledge is present in the ever-growing scientific literature. In order to grasp this massive amount knowledge, various computational tasks are proposed for training computers to read and analyze scientific documents. As one of these task, semantic relationship classification aims at automatically analyzing semantic relationships in scientific documents. Conventionally, only a limited number of commonly used knowledge bases such as Wikipedia are used for collecting background information for this task. In this work, we hypothesize that scientific papers also could be utilized as a source of background information for semantic relationship classification. Based on the hypothesis, we propose the model that is capable of extracting background information from unannotated scientific papers. Preliminary experiments on the RANIS dataset [1] proves the effectiveness of the proposed model on relationship classification in scientific articles.

Keywords: Semantic relationship · Scientific document · Lexical chain

1 Introduction

In recent years, with an increase in the number of scientific papers, it is prohibitively time-consuming for researchers to review and fully-comprehend all papers. To effectively and quickly access a large amount of scientific papers and grasp useful knowledge, a wide variety of computational studies for structuralizing scientific papers have been conducted, such as Argumentative Zoning [2], BioNLP Shared Task [3] and ScienceIE Shared Task [4].

One fundamental and promising task is *relation classification*. In this work, we address the task of *relation classification* in Computer Science (CS) papers [1], as an approach for automatic scientific paper analysis. Specifically, given a sentence from a CS paper marked with two *target entities* (represented by A, B), the task is to identify the type of predefined semantic relationship for the marked

target entity¹ pair. For notational convenience, we refer to the sentence where target entity pairs are extracted from as *target sentence*. Suppose that the following target sentence from a CS paper is given:²

- (1) *We introduce referential translation machines (\underline{RTM}_A) for $\underline{quality\ estimation}_B$ of translation outputs of sentence-level and word-level statistical machine translation (SMT) quality.*

The goal is to train a model to correctly classify the semantic relationship as *Apply-to*(\underline{RTMs}_A , $\underline{quality\ estimation}_B$), which namely means that \underline{RTMs}_A is used as a method for the purpose of $\underline{quality\ estimation}_B$. Many applications, such as scientific question answering (QA) and scientific paper summarization, could benefit from this level of fine-grained analysis.

There is a large body of work on relation classification in a general domain [5,6]. These approaches depend on complex features such as manually prepared lexical-syntactic patterns [7–9, etc.]. Recently, Neural Network (NN)-based approaches achieve close or even better performance than previous approaches without the need for complicated, manually prepared features [10–12]. In the context of scientific relation identification, Ammar et al. [13] enhanced Miwa and Bansal [14]’s end-to-end general relation extraction model by incorporating external knowledge such as gazetteer-like information extracted from Wikipedia.

However, no previous research leverages prevalent scientific papers as a source of background information for relation classification. We hypothesize that the scientific paper where a target sentence is placed could be utilized as a source of background information to facilitate relation classification. We call such background information that is depicted in the document where a target sentence is placed as *document-specific information*. For example, for a machine, the target sentence (1) might literally mean that “RTM is introduced for quality estimation”, but does not necessarily imply that “RTM is used for quality estimation”. However, if we present the machine with some document-specific information from the scientific paper where the target sentence (1) is extracted from, such as sentence (2), it would be easier for the machine to correctly interpret the target sentence (1) as that “RTM is used for quality estimation”. This clarification thereby could help the machine to correctly identify the relationship between \underline{RTM}_A and $\underline{quality\ estimation}_B$ as *Apply-to*, because comparing to sentence (1), sentence (2) clearly denotes this relationship via some informative key words like “...be used for...”.

- (2) *RTM can be used for predicting the quality of translation outputs.*

In this work, we propose a novel scientific relation classification model that extends a state-of-the-art neural relation classification model by leveraging scientific papers as a source of document-specific information. We propose three

¹ The phrase *target entity* refers not merely the concept denoted by noun or noun phrase, but it could be an action denoted by a verb or verb phrase and some quality denoted by an adjective, adverb, etc.

² This example is taken from W13-2242, ACL anthology (<http://aclanthology.info>).

methods to extract document-specific information from scientific papers. Our evaluation empirically demonstrates that incorporating document-specific information improves the performance of scientific relation classification on Tateishi et al. [1]’s RANIS corpus, the scientific semantic relationship-annotated corpus collected from CS paper abstracts.

2 Related Work

Conventional approaches to relation classification rely on human-designed complex lexical syntactic patterns [7], statistical co-occurrences [8] and a structuralized knowledge base such as WordNet [9,15]. In recent years, Neural Networks (NNs) are the dominant approach in the field. Zeng et al. [10] proposed a deep Convolutional Neural Network (CNN)-based framework, which depends on sentence-level features collected from a target sentence and lexical-level features acquired from lexical resources such as WordNet [16]. Santos et al. [12] proposed a ranking CNN model, which is trained by a pairwise ranking loss function. To improve the capability of sequential modeling, Zhang and Wang [11] proposed a recurrent neural network (RNN)-based model for relation classification. Other variants of RNN-based models have been proposed, such as Miwa et al. [14], who proposed a bidirectional tree structured LSTM model. Opposed to our work, none of the above approaches have leveraged document-specific information for relationship classification.

In the domain of scientific relation classification, Gu et al. [17] utilize a CNN-based model for identifying chemical-disease relationships from the abstracts of MEDLINE papers. Ammar et al. [13] enhanced Miwa and Bansal [14]’s relation extraction model via some extension such as utilizing gazetteer-like information extracted from Wikipedia. Hahn-Powell et al. [18] proposed an LSTM-based RNN model for classifying *causal precedence* relationships between two event mentions in biomedical papers. In [18], researchers take the surrounding context of a target event pair as a source of background information for relation classification. Unlike their approaches, which only utilize surrounding word or sentences, in our work, we search for useful background information from the entire document.

3 Baseline Models

Relation classification has significantly benefited from neural networks due to their high performance and less need for feature engineering. There are two main NNs architectures [19]: recurrent neural networks (RNN) [20] and convolutional neural network (CNN) [21]. RNN-based framework [11,14] and CNN-based framework [10,12] have been proved to be effective on relation classification. The main focus of this paper is to validate the effectiveness of document-specific information for scientific relation classification, not primarily to optimize the NNs learning architecture. We thus implement two relatively simple but strong

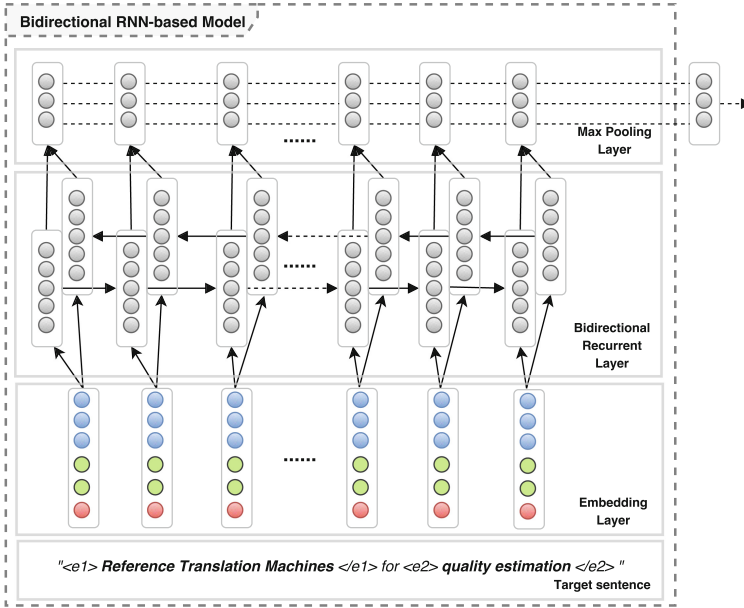


Fig. 1. Architecture of RNN-based baseline model.

and representative neural network models, from each of the two NNs architecture. By comparing their performance on RANIS corpus [1], we will select the base model that will be extended by the document-specific information.

3.1 RNN-Based Baseline Model

The first candidate baseline model is proposed by Zhang and Wang [11]. As shown in Fig. 1, it mainly consists of three layers.

The first layer is an embedding layer. This layer maps each word of the target sentence into a low dimensional word vector representation. For indicating the position of target entities, the model wraps the target entity pair with the special symbols $\langle e1 \rangle$, $\langle /e1 \rangle$, $\langle e2 \rangle$ and $\langle /e2 \rangle$. These special symbols are treated as normal words, and their embeddings will be learned during training. The embedding layer is calculated via Eq. (3), where W_{emb}^w is a word embedding projection matrix, and W_{emb}^{et} entity type projection matrix, x_t^w one-hot word representation, x_t^{et} one-hot entity type representation.

$$e_t^w = W_{emb}^w x_t^w \tag{1}$$

$$e_t^{et} = W_{emb}^{et} x_t^{et} \tag{2}$$

$$e_t = concat(e_t^w, e_t^{et}) \tag{3}$$

The second layer, or the hidden layer, is a bidirectional recurrent layer, which extracts word-level features from each word embedding. The hidden layer is calculated via Eqs. (4), (5), and (6), where h_t is the final output of the bidirectional RNN at t -th time step, e_t is the word vector from previous layer, h_t^{fw} (h_t^{bw}) is the output of forward (backward) RNN at the t -th step, W_{fw} , W_{bw} , U_{fw} , U_{bw} , b_{fw} and b_{bw} are model parameters.

$$h_t^{fw} = \tanh(W_{fw}e_t + U_{fw}h_{t-1}^{fw} + b_{fw}) \quad (4)$$

$$h_t^{bw} = \tanh(W_{bw}e_t + U_{bw}h_{t+1}^{bw} + b_{bw}) \quad (5)$$

$$h_t = h_t^{fw} + h_t^{bw} \quad (6)$$

The third layer is a max pooling layer, which selects the maximum value from each dimension of the outputs hidden layers, and merges them into a sentence-level feature. Finally, the sentence-level feature vector will be fed into a logistic regression model based relation classifier. See the original paper for further details.

3.2 CNN-Based Baseline Model

The second candidate baseline model is proposed by Santos et al. [12]. As shown in Fig. 2, it is also composed of three layers.

Similar to the RNN-based model, the first layer creates word representation for each word in a target sentence. This layer concatenates word embedding, entity type embedding and word position embedding to create the final word representation $[e_t^w, e_t^{et}, e_t^{wp}]$, where e_t^{wp} is the word position embedding. The position vector encodes the relative distance between the current word and the head of target entity pair. For instance, in Fig. 2, the position vector of word *for* in the target sentence is $[-1, 2]$. These two relative positions will be encoded into a position vector via the Eq. (7), where W_{emb}^{wp} is word position embedding projection matrix, and x_t^{wp} is one-hot representation of the relative distance.

$$e_t^{wp} = W_{emb}^{wp} x_t^{wp} \quad (7)$$

The next layer is a convolutional layer, which generates a distributed convolutional window level feature. The third layer is a max pooling layer, which chooses the maximum value from each dimension of convolutional window level feature and merges them as the ultimate sentence level feature.

Finally, the model predicts semantic relationship between target terms, by computing the score for a class label $c \in C$ via dot product:

$$S_\theta(x)_c = r_x^T [W^{class}]_c \quad (8)$$

where C is a set of semantic relationships, r_x is the sentence level feature vector and W^{class} is the class embedding matrix. The column of W^{class} represents the distributed vector representation of different class labels. It is worth pointing out that the model uses a logistic loss function below:

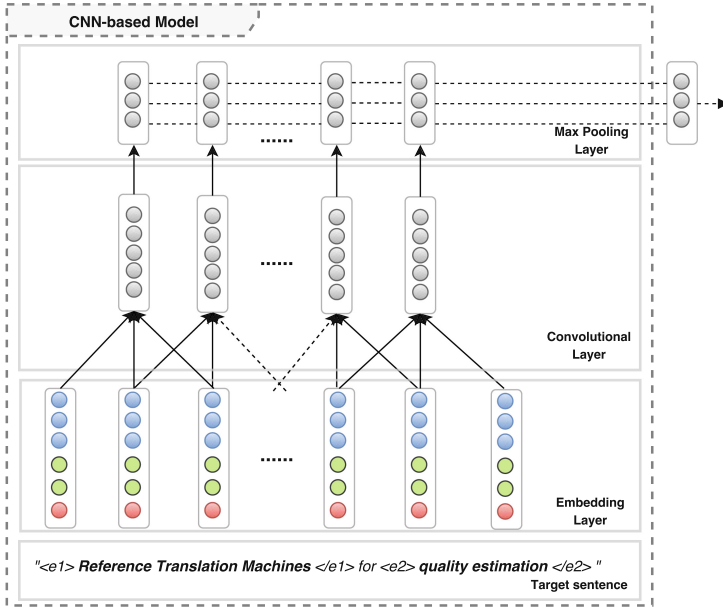


Fig. 2. Architecture of CNN-based baseline model.

$$L = \log(1 + \exp(\gamma(m^+ - s_\theta(x)_{y^+}))) + \log(1 + \exp(\gamma(m^- + s_\theta(x)_{c^-}))) \quad (9)$$

where, $s_\theta(x)_{y^+}$ is the score of correct class label, $s_\theta(x)_{c^-}$ is the score of the most competitive incorrect class label, m^+ and m^- are margins, and γ is a scaling factor. In our experiment, we use $m^+ = 2.5$, $m^- = 0.5$ and $\gamma = 2$. See the original paper for further details.

4 Proposed Model

In this paper, we assume that scientific papers could be utilized as a source of document-specific information to facilitate relation classification in scientific document. Therefore, we propose a new relation classification model that, on the one hand, could extract feature from target sentence like the baseline model does, on the other hand, it could automatically acquire the document-specific information, as illustrated in Sect. 1, from scientific papers.

We test our hypothesis by the annotated corpus of Tateishi et al. [1], which is created by CS paper *abstracts* with scientific semantic relationship annotation. We attribute this to the fact that the annotation scheme of RANIS corpus is based on CS related concepts [1], such as “Input” and “computational model”, additionally the background information about these CS related concepts are described in CS papers [1] like that “X is a computational model for Y”. Therefore we consider a problem setting like that we treat the annotated sentence in the abstract of paper as the target sentence, and the unannotated body of

paper (henceforth, *paper body*) as the source document-specific information, and we assume that the document-specific information extracted from paper body could facilitate relation classification in paper abstract. We believe that our method can be easily adapted to a more general task setting, e.g. analyzing semantic relationship in a whole document (not just in an abstract) via considering the entire document as a source of background information. To create such a framework, we need to address the following challenges:

1. How to automatically extract the document-specific information from a unannotated paper body?
2. How to encode the extracted document-specific information into a vector representation for relationship classification?

This paper proposes several automatic approaches to extract document-specific information, which are described in the rest of this section.

4.1 Retrieving Document-Specific Information

For the first question, we propose three methods.

Method 1: treats all the paper body sentences that contains a target entity as the representation of document-specific information, which is called *term sentence* (TS) henceforth. For example, given a target entity *RTM*, we would find the following TS in its corresponding paper body:

- (3) *RTM is a computational model for identifying the acts of translating between any given two data sets with respect to a reference corpus in the same domain.*

Given multiple TSs for a target entity, this method simply concatenates all of them into a long sequence of words as the final TS and feed it into our proposed NNs model. The intuition behind the method is that TS could contain document-specific background information about target entity, for example, TS (3) explains the type of target entity *RTM* as “*a computational model*”.

Method 2: uses Lexical Chain (LC) of content words (i.e., nouns, verbs and adjectives) as the representation of document-specific information. This approach is inspired by the usage of LC on word sense disambiguation [22]. LC is a sequence of semantically related words [23]. In this work, we define it as the sequences of contents words from paper body that are semantically close to a given target entity. For instance, the two chain of words listed below, are practical case of LC extracted from a paper body for the two target entities, *algorithm* and *research*, respectively.

LC1: *algorithm* → *probabilistic* → *parameter* → *method* → *computational* → *disambiguation*

LC2: *research* → *study* → *analysis* → *experiments* → *science* → *biomedical*

The process of creating LC in this work is similar to the approach proposed by [24]. Specifically, based on word embedding, we calculate cosine distance between a given target entity and each content word in paper body, and then use a predefined criteria to select the member for LC. We manually set the criteria cosine distance as 0.65, and only recruit the word whose cosine distance to the target entity is lower than this criteria as the member of the LC for the target entity, and put these recruited words in cosine distance increasing order.

About the significance of LC, it could be understood from the following two short texts. The specific entity type of “*drink*” in text1 is different with the one in text2, and the difference could be illustrated by creating LC of “*drink*” from each text, which is denoted in parenthesis.

text1: *Tony Blair’s favorite **drink** is tea, he prefers Oolong.*

(**drink**→tea→Oolong)

text2: *Tony Blair’s favorite **drink** is alcohol, he prefers wine.*

(**drink**→alcohol→wine)

Based on the LC, we can partly specify the entity type information about the target entity, for instance, according to LC in text1, “*drink*” refers to a type of “*tea*” called “*Oolong*”, but in text2, the specific entity type of “*drink*” is “*wine*”.

Since entity type information closely interact with relation classification [14, 25], we assume that LC could illustrate the entity type information about target entity, thereby facilitate relation classification.

Method 3: uses Semantically Related Sentence (SRS) as a representation of document-specific information. SRS is defined as that, among all sentences in paper body, its semantic representation (or vector representation) is most similar to the given target entity pair.

The semantic representation of a sentence is calculated by averaging weighted word vector of each word in the sentence, specifically by the equation $\frac{1}{n}(\sum_{w \in s} \frac{a}{p(w) + a} v_w)$ proposed by [26]. n is the number of word in the sentence, v_w is the vector of word, a is a hyper parameter, here, we set it as $a = 1$. $p(w)$ stands for the probability or frequency of the word w in a given paper body. Note that for more frequent word w , the weight $\frac{a}{p(w) + a}$ become lower, so that the vector representation of sentence will keep more of sentence-specific local meaning.

We explain the motivation of using SRS by the two sample sentences listed below, which are target sentence and SRS respectively.

Target Sentence: *We introduce referential translation machines(RTM_A) for quality estimation_B of translation outputs.*

SRS: *RTM can be used for predicting the quality of translation outputs.*

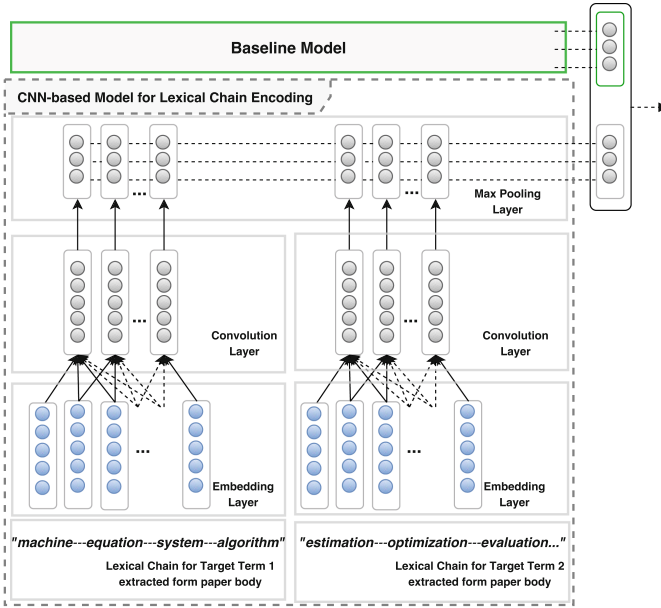


Fig. 3. The architecture of proposed model enhanced by LC or TS encoding.

The target Sentence does not literally express that *RTM* is applied for *quality estimation*, but the extracted SRS represents this meaning directly by the expression “*be used for*”. As we know, sentences in abstract tend to spare some details and sentences in paper body could supplement the omitted details. Therefore, the main purpose of extracting SRS from paper body is to complement and interpret the omitted details of target sentence in paper abstract, so that system will be easier to identify the relationship.

4.2 Architecture

After figuring out how to extract document-specific information, the next question is how to encode those extracted document-specific information into vector representation. As shown in Figs. 3 and 4, we propose two CNN based frameworks to encode document-specific information. It can be seen that they are structurally very similar to CNN-based baseline model. The reason for choosing CNN over RNN based model is because, empirically, the CNN-based model performs much higher than RNN-based model on our specific task, which will be explained in next section.

The framework in Fig. 3 is used for encoding LC pair (or TS pair), therefore it has a parallel structure, but with shared parameters such as word embedding matrix. As shown in Fig. 3, the model consists of 3 components: the word embedding layer, the convolutional layer and the last one, the max pooling layer that chooses maximum value across LC pair (or TS pair), and merge them into

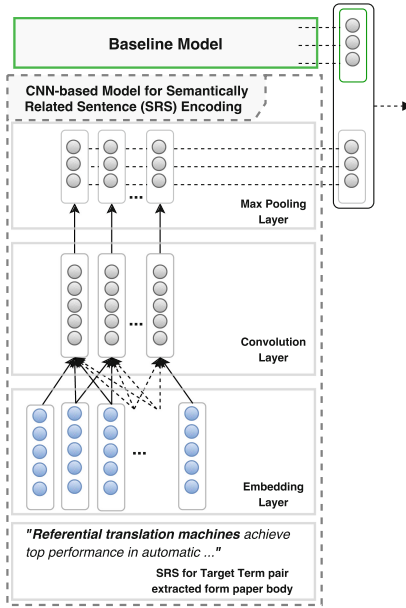


Fig. 4. The architecture of proposed model enhanced by SRS encoding.

one vector representation. Finally, this final representation will be concatenated with the final one of baseline model and be provided to relationship classification. Similar to the framework in Fig. 3, the model proposed in Fig. 4, also contains three layers, embedding layer, convolutional layer, max pooling layer. The only difference is the non-parallel structure because its input is a single SRS. In addition, both baseline model and the extension model shares the same word embedding matrix.

Generally, the whole model contains two main parts: the part of baseline model and the part of extension model, which have been discussed before. The baseline model aims at extracting feature from target sentence, and the extension model is responsible for collecting document-specific feature from paper body. This architecture illustrates the thinking that, to analyze scientific relationship, it would be helpful to comprehensively consider both the target sentence and the document-specific background information that is depicted in the document where the target sentence is placed. We use f-score based early-stopping by validation data, and apply the back-propagation algorithm to train the whole model, and choose the logistic loss function proposed by [12] as the objective function of the whole model. In our experiments, we implement the baseline model, the proposed extension model and the back-propagation algorithm by using Theano [27].

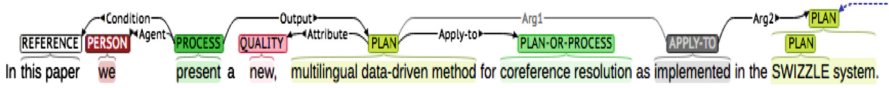


Fig. 5. Anotation example shown in brat rapid annotation tool

5 Experiments

5.1 Dataset

We use Tateisi et al. [1]’s RANIS corpus, which is a set of paper abstracts in Computer Science domain. In the dataset, the type of entity (called Entity Type (ET) hereafter), such as “QUALITY”, “PROCESS” and “DATA-ITEM” as well as the domain specific scientific relationships among them such as *INPUT*, *OUTPUT* and *APPLY_TO* have already been annotated, as shown in Fig. 5, based on the annotation scheme proposed by [1].

The RANIS corpus contains 250 abstracts collected from ACL Anthology (230 abstracts in the development set and 20 abstracts in the testing set) and 150 abstracts from ACM Digital Library. In this experiment, we only used the 250 abstracts from ACL Anthology and collected the corresponding paper bodies from ACL Anthology Reference Corpus [28].

In the RANIS corpus, there are 22 classes of semantic relations (e.g. *Input*, *Is-a*, *Apply-to*, *In_Out*). Since each relation has its direction, we add the direction to the relation names, such as *L-Input*, *R-Input*. This yields $22 \times 2 = 44$ classes. For the further details such as definition and difference of relationship etc., see the original paper [1].

5.2 Setup

We extract 12,807 relations from the ACL development set and 1,264 relations from the ACL testing set. From the development set, we randomly select 90% of samples as training data and the rest as validation data for tuning hyper parameter such as the number of epoch in early-stopping, learning rate etc. In Table 1, we show the selected hyper parameter values.

It has been showed that pre-trained word embedding can improve the training of relation extraction model [10]. Therefore, we initialized the word embedding layer with the pre-trained Google News word2vec embedding.³

5.3 Results and Discussion

Table 2 represents the performance of the two implemented baseline models, with the contribution of semantic information extension including the Entity Type Information, and part of speech information⁴. It can be observed that the

³ <https://code.google.com/archive/p/word2vec/>.

⁴ We used the Stanford CoreNLP. <https://stanfordnlp.github.io/CoreNLP/>.

Table 1. Selected hyper parameters

Parameter name	Value
Word Emb. size	300
Word POS Emb. size	50
Word Entity Type Emb. size	50
Word Position Emb. size	100
Convolutional Units (Baseline model)	1000
Context Window size (Baseline model)	3
Convolutional Units (Extension model)	200
Context Window size (Extension model)	5
Maximum Epoch	20
Learning Rate	0.01

CNN-based baseline model performs much better than the RNN-based baseline model in this specific task. It is known that, RNN is used for sequence modeling and CNN is used for detecting some key phrases [19]. The better performance of CNN-based baseline model indicates one property of the RANIS corpus that comparing to considering the whole sequence of words in a target sentence, just focusing on some key words is more effective to identify the semantic relationship in this dataset. For example, in the target sentence (4),

- (1) *In this paper we present a new, multilingual data-driven method_A for coreference resolution_B as implemented in the SWIZZLE system.*

for identifying the semantic relationship, *Apply-to*, between “*multilingual data-driven method*” and “*coreference resolution*”, it is unnecessary to considering the whole sequence of words in the sentence like “*In this paper, we present ...*” etc. In contrast, it would be more effective to concentrate on some key words like “*... for ...*” in this sentence. This result makes us to choose the CNN-based model as the baseline model for our task, and it also inspires us to apply CNN-based architecture for encoding document-specific information into vector representation.

This table also indicates that the extension of semantic information especially the entity types information significantly contributes to the improvement of performance for both baseline models. This can be attributed to the interdependency between relationship type and entity type [14, 25]. For instance, according to the definition of relationship by [1], “*Apply-to(A, B) means the Method A apply to purpose B*”, if one target entity belongs to Method-like entity type like *algorithm* in computer science, it would have high tendency to participate in *Apply-to* relationship.

To gain more understanding towards the disadvantage of baseline model and improve its performance, we did manual error analysis in 5 randomly selected

Table 2. Comparison of Baseline models

Model	Precision	Recall	F-score
RNN-based Baseline Model	54.71	54.78	50.90
RNN-based Baseline Model + POS	57.54	56.49	53.12
RNN-based Baseline Model + Entity Type	69.01	64.12	61.89
RNN-based Baseline Model + POS + Entity Type	66.38	64.52	61.24
CNN-based Baseline Model	61.78	60.21	59.54
CNN-based Baseline Model + POS	66.10	61.08	62.21
CNN-based Baseline Model + Entity Type (Baseline)	72.83	72.47	72.20
CNN-based Baseline Model + POS + Entity Type	73.40	71.32	71.39

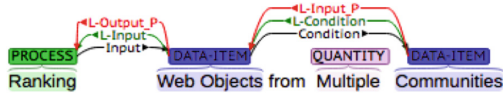


Fig. 6. Relationship identification error from baseline model, where red line indicates the error from baseline model, green and black shows correct one, the only difference between them is the green one has directional mark "L-" (means left). (Color figure online)

abstracts from the result of the baseline model like Fig. 6. Through the analysis, we found that the majority of mistake happens between definitionally similar relationships, for instance between *Input* and *Output*, *Input* and *In-Out*, *Apply-to* and *Input*, *Attribute* and *Condition*. We assume that these mistakes could be partially fixed by the document-specific information from paper body. For instance, the syntactic pattern “*use ... for ...*” could represent distinct relationship. In target sentence “*use X_A for training_B*”, the relation between “ X_A ” and “*training_B*” could be *Input* or *Apply-to*. If “ X_A ” is a *method* or *algorithm*, it will be *Apply-to* relation, but if “ X_A ” is *training data*, the relation will be *Input*, because *training data* is not used as a *method*, but as *input data*. Therefore, just using the given target sentence, without further specification via document-specific information like “ X_A is an algorithm built upon ...”, the model would be easy to confuse between *Apply-to* and *Input*. In the next experiment, we evaluate the effect of paper body on classifying relationships, especially on these definitional similar relationships.

We have argued that paper body could be used as a knowledge base for scientific relationship classification. To verify this arguments, we compare the performance of the proposed extension, TS, LC and SRS with the baseline model. To minimize the affect of random initialization of neural network on result, we run the experiment of evaluation for each method 5 times and take their average value for the comparison. Table 3 represents the result of each extension. It can be seen that the extension of SRS and LC get better performance than baseline model in all metrics, especially the extension of LC, which achieved the

Table 3. Impact of proposed method on overall performance (mean value \pm standard deviation)

Model	Precision	Recall	F-score
CNN-based Baseline Model + Entity Type (Baseline)	73.14 \pm 0.47	72.63 \pm 0.40	72.27 \pm 0.27
Baseline + POS + TS (Method 1)	73.86 \pm 0.68	70.68 \pm 1.22	71.15 \pm 0.97
Baseline + POS + SRS (Method 3)	74.51 \pm 0.77	73.51 \pm 0.56	73.33 \pm 0.68
Baseline + POS + LC (Method 2)	75.23 \pm 0.18	73.72 \pm 0.24	73.81 \pm 0.26

highest scores. The result shows that firstly paper body is an useful resource of background information for scientific relationship classification. Secondly, SRS and LC could be an effective approach to extract document-specific background information for scientific relationship classification. However, TS didn't increase the performance (on Recall and F-score), one of possible explanation for this phenomenon is that although paper body contains background information, it does not necessarily mean all part of the paper body is informative to scientific relationship classification. For instance, as in sentence (3), some words like “*with respect to a reference corpus . . .*” is semantically irrelevant with the target term, even they exit in a same sentence, therefore TS needs further semantic pruning. From the perspective learning architecture, the result also shows that CNN-based model is not only effective to encode target sentence, but also useful for encoding document-specific information extracted from paper body.

As discussed before, the baseline model tends to confuse among particular relationships includes *Input*, *Output*, and so on. We assumed that the paper body based extension especially LC could improve the systems performance on theses types of relationships. As we can see in Table 4, the LC outperforms the baseline model on almost all of these relationships except *In_Out* relationship. This comparison once again indicates the effectiveness of paper body, specifically the LC from paper body, on relationship classification in scientific document.

Table 4. Performance on selected types of relationship

Relationship	Baseline	Baseline + POS + LC
<i>Input</i>	52.58 \pm 1.18	56.12 \pm 2.39
<i>Output</i>	68.56 \pm 2.03	68.74 \pm 2.12
<i>In_Out</i>	54.72 \pm 3.51	54.16 \pm 2.39
<i>Target</i>	25.10 \pm 3.81	29.80 \pm 2.07
<i>Apply-to</i>	79.52 \pm 2.05	81.76 \pm 3.13
<i>Condition</i>	49.44 \pm 1.78	49.72 \pm 2.07
<i>Attribute</i>	86.26 \pm 0.52	86.64 \pm 0.24

6 Conclusion

In this work, we tackle the relationship classification task in scientific documents by extending baseline models via document-specific information extracted from paper body. Compared to the baseline models, the extend model not only collects feature from target sentence like most base models do, but also extracts document-specific background information from paper body. To extract the document-specific information from paper body, two powerful approaches, SRS and LC are proposed. Experimental result on RANIS corpus demonstrated that paper body could be used as a source of background knowledge for scientific relationship classification, and the proposed SRS and LC based model could be effective approaches for extracting document-specific information from paper body. However, the experimental result also shows that the proposed model achieved small progress on some types of relationship such as *Output* and *Condition*. We believe, this will be solved by refining the current approach such as by improving sentence vector calculation and lexical chain construction, and it will be left as future work.

Acknowledgement. This work was supported by JST CREST Grant Number JPMJCR1513, Japan and KAKENHI Grant Number 16H06614.

References

1. Tateisi, Y., Shidahara, Y., Miyao, Y., Aizawa, A.: Annotation of computer science papers for semantic relation extraction. In: LREC, pp. 1423–1429 (2014)
2. Teufel, S., et al.: Argumentative zoning: information extraction from scientific text. Ph.D. thesis, University of Edinburgh (2000)
3. Cohen, K.B., Demner-Fushman, D., Ananiadou, S., Tsujii, J.: BioNLP 2017 (2017)
4. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 Task 10: ScienceIE-extracting keyphrases and relations from scientific publications. arXiv preprint [arXiv:1704.02853](https://arxiv.org/abs/1704.02853) (2017)
5. Kumar, S.: A survey of deep learning methods for relation extraction. arXiv preprint [arXiv:1705.03645](https://arxiv.org/abs/1705.03645) (2017)
6. Zhou, D., Zhong, D., He, Y.: Biomedical relation extraction: from binary to complex. *Comput. Math. Methods Med.* **2014**, 18 p. (2014). <https://doi.org/10.1155/2014/298473>. Article ID 298473
7. Boschee, E., Weischedel, R., Zamanian, A.: Automatic information extraction. In: Proceedings of the International Conference on Intelligence Analysis, vol. 71. Cite-seer (2005)
8. Suchanek, F.M., Ifrim, G., Weikum, G.: Combining linguistic and statistical analysis to extract relations from web documents. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 712–717. ACM (2006)
9. Chan, Y.S., Roth, D.: Exploiting background knowledge for relation extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 152–160. Association for Computational Linguistics (2010)
10. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al.: Relation classification via convolutional deep neural network. In: COLING, pp. 2335–2344 (2014)

11. Zhang, D., Wang, D.: Relation classification via recurrent neural network. arXiv preprint [arXiv:1508.01006](https://arxiv.org/abs/1508.01006) (2015)
12. dos Santos, C.N., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. arXiv preprint [arXiv:1504.06580](https://arxiv.org/abs/1504.06580) (2015)
13. Ammar, W., Peters, M., Bhagavatula, C., Power, R.: The AI2 system at SemEval-2017 Task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 592–596 (2017)
14. Miwa, M., Bansal, M.: End-to-end relation extraction using LSTMs on sequences and tree structures. arXiv preprint [arXiv:1601.00770](https://arxiv.org/abs/1601.00770) (2016)
15. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 427–434. Association for Computational Linguistics (2005)
16. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
17. Gu, J., Sun, F., Qian, L., Zhou, G.: Chemical-induced disease relation extraction via convolutional neural network. Database **2017**, bax024 (2017). <https://doi.org/10.1093/database/bax024>
18. Hahn-Powell, G., Bell, D., Valenzuela-Escárcega, M.A., Surdeanu, M.: This before that: causal precedence in the biomedical domain. arXiv preprint [arXiv:1606.08089](https://arxiv.org/abs/1606.08089) (2016)
19. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint [arXiv:1702.01923](https://arxiv.org/abs/1702.01923) (2017)
20. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
22. Galley, M., McKeown, K.: Improving word sense disambiguation in lexical chaining (2003)
23. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Comput. Linguisti. **17**(1), 21–48 (1991)
24. Mascarell, L.: Lexical chains meet word embeddings in document-level statistical machine translation. In: Proceedings of the Third Workshop on Discourse in Machine Translation, pp. 99–109 (2017)
25. Wang, T., Li, Y., Bontcheva, K., Cunningham, H., Wang, J.: Automatic extraction of hierarchical relations from text. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 215–229. Springer, Heidelberg (2006). https://doi.org/10.1007/11762256_18
26. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings (2016)
27. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a CPU and GPU math compiler in Python. In: Proceedings of the 9th Python in Science Conference, pp. 1–7 (2010)
28. Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M.Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F., et al.: The ACL anthology reference corpus: a reference dataset for bibliographic research in computational linguistics. In: LREC (2008)

kNeXI2017

Knowledge Explication for Industry (kNeXI 2017)

Satoshi Nishimura

National Institute of Advanced Industrial Science and Technology
satoshi.nishimura@aist.go.jp

1 The Workshop

The 1st International Workshop on Knowledge Explication for Industry (kNeXI 2017) was successfully held on November 14–15, 2017 at University of Tsukuba, Tokyo Campus in Tokyo, Japan. This was held as a workshop of the 9th JSAI International Symposia on AI (JSAI-isAI 2017), sponsored by the Japanese Society for Artificial Intelligence (JSAI).

The aim of kNeXI is to create an opportunity to discuss about knowledge in the industry with interdisciplinary researchers and practitioners. The main themes of this workshop are following two: 1. Enhancement of awareness, decision-making, teamwork based on shared knowledge 2. Automation of simple work based on shared knowledge.

We had four keynotes and nine oral presentations in this workshop. All presentations are relevant to knowledge building or its use for empowerment of human ability. Specialties of the presenters are quite interdisciplinary from elderly care services, agriculture, sustainability science and life, higher education, brain science, and art. The commonality of these domain and their focus is “knowledge” which the key person has, in each domain.

2 Papers

The organizers selected three papers in the kNeXI part of the present volume. The first paper is “Investigating Classroom Activities in English Conversation Lessons based on Activity Coding and Data Visualization” which is authored by Zilu Liang, Satoshi Nishimura, Takuichi Nishimura, Mario Alberto Chapa-Martell. They propose data collection, processing and visualization method for reflection of teacher and students’ activity in English conversation lessons. The data processing is based on the activity coding list which is constructed from the collected data in the lessons. The paper includes meaningful case information to apply the proposed method.

The second paper is “On-site Knowledge Representation Tool for Employee-Driven Innovation in Services” which is authored by Kentaro Watanabe. He introduces that knowledge of service employees contributes to service productivity and customer satisfaction in the service fields with frequent interactions among customers and employees, such as care services. This paper proposes brand-new knowledge

representation tool, called DRAW2. The DRAW2 tool was tested in elderly care facility and it works well. Watanabe also provides comparison among knowledge representation tools at work. Such discussion is fruitful for the researchers who will develop the tools using at workplaces.

The third paper is “Consideration of application cases of structured manual and its utilization” is authored by Satoshi Nishimura, Ken Fukuda, Takuichi Nishimura. The paper provides the consideration of application cases of knowledge explication which was provided by the authors. They analyse the effects related to knowledge explication and found five different effects are occurred before and after knowledge explication. The comparison among application cases are provided after the analysis. They employed the insight from service science to compare each case. They also point out their future work about knowledge explication and its use at the last of the paper.

Acknowledgement. As the organizing committee chair, I would like to thank the program committee members, Dr. Takuichi Nishimura, Dr. Ken Fukuda, Dr. Kentaro Watanabe, Dr. Yasuyuki Yoshida, Ms. Nami Iino. The organizers would like to thank Mr. Kazuo Barada, Mr. Yuki Koyama for their support. The organizers also would like to thank to keynote speakers, Prof. Riichiro Mizoguchi, Japan Advanced Institute of Science and Technology, Mr. Hiroshi Ohtani, Medical Corporation Hanamaru-kai, Prof. Kristiina Jokinen, National Institute of Advanced Industrial Science and Technology, Prof. Hideaki Takeda, National Institute of Informatics. The workshop is based partly on results obtained from the “Future AI and Robot Technology Research and Development Project” commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and I also would like to thank to the organizers of International Symposia on Artificial Intelligence for their great support and giving the opportunity to hold the workshop.



Investigating Classroom Activities in English Conversation Lessons Based on Activity Coding and Data Visualization

Zilu Liang¹(✉), Satoshi Nishimura², Takuichi Nishimura²,
and Mario Alberto Chapa-Martell³

¹ Faculty of Engineering, The University of Tokyo, Tokyo, Japan
z.liang@cml.t.u-tokyo.ac.jp

² AI Research Centre, National Institute of Advanced Science and Technology, Tokyo, Japan
{satoshi.nishimura,takuichi.nishimura}@aist.go.jp

³ CAC Corporation, Tokyo, Japan

Abstract. Reflective teaching has become dominant paradigm in second language teacher education, as critical reflection helps teachers achieve a better understanding of teaching and learning processes. Critical reflection begins from classroom investigation. Several methods such as questionnaire, lesson report, teaching journal and audio/video recordings are widely used for classroom investigation. However, these methods are either susceptible to memory bias or are hard to be continued on a day-to-day basis. In this study, we proposed an approach for effective classroom investigation in second language education using activity coding in combination with data visualization technology. The proposed method consists of three stages. In the first stage, a smartphone application was used to record the activities that happen in a class following a slightly modified experience sampling method. In the second stage, the activities were quantified using our proposed 2-level activity coding scheme, and each activity was assigned a colour code. In the third stage, a data visualization tool D3.js was used to create heat maps of the classroom activities. We applied the proposed method to investigating the classroom activities in five English conversation lessons given by native-speaking teachers. The visual feedback led to the answering of some key questions that critical reflection aims to address, including teachers' time management, lesson structure, and the characteristics of teacher-student interaction. Based on the results obtained, we highlighted the potential of the proposed approach for involving different sectors in second language education and pointed out the directions for future research.

Keywords: Second language education · Data visualization · D3.js
Experience sampling method · Reflective teaching

1 Introduction

Nurturing the skills of critical reflection has become the underpinnings of most second language teacher education program worldwide [1, 2]. Previous research indicated that

critical reflection makes a significant impact on a teacher's knowledge, skills and effectiveness [3–5]. Critical reflection helps teachers understand their teaching practices and enables them to look at the underlying principles and beliefs that define the way they teach [6].

Critical reflection begins in gathering information about what happens in the class. Keeping tracking and reflecting on classroom activities can help teachers develop deeper understanding of their teaching practices and strategies. A number of approaches have been employed for classroom investigation. Surveys and questionnaires collect information on a particular aspect of teaching but tell little about the actual teaching procedures [6]. Teachers' diary records the routine and conscious actions in the classroom for later reflection but is susceptible to memory bias. Previous studies suggest that people are not good at reconstructing their experience after the fact and they cannot provide reliable assessment of complex dimensions of their experiences [7, 8]. Peer observation and audio/video recording capture the details of classroom activities. However, these approaches generate large amount of qualitative data. Reviewing these data requires a lot of time and therefore can never become an activity continued on a day-to-day basis [9].

In this study, we proposed a novel approach for classroom investigation using activity coding and data visualization technology. The proposed approach enables the recording of classroom activities on the spot to address the problem of memory bias in existing approaches. Additionally, data visualization technology helps to generate visual feedback that leads to insights on the teaching procedures. The proposed method consists of three stages: data collection, data processing, and data visualization. In the first stage, a smartphone application named aTimeLogger [10] was used to record all activities that occur in a class following a slightly modified procedure of experience sampling method. In the second stage, a 2-level activity coding scheme that we proposed is used to map each recorded classroom activities to a corresponding color code. This stage is essential for converting qualitative description of classroom activity to quantitative color code. In the third stage, the sequence of color code generated in the previous stage is used as input to render a heat map [11] of the classroom activities using a JavaScript library named D3.js.

Employing the proposed approach, we investigated five English conversation classes given by native-speaking teachers. We plotted heat maps of the classroom activities recorded in the five classes and calculated several metrics that characterizes the teaching procedures. The visualization and quantitative results helped us answer important questions in critical reflection such as “was the class teacher dominated?”, “how was teacher's time management?”, “how did teacher and students interact?”. Acknowledging the limitations of the proposed approach, we highlighted the opportunities for future research.

Note that we do not intend to replace existing approaches with the proposed one. Instead, we suggest that the proposed approach can be used in combination with existing approaches to create synergistic effect. This approach has many potential applications to involve different sectors in second language education. It empowers student teachers as well as in-service teachers to move beyond the level of routinized responses to classroom situations or to investigate the effect of changes in their teaching. Educational

institutions can apply this approach in their teacher education programs. Education authorities can employ and extend this approach to quantitatively compare and evaluate teaching practices.

2 Related Work

Critical reflection on ones' own teaching procedures help teachers achieve a better understanding of teaching and learning processes [6]. Therefore, reflective teaching has become a dominant paradigm in second language teacher education in recent years [1, 2].

Reflective teaching commences when one investigates the activities that occur in class. A number of simple procedures are widely used for classroom investigation. Depending on when the information is collected, existing approaches for classroom investigation can be classified into two categories: after-event approaches and in-event approaches. Typical after-event approaches include surveys and questionnaires, lesson reports [12] and teaching journal (also called teachers' diary) [13]. Typical in-event approaches include peer observation [14, 15] and audio/video recording [6].

Surveys and questionnaires are often used at the end of a class or a semester to gain general impression on certain aspects of teaching. Though this approach has been widely used to obtain feedback from students, it collects little information on the actual teaching procedures and activities in class [6]. Alternatively, teachers can use a diary to describe their recollections of what happened during a lesson for later reflection [13]. However, this approach dominantly relies on teachers' recalling on the activities occurred in class and thus are susceptible to memory bias. Previous studies suggest that people are not good at reconstructing their experience after the fact and they cannot provide reliable assessment of complex dimensions of their experiences [7, 8]. In addition, this approach can capture only recollections and interpretations of the activities happened during a lesson but not the actual activities themselves.

In-event approaches such as peer observation and audio/video recording solve the problems of memory bias and enable the collection of detailed information on classroom activities. These approaches also have problems. For one thing, peer observation generates mostly qualitative data which is hard to analyze quantitatively. Comparing the observation notes of different lessons also imposes great challenges to teachers. For another, reviewing an audio or video recording requires large amount of time and therefore can never become an activity continued daily given the busy schedule of teachers [9].

In our study, we employed activity coding in combination with data visualization technology to address the limitations of existing approaches. Classroom activities are recorded in real time to avoid memory bias, and the captured data is coded and then visualized in heat maps that render intuitive visual cues for insights and for further analysis. We will present the proposal in detail in Sect. 3.

3 Proposed Method

The proposed approach follows a procedure that resembles individual's reflection on personal data [18]: data collection following experience sampling method, data processing for classroom activity coding, and data visualization of classroom activities using JavaScript library D3.js.

3.1 Data Collection: Recording Activities Using Experience Sampling Method

Critical reflection begins in gathering information on what happen in class. Therefore, the first stage of the proposed approach is to record all the classroom activities to form an archival information on the teaching procedure in a lesson. To avoid the problem of memory bias in existing approaches, we adopted a modified procedure of the Experience Sampling Method (ESM) [16, 17] to capture teaching as it is delivered. Given that it is not feasible for teachers to provide self-reports during their teaching, we created logs from a third person's point of view following the procedures recommended in [19–21]. This process can be done either on paper or digitally. In this study, we used a mobile time tracking app named aTimeLogger to log all the activities happened during a class, including the teacher's and the students' activities and the time stamp of those activities. The aTimeLogger was originally developed to help individuals manage their daily time usage and to enhance productivity. It was designed to record different type of activities. We used the activity type "work" and "study" to record the activities of teachers and students respectively.

Considering each activity as an event, we selected event-based sampling strategy [22] to record the classroom activities. An entry was created when a new activity started, and the time stamp was recorded as the start time of this activity. When this activity was completed, the observer simply tapped the stop button to record the end time. The aTimeLogger also allowed for writing comments for each event. Therefore the observer logged the content of the activities and students' reactions in detail. When the lesson finished, the recorded data was exported into a CSV file for further processing.

3.2 Data Processing: Classroom Activity Coding

In the CSV files exported from aTimeLogger app, each entry comprised the following information of an activity: activity type, duration of the activity in minutes, start time, end time, and observer's comments. All activities were listed in chronological order starting with the most recent one. A screenshot of an exported CSV file is shown in Fig. 1.

	A	B	C	D	E
	Activity type	Duration	From	To	Comment
1	Work	0:02	20:57:00	21:00:00	Gave homework. A writing task.
2	Study	0:10	20:48:00	20:56:00	Students playing 20 questions game. It was very fun but challenging.
3	Study	0:16	20:32:00	20:47:00	Group discussions on the travel packages that they created.
4	Work	0:01	20:31:00	20:31:00	Gave structures for giving a presentation and biz pitch.
5	Study	0:14	20:17:00	20:30:00	Group discussions to prepare travel packages. One group was very ac
6	Work	0:03	20:14:00	20:16:00	Explain pronunciation of some words. Humor, theme, thesis, genre.
7	Study	0:04	20:10:00	20:13:00	The whole class compared answers.
8	Study	0:03	20:07:00	20:09:00	Then compare answer in groups.
9	Study	0:03	20:04:00	20:06:00	Listening comprehension. All students requested to listen a second t
10	Study	0:12	19:52:00	20:03:00	Students made further preparation for the role play during the break
11	Study	0:05	19:47:00	19:51:00	Teacher suggested alternative ways to express the meaning more na
12	Study	0:03	19:44:00	19:46:00	Role play on seeking help in a hotel. "Open the key" -> open the doc
13	Work	0:08	19:37:00	19:43:00	Explain the difference between "would you" and "could you-do you
14	Study	0:07	19:30:00	19:36:00	In group discuss the problems encountered in a hotel and the phrase
15	Work	0:01	19:28:00	19:29:00	Matthew wrapped up the presentation and corrected mistakes.
16	Study	0:01	19:27:00	19:28:00	Correct pronunciation "walk" and "work".
17	Study	0:09	19:18:00	19:26:00	Group presentation. One group presented and the other group gave
18	Study	0:10	19:08:00	19:17:00	Matthew wrote down some questions about travel on the white board
19	Study	0:02	19:05:00	19:07:00	Matthew corrected grammar mistakes. Students tend to forget using
20	Study	0:08	18:57:00	19:04:00	Then the whole class compared answers. Student took turns to com
21	Study	0:05	18:51:00	18:56:00	Students were doing group discussions on their homework (exercise
22					
23					
24					

Fig. 1. A screenshot of the CSV file exported from aTimeLogger app. Each data entry comprised information of an activity occurred in the class, including activity type, time stamps and observer's comments.

The raw data was qualitative and thus was not suitable for further analysis. We converted the qualitative data to quantitative through activity coding. Activity coding starts with the creation of a list of initial codes of classroom activities such as "having group discussions", "giving presentations", "explaining new words and phrases" and "checking answers with the whole class". The activities occurred in second language classes are highly repetitive and share common features across classes. A list of common activities used in second language teaching was presented in [23]. However, this list served for a more general purpose in language teaching, which covered reading, writing, listening and grammar. Most of the activities mainly targeted at young learners rather than university students. Therefore, the activity coding list in [23] was not suitable for our purpose of understanding conversation classes for adults (i.e. university students and staff) in particular, and we thus developed our own activity list which is presented in Sect. 4.

3.3 Data Visualization: Creating Heat Map of Classroom Activities

After activity coding, the sequence of color code that corresponds to the whole set of activities occurred in a class was used to create a heat map using a JavaScript library D3.js [11]. A heat map of a class illustrates the sequence of classroom activities in chronological order using colored spectrum. The level-1 activity coding scheme maps the activities of students and teachers into red and blue segments on the colored spectrum, while the level-2 activity coding scheme further differentiate activities under each category using different shades of red or blue.

Base on the heat maps, we also deduced several quantitative features to characterized teaching styles, including the *Ratio of Students' Activities* (RSA), the *Ratio of*

Teachers' Activities (RTA), and the *Normalized Class Dynamics* (NCD). The metrics were calculated using Eqs. (1)–(3).

$$RSA = \frac{\sum_{i=1}^I t_i^{student}}{T} \quad (1)$$

$$RTA = \frac{\sum_{j=1}^J t_j^{teacher}}{T} \quad (2)$$

$$NCD = \frac{N_{student \leftrightarrow teacher}}{T} \quad (3)$$

where $t_i^{student}$ and $t_j^{teacher}$ represent the duration of the i -th students' activity and the duration of the j -th teachers' activity in minutes respectively. T is the total duration of the whole class in minutes. I and J are the total amount of students' and teachers' activities. $N_{student \leftrightarrow teacher}$ is the total number of switches between students' and teachers' activities.

4 Results

4.1 Data Collection and Processing

Using the proposed method, we collected and analyzed the classroom activities in five English conversation classes given by five native-speaking teachers. The first author presented at the classes and logged all the activities using aTimeLogger in a manner as unobtrusive as possible. The second and the third author also presented in part of the classes. The first author then discussed with the second and the fourth author to develop the activity coding list which is summarized in Table 1. In this study, we adopted a hybrid approach inspired by thematic analysis [24] to create the activity coding list. Firstly, we read through all the comments in the classroom observation logs to get familiar with the classroom activities occurred in the observed lessons. Second, we identified initial codes by repeatedly going through the comments. Routine activities such as “having group discussions”, “giving presentations”, “explaining tasks and taking questions”, “giving pronunciation drill”, “correcting grammar mistakes” were first extracted and then ranked according to their perceived educational benefits. Third, all the initial codes were assigned a numerical color code, i.e. red for students activities and blue for teachers activities.

Main activities are marked bold in the list. An activity was considered as main activity if its total duration last more than 10% of the total class time in at least one of the five classes observed. In line with the coding scheme of the d3.js library, the color codes assigned to students' activities were within [50, 100], and higher values corresponded to darker red. The color codes assigned to teachers' activities were within [0, 50], and lower values corresponded to darker blue. The visualization results are presented in following sections.

Table 1. Activity coding list.

Category (level-1 coding)	Activity name (level-2 coding)	Color code
Students' activities	#Having group discussions with teacher's interaction	100
	#Having pair discussions with teacher's interaction	97.5
	#Having group discussions without teacher's interaction	95
	#Having pair discussion without teacher's interaction	92.5
	#Having role plays or playing a game	90
	#Asking questions to teacher	85
	#Asking questions to presenter	82.5
	#Asking questions to other students	80
	#Having casual chat with teacher	77.5
	#Comparing answers in group	75
	#Comparing answers in pair	72.5
	#Giving a presentation in group	70
	#Giving a presentation alone	67.5
	#Brainstorming alone	65
	#Reading out	62.5
	#Listening to an audio and answering questions	60
	#Doing a quiz or an exercise	57.5
#Preparing for the classes	55	
#Having a break	50	
Teachers' activities	#Checking attendance	49
	#Assigning homework	48
	#Explaining tasks and taking questions	44
	#Explaining new words and phrases	40
	#Explaining grammar points	36
	#Wrapping up a discussion	32
	#Checking answers with the whole class	28
	#Asking questions to students individually	24
	#Sharing techniques and tips	20
	#Correcting grammar mistakes	16
	#Giving comments to the whole class	12
	#Asking follow-up questions	4
#Sharing personal/cultural experiences with students	0	

4.2 Heat Map Visualization of Classroom Activities

Heat maps of the five classes are shown in Fig. 2. The color and length of a segment indicate the type and length of an activity. Red segments represent student activities such as *having group discussions*, *giving presentations*, *comparing answers*. Blue segments represent teacher activities such as *explaining new words and phrases*, *explaining tasks*, *giving learning tips*. As shown in Fig. 2, the class of Teacher_3 had

more blue segments than red ones, indicating that the teacher was dominating the class and the students were not given sufficient time to talk. As a contrast, student activities consumed more time than teacher activities in the classes of Teacher_4 and Teacher_5. Especially, the class of Teacher_5 manifested good rhythm as the activities of students and teachers were almost uniformly distributed. This matches our observation that Teacher_5 had very good time management in his classes.

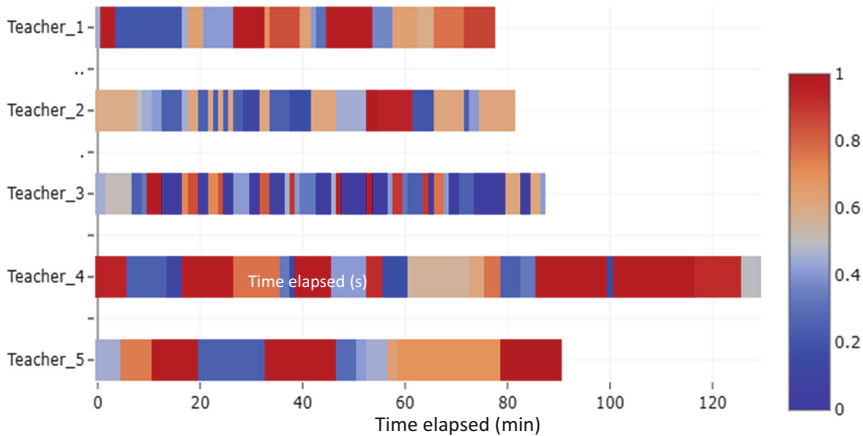


Fig. 2. Heat maps of the observed classes. (Color figure online)

The *Ratio of Students' Activities (RSA)* and the *Ratio of Teachers' Activities (RTA)* are plotted in Fig. 3. Given the objectives of the classes, it is expected that the teachers should not dominate the class and instead should give sufficient time for students to practice speaking. Figure 3 shows that the *RSA* of the five observed classes ranged from as low as 27.3% in the class of Teacher_3 to as high as 70.8% in the class of Teacher_4. The *Normalized Class Dynamics (NCD)* is plotted in Fig. 4. The class of Teacher_3 peaked in the plot with one activity switch between students and the teacher every 3 min. The *NCD* of Teacher_5 was the lowest with approximately one switch every 20 min. Classes of the other teachers have *NCD* in between.

Visualizing the nominal coding helped answer some of the key questions that critical reflection aims to address [6].

1. *How did the teacher structure the lesson?* Different teachers have their own ways of structuring a class. For example, Teacher_2 started her lesson with a quiz exercise, whereas other teachers started their lessons by explaining the tasks and taking questions from students. In addition, Teacher_2 integrated listening practice into the conversation class.

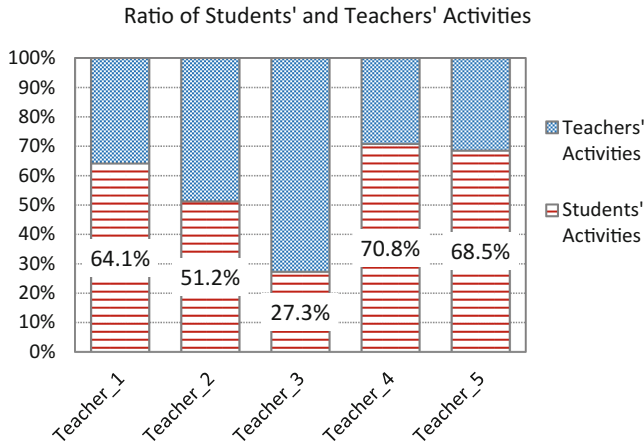


Fig. 3. The ratio of students' and teachers' activities in five classes.

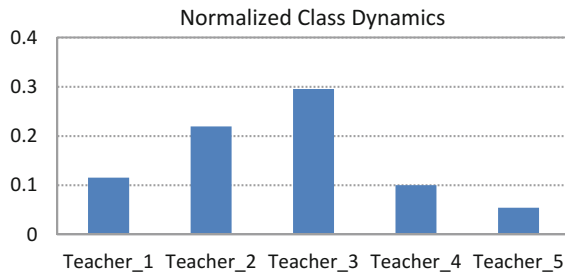


Fig. 4. Normalized frequency of switch between students' and teachers' activities.

- What kind of teacher-student interaction occurred?* The visualization of level-2 coding provided fine-grained information on the context in which teacher-student interaction happened. Teacher_2, Teacher_3 and Teacher_5 all interacted with students during pair discussions, whereas Teacher_1 and Teacher_4 observed students' group discussions but did not have verbal interaction with them. The pattern of teacher-student interaction in Teacher_3's class was characterized by teacher sharing personal experience and asking follow-up questions after the students summarized their group discussion. Indeed, our observation suggested that Teacher_3 asked many follow-up questions to help the students further develop their discussions. In four out of the five classes, teachers interacted with students by comparing answers with the whole class.
- What are the key activities?* Table 2 summarizes the top 3 activities in each class according to the total amount of time. The total amount of time spent on the top 3 activities occupied at least 40% of the total class time. Students having group discussion without teachers interaction was the dominant activity in the class of Teacher_1,

Teacher_4 and Teacher_5. Teacher_2 allocated abundant time for students to practice listening, while Teacher_3 spent much time asking follow-up questions to students.

Table 2. Top 3 activities in each class

Instructor	Top 3 activities	Ratio
Teacher_1	1. Students having group discussions without teacher's interaction	23.1%
	2. Teacher asking questions and students answer one by one	16.7%
	3. Students reading out	12.8%
Teacher_2	1. Students listening to an audio and answer questions	30.5%
	2. Teacher checking answers with the whole class	18.3%
	3. Teacher explaining tasks and took questions	11.0%
Teacher_3	1. Teacher asking follow-up questions	21.6%
	2. Teacher sharing personal/cultural experience with students	12.5%
	3. Teacher checking answers with the whole class	9.1%
Teacher_4	1. Students having group discussions without teacher's interaction	36.2%
	2. Students having role play/game	9.2%
	3. Students comparing answers in group	9.2%
Teacher_5	1. Students having group discussions without teacher's interaction	25.0%
	2. Students giving a presentation alone	21.7%
	3. Teacher checking answers with the whole class	15.2%

4. *How did the classes differ from one another?* To investigate inter-class difference as well as commonness, we plotted the main activities in Fig. 5. An activity was considered as main activity if the total duration of this activity last more than 10% of the total class time in at least one of the five classes observed. Figure 5 shows that asking follow-up questions (21.6%) and sharing personal/cultural experience with students (12.5%) seemed to be the favorite of Teacher_3, which occupied approximately one third of the total class time. Teacher_2 spent 30% of the total class time doing listening practice and 18% of the time checking answers with the whole class, with only a small portion of time allocated to group or pair discussions. Teacher_5 sometimes interacted with the students during their discussions, whereas such interaction was absent in the classes of Teacher_1 and Teacher_4.

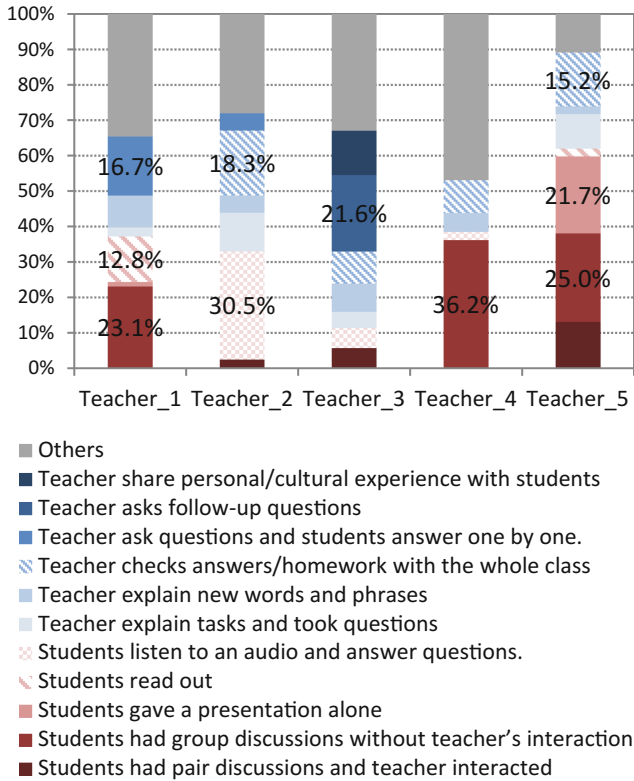


Fig. 5. Comparison of the ratio of main activities in each class

5 Discussions and Potential Applications

5.1 Principal Findings

We have shown that the heat maps of classroom activities provided useful visual feedback on teaching procedures and practices. The visualization of level-1 coding led to the answers to questions such as “Was the lesson teacher dominated?” “How often did teacher-student interaction occur?” “How was teacher’ time management?”. The visualization of level-2 coding rendered fine-grained information to help answer questions such as “How did the teacher structure the lesson?” “What kind of teacher-student interaction occurred?”. We found that group discussion was a commonly used strategy by all teachers to help students develop conversation skills.

The visual cues deepened our understanding on how the classes differed from one teacher to another. Through analyzing and visualizing the lessons, we achieved a better understanding on the strong and weak points of each teacher. For example, the class of Teacher_3 was featured by quick pace switches between students and teachers activities.

Indeed, Teacher_3 frequently interrupted the students to ask follow-up questions during their group discussions. Though sharing personal and cultural experience with students was no doubt very beneficial for students' English learning. Given that many Japanese students were shy and were not used to the western-style communication, however, it was likely that some students might feel overwhelmed by the dynamics of the class. This is supported by our observation that some students felt lost when Teacher_3 suddenly changed to a random new topic in the course of their discussions. As a contrast, the class of Teacher_4 had a steady pace, and he allocated abundant time for group discussions. However, our observations revealed that the students were usually left unattended in the second half of the class. They were asked to discuss on a certain topic, but the teacher didn't interact with the students. Teacher_2 differentiated her class from others' by integrating listening practice into conversation class. This strategy was supported by previous studies that aural and oral skills are close related [25, 26] and that enhancing listening could support the improvement on speaking in English learning [27]. Yet, we noticed that students spent more time doing exercise or individual tasks rather than practicing conversation skills in pairs or groups.

Visualizing classroom activities may also uncover interesting phenomenon that are otherwise not easy to perceive, which may eventually lead to new hypothesis in second language teaching. For example, the class of Teacher_5 demonstrated less dynamics and less frequent activity switches in comparison with other classes. However, all students of this class acknowledged that his teaching style was very impressive, and his classes had good rhythm. This challenges our stereotype that second language teachers should be entertaining and should bring fun to classes. Instead, we noticed that Teacher_5's class had the traits that all activities were almost uniformly distributed. There was a clear goal for each of the students' activity and the students were given sufficient time to practice English speaking in each activity. Based on this observation, we hypothesized that the rhythm of a class may be an important metric that impacts learning outcomes in second language teaching.

The proposed approach can also be employed to compare different classes given by the same teacher. If a teacher keeps tracking his/her classes to collect longitude data spreading over months or even years, he/her could gain insightful information on other key questions in critical reflection such as "Did I do anything differently than usual in my class?", "Did the change make things better or worse?", "what else can I change to make my teaching better?"

The heat maps of classroom activity rendered rich information that helped us gain insights on many important aspects of teaching in English conversation classes. However, we also found that the current design failed to manifest how students performed and reacted in each activity. The 2-level activity coding scheme only focused on classroom activities per se and provided little information on the performance and cognitive response of the students. The impact of these classroom activities on students' learning process therefore remains unknown. For example, Teacher_4 and Teacher_5 allocated almost the same ratio of class time for students to discuss in groups, yet the effect differed. According to our observations, the students of Teacher_5 made full use of the time and actively participated in the discussions, whereas the students of

Teacher_4 were very shy and were not actually on the task for full time. Future research is needed to address this problem.

5.2 Limitations

In addition to lacking focus on students' performance as explained above, the current study has several limitations presented below. First, capturing data on the activities happened during a class was done in a semi-automatic way. Whereas using smartphone applications significantly reduced the burden of data collection compared to paper-based approach, a complete automatic data collection scheme will further improve the scalability of the proposed method. Second, the current coding list only captures activities that are accompanied with verbal exchanges. Non-verbal activities such as teachers wandering around and observing group discussion were not considered in creating the activity coding list. These non-verbal activities may have educational benefit, but it is not possible to accommodate these activities following the existing line of how we created the activity coding list. A promising solution is to model English teaching using structured ontology engineering [28, 29]. Furthermore, we only used activity logs for analysis and visualization. Although activity logs provided an archival report on what happened in a class, the effect and impact of these activities remained unknown. Therefore, it will be preferable to include multi-modal data from several sources ranging from the evaluation on students' improvement, students' evaluation of their teachers, to biometrics such as heart rate and cognitive load during study and teaching [30, 31]. Third, the activity coding list that we extracted from the five observed classes did not exhaustively cover all possible classroom activities. This list should be extended as more classes are observed in the future. Last but not the least, it is important to investigate the impact of such visualization on teaching practices. The questions of interest include what changes have been trigger in teaching strategies and procedures and if such changes can eventually lead to positive outcomes in teaching and learning in second language classes. Future researches should focus on addressing these limitations.

5.3 Potential Applications

Classroom observation for critical reflection have immediate practical benefits for individual teachers as well as longer term benefits to second language teacher development programs. The method proposed in this study enables a lower-burden way for recording and analyzing classroom activities and can be employed in different sectors of second language education to support critical reflection. The proposed approach can help experienced teachers to move beyond the level of automatic or routinized responses to classroom situations and to investigate the effect of changes in their teaching strategies and procedures. New student teachers can use this approach to reflect their teaching practices to achieve a higher level of awareness of how they teach. Educational institutions can use this method in their training programs to foster critical reflection in student teachers. Education authorities can employ and extend this method to quantitatively compare and assess teachers' classroom management and teaching procedures. In the next step, we plan to conduct a large-scale study using the proposed approach to investigate how the

visualization of classroom activities facilitates teachers' reflection and teaching practices.

6 Conclusions

We have proposed a novel approach for low-burden classroom activity investigation based on experience sampling method and data visualization technology. We applied the proposed method to record and visualize the classroom activities of five English conversation classes given by native speakers. The visualization provided rich information on teachers' time management, class structure, and teacher-student interaction patterns, which led to the answering of some important questions that critical reflection aims to address. We highlighted four limitations of the current study to guide future research on this topic. The proposed method has the potential for large-scale implementation in educational institutions and will benefit both teachers and students through facilitating critical reflection on teaching and learning.

References

1. Clark, M., Otaky, D.: Reflection "on" and "in" teacher education in the United Arab Emirates. *Int. J. Educ. Dev.* **26**(1), 111–122 (2006)
2. Sze, P.: Reflective teaching in second language teacher education: an overview. *Educ. Res. J.* **14**(1), 131–155 (1999)
3. Willis, P.: Looking for what its' really like: phenomenology in reflective practice. *Stud. Contin. Educ.* **21**(1), 91–112 (1999)
4. Harris, A.: Effective teaching: a review of the literature. *Sch. Leadersh. Manag.* **18**(2), 169–183 (1998)
5. Cranton, P.: *Professional Development as Transformative Learning: New Perspectives for Teachers of Adults*. Jossey Bass, San Francisco (1996)
6. Richard, J.C., Lockhart, C.: *Reflective Teaching in Second Language Classrooms*. Cambridge University Press, New York (2013)
7. Yarmey, D.: *The Psychology of Eyewitness Testimony*. Free Press, New York (1979)
8. Fiske, D.: *Measuring the Concept of Personality*. Aldine, Chicago (1971)
9. Schratz, M.: Researching while teaching: an action research in higher education. *Stud. High. Educ.* **17**(1), 81–95 (1992)
10. <http://www.atimelogger.com/>
11. Wilkinson, L., Friendly, M.: The history of the cluster heat map. *Am. Stat.* **63**(2), 179–184 (2009). <https://doi.org/10.1198/tas.2009.0033>
12. Pak, J.: *Find Out How You Teach*. National Curriculum Resource Centre, Adelaide (1986)
13. Bartlett, L.: Teacher development through reflective teaching. In: Richards, J.C., Nunan, D. (eds.) *Second Language Teacher Education*, pp. 202–214. Cambridge University Press, New York (1990)
14. Murphy, J.M.: An etiquette for the non-supervisory observation of L2 classrooms. In: *Proceedings of the 1st International Conference on Teacher Education*, City Polytechnic of Hong Kong (1991)
15. Richards, J.C., Lockhart, C.: Teacher development through peer observation. *TESOL J.* **1**(2), 7–10 (1991–1992)

16. Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M.: *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage, Thousand Oaks (2007)
17. Larson, R.W., Csikszentmihalyi, M.: The experience sampling method. *New Dir. Methodol. Soc. Behav. Sci.* **15**, 41–56 (1983)
18. Liang, Z., Chapa Martell, M.A.: Framing self-quantification for individual-level preventive health care. In: *Proceedings of HEALTHINF*, pp. 336–343 (2015)
19. Pennington, M., Young, A.: Approaches to faculty evaluation for ESL. *TESOL Q.* **23**(4), 619–646 (1989)
20. Good, T., Brophy, F.: *Looking in Classrooms*. Harper and Row, New York (1987)
21. Freeman, D.: Observing teachers: three approaches to in-service training and development. *TESOL Q.* **16**(1), 21–28 (1982)
22. Wheeler, L., Reis, H.T.: Self-recording of everyday life events: origins, types, and uses. *J. Pers.* **59**(3), 339–354 (1991)
23. Edwards, C., Shortall, T., Willits, D., et al.: *Language Teaching Methodology*. Center for English Language Studies, Birmingham (2000)
24. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. *Int. J. Qual. Methods* **5**(1), 80–92 (2006)
25. Vandergrift, L.: Recent development in second and foreign language listening comprehension research. *Lang. Teach.* **40**, 191–210 (2007)
26. Bozorgian, H.: The relationship between listening and other language skills in international English language testing system. *Theory Pract. Lang. Stud.* **2**(4), 657–663 (2012)
27. Astorga-Cabezas, E.D.: The relationship between listening proficiency and speaking improvement in higher education: considerations in assessing speaking and listening. *High. Learn. Res. Commun.* **5**(2), 34–56 (2015)
28. Mizoguchi, R., Bourdeau, J.: Using ontological engineering to overcome common AI-ED problem. *J. Artif. Intell. Educ.* **11**, 107–121 (2000)
29. Munn, K., Smith, B. (eds.): *Applied Ontology: An Introduction*, vol. 9. Walter de Gruyter, Berlin (2008)
30. Mills, C., Fridman, I., Soussou, W., et al.: Put your thinking cap on: detecting cognitive load using EEG during learning. In: *Proceedings of LAK 2017* (2017)
31. Di Mitri, D., Scheffel, M., Drachsler, H., et al.: Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data. In: *Proceedings of LAK 2017* (2017)



On-site Knowledge Representation Tool for Employee-Driven Service Innovation

Kentaro Watanabe^(✉)

National Institute of Advance Industrial Science and Technology, Tokyo, Japan
kentaro.watanabe@aist.go.jp

Abstract. Knowledge of service employees is an important source of innovation in the service sectors. It is a challenge to collect and share knowledge among service employees for utilization on a daily basis. In this study, several on-site knowledge representation tools and one test case are introduced to illustrate the effectiveness of knowledge representation tools applied at work. In addition, the advantages and effective use cases of different types of knowledge representation are also discussed.

Keywords: Knowledge representation · Service field
Employee-driven innovation

1 Introduction

A service is an important economic and social activity used to create value for and with customers [1–3]. Service activities are becoming more complex owing to diversified customer requirements and a specialization of tasks. This tendency is particularly outstanding in interpersonal services with frequent interactions among customers and employees, such as care services [4].

In such types of service fields, knowledge of the service employees strongly contributes to the service productivity and customer satisfaction [5]. Knowledge sharing and its support system have been an important research topic in the research domain of information systems [6]. In innovation research, to tackle this issue by utilizing the knowledge of employees regarding service situations and practices, employee-driven innovation (EDI) has recently attracted attention [7, 8]. Service practices and interactions among employees and customers in the service fields are not clearly visible from a management viewpoint, and it is therefore difficult to manage service activities using only a top-down management approach. It is important to not only create innovative ideas from the employees' perspective, but to also promote organizational learning to increase the impact of innovation [8]. Although various EDI cases based on workshops have been proposed [9, 10], a means to represent and share such ideas as knowledge for innovation in a daily busy work setting have yet to be established.

This study focuses on the potential of technological support for knowledge representation by service employees. Studies on knowledge representation tools and methodologies have been conducted by the present author with regard to autonomous design

activities by employees in service fields [11, 12]. In addition to existing systems, a new knowledge representation tool is introduced herein to represent and share practical knowledge in service fields based on previous research results. Based on a test case of the proposed system and comparison among representation tools, the advantages and effective use cases of different types of knowledge representations are discussed.

2 Knowledge Representation at Work

2.1 Employee Knowledge and Its Representation

Employee knowledge has been an important topic in several research fields. In the field of informatics, Computer-Supported Cooperative Work (CSCW) has been the main research area, tackling issues regarding knowledge sharing among employees [6]. The sharing of knowledge in an organization has been considered an important source of innovation. Recent technologies, such as mobile devices, have become available for service employees who frequently interact with customers and other employees while on the move [14]. Utilizing these types of ICTs, digitalization has been anticipated to increase the productivity of service sectors [15]. Knowledge sharing through digitalization is also a key issue among this trend.

To realize effective knowledge sharing, not only databases or any information systems used to support sharing knowledge, but also the active participation of organizational members, has been emphasized [13]. The active role of employees has also been highlighted in organizational and innovation studies. For better organizational management, how to circulate an employee's knowledge in an organization is essential. EDI is considered part of organizational learning used to promote and manage innovation processes by employees [7, 8]. The role of ICT has also been emphasized in this innovation process in different service firms [5].

Reflection is an important process for organizational learning [8]. Expressing thoughts through reflection is meaningful in establishing a professional mindset, and for creating organizational knowledge in the workplace. Nishimura et al. [16] emphasized the importance of knowledge explication, which allows employees to explicate their own knowledge for standardizing and refining their work processes. In the process of knowledge explication, active participation of care professionals is essential, and a workshop approach has been adopted, as is common in other knowledge sharing approaches. Meanwhile, one research challenge in terms of knowledge sharing and knowledge explication is how to represent and share knowledge in daily work. This is particularly challenging for the employees of interpersonal services, who have limited time to represent knowledge while at work. Technological and methodological support for representing knowledge while at work is expected to promote innovation.

2.2 Previous Development

In this paper, two cases regarding the development of knowledge representation tools applied at work are introduced.

Zuzie Poetry. Zuzie Poetry is a representation tool designed to allow nurses to share their experiences in a hospital [11]. The problem tackled by this system is how to share the experiences of nurses who collaborate with each other in a daily work setting. In a busy work environment, it has become difficult for nurses to share their view of each patient case when further collaboration is required. Unlike ordinary knowledge representation based on a structured representation, such as a resource description framework [17], this system adopts an unstructured, collaborative representation among the nurses. As a result of a literature survey on a design representation, it was assumed that an unstructured representation approach, for example in combination with text and pictorial representations, would be effective for use by non-technical employees [11]. Through an MED project, which is an interdisciplinary research project among researchers and practitioners in the areas of medicine, design, and informatics, a representation tool used by nurses at a university hospital was co-designed, which eventually became Zuzie Poetry.

Figure 1 shows a screenshot of Zuzie Poetry. This system operates on a PC. Users can represent humans using avatars or photographs, and their actions and emotions through texts and emoji. Detailed situations such as relations among humans and their environment can be described by drawing lines. Changing scenes can also be represented using several different sheets. Thus, users can describe, for example, how the situations of their patients and other surrounding individuals have changed. Several nurses joining the project tested the system and evaluated it as an effective tool for sharing multiple viewpoints toward a particular case [11]. The main goal of the system is to develop mutual understanding and empathy among coworkers, which will improve the sense of teamwork and decrease stress in the workplace [18].

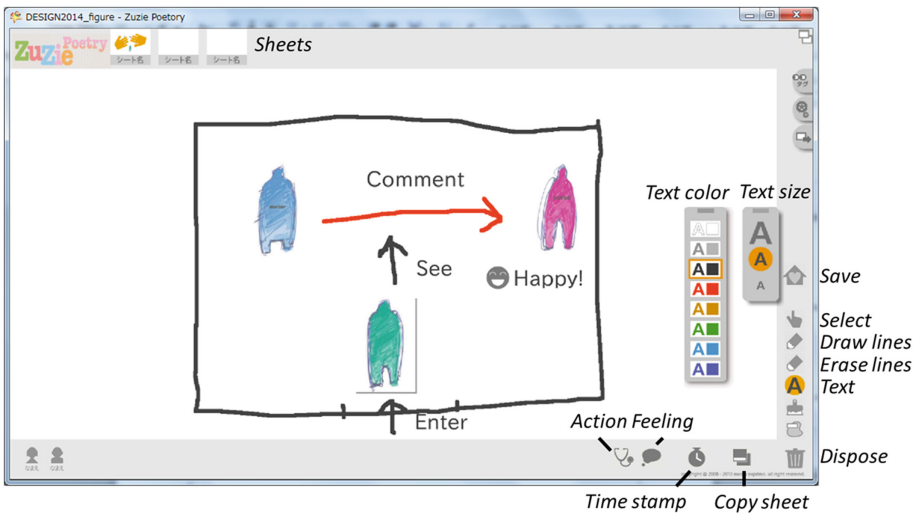


Fig. 1. Zuzie Poetry [11].

DRAW. Another example of a knowledge representation tool development is the Design Representation tool for Autonomous Workplaces (DRAW) [12]. DRAW was developed through collaboration with care professionals at a care facility. Compared to Zuzie Poetry, DRAW is aimed at representing ideas and knowledge of care professionals more immediately. This tool operates on tablet devices. DRAW was designed to share not only practical knowledge at work, but also contexts of knowledge, such as why an idea for improvement arose and under what conditions the idea can be successfully applied. By sharing such information, care professionals can utilize knowledge more flexibly and adequately toward similar problems and situations.

Figure 2 shows a screenshot of DRAW exemplifying a risk management report in a care facility. DRAW has several functions for representing particular issues and design results in different service fields, such as adding texts, photographs, and line drawings of various colors. The simple interfaces of DRAW enable users to input their ideas intuitively. In particular, the photo function supports instant and effective information sharing. Representation results can be shared with other employees through the use of another communication system. This tool was used for knowledge sharing at the care facility where the tool was co-designed.

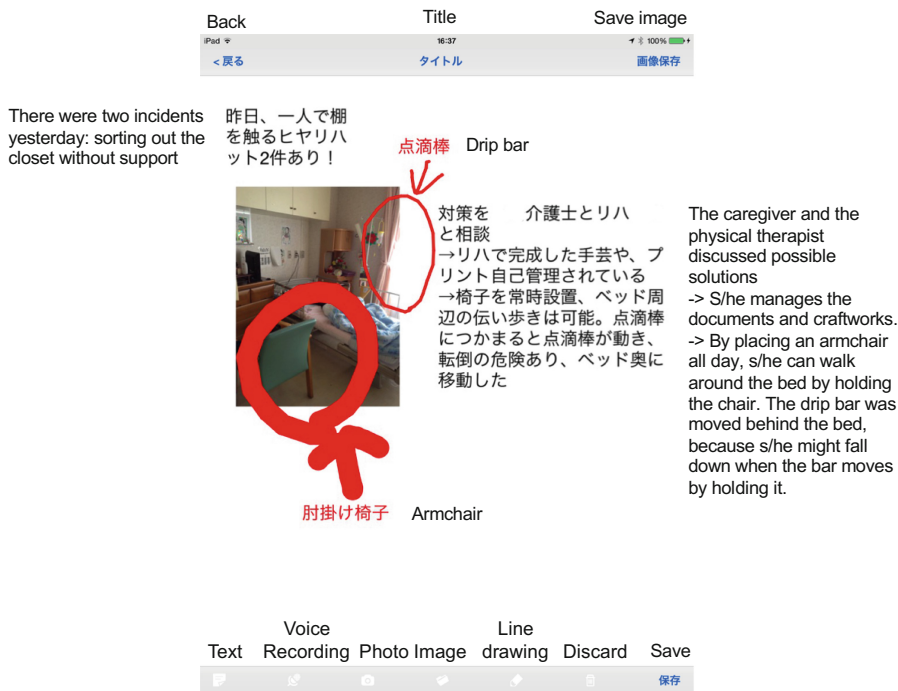


Fig. 2. DRAW [12].

3 DRAW2

3.1 Concept

In this section, a new knowledge representation tool, called DRAW2, is introduced. DRAW2 was developed to tackle the following two challenges experienced during the development of the previous version.

Reusability of Previous Knowledge. The systems introduced above apply an unstructured representation. This type of representation allows knowledge input without specific skills; however, this makes it difficult to reuse knowledge with computational support. For example, one of the common approaches for knowledge utilization is a data search. It is inefficient to search data without the data structure. In particular, knowledge representation in such systems includes photos, figures, and voice records, which are difficult to search using the small amount of computer resources available at a typical workplace.

Tagging to a represented element is an effective approach to knowledge utilization. However, it is difficult for service employees, many of which are novices with regard to new technologies, to add tags to a knowledge representation. Therefore, a new mechanism for adding tags to a knowledge representation is required.

Efficient Knowledge Representation at Work. Users of both Zuzie Poetry and DRAW need to input data on a blank sheet. This feature allows users to represent their ideas freely. However, it is not necessarily acceptable for all employees. To collect many pieces of knowledge, an easier method for knowledge representation is needed.

3.2 Functions

Figure 3 shows a screenshot of DRAW2. Most of the representation functions are from the original DRAW, such as the text input, line drawing, voice recording, and photographs. DRAW2 has two main additions.

Template. One addition is a template function. Users can create the basic framework of a knowledge representation through specific items users need to fill in, as shown in Fig. 3. This allows users who are unfamiliar with representing information to represent their ideas more easily. Meanwhile, users of templates can still modify the positions of the elements and add new elements onto a sheet. Therefore, freedom of expression is still maintained with this system.

Tagging as a Caption. The second addition is a caption feature for each element. A caption attached to an element, such as a text box and a photo area, as indicated in Fig. 3, works as a tag for a later search. The captions can be arranged in the templates in advance. By doing so, the search for knowledge elements becomes more efficient.

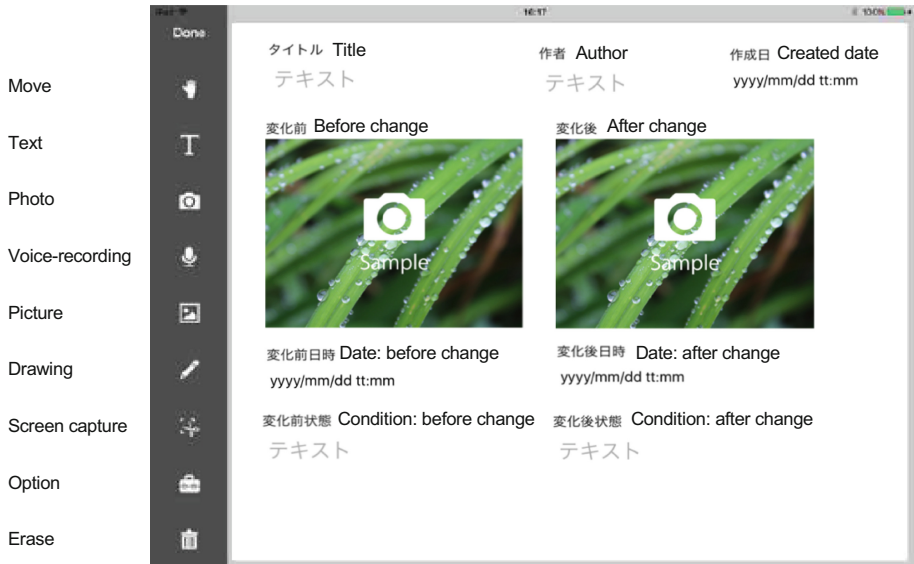


Fig. 3. DRAW2.

3.3 Test Case

To clarify the effectiveness of DRAW2, two workshops were conducted with care professionals. The participants were from the facility where DRAW was developed. Both workshops were held in the meeting room of the facility.

In the first workshop, the participants were asked to test DRAW2 after receiving brief instructions regarding its operation. They were asked to represent their ideas. All information input during the workshop was fictive. After the test use, group interviews were conducted with the participants regarding the usability of DRAW2 and its potential use at work. Five care professionals joined the workshop during two separate sessions.

In the second workshop, the participants were asked to represent knowledge using templates developed by the author based on the usage ideas collected during the first workshop. Some photographs were prepared in the test devices to represent their ideas in a more realistic manner. After the workshop, the participants were interviewed again about their impressions regarding the template function, usability, and related issues. Three participants joined the second workshop.

3.4 Results

During the first workshop, the participants were able to use DRAW2 rather smoothly, with some advice given by the facilitator. Although the data they were able to input in the workshop place were fairly limited, they represented what they felt was interesting about their work.

After the workshop, they provided the following potential usages during the interview.

- Recording changes in residents
- Summarizing the daily situations of residents, including good conditions and cases of anxiety
- Evaluation of home conditions for independent living of the residents
- Swallowing evaluation

In addition, the development of training materials through video recordings, which was not available in DRAW2, was taken up as a potential usage.

In the second workshop, the participants input fictive knowledge based on their assumptions, utilizing the templates and photographs attached in the test devices. After the test, the participants answered that it is easier to input knowledge compared to filling in a blank sheet. In addition to the requirements regarding the appearance and usability, such as the font, tables, and lists, a search function used to find a previous representation for reference was taken up as an expected function. This function has already been implemented in the updated version of DRAW2.

4 Discussion

4.1 Semi-structured Representation

In comparison to Zuzie Poetry and DRAW, DRAW2 provides a template function as a scheme for inputting information. While required information is explicitly presented in the template. Users can add and modify templates, and also input different type of information in addition to selected template. This point provides a significant difference from tools with an unstructured representation. In this sense, knowledge representation using DRAW2 is semi-structured.

This characteristic is also highlighted in the workshop results. Some of the usages taken up during the first workshop were considered in the previous project development of DRAW. The participants might consider DRAW2 to be a representation tool used to accumulate focused sets of information meaningful at work. One interesting result from the first workshop was the interest of the participants regarding the daily situations of the residents. Although care professionals take care of the residents during their daily work, they were still interested in the reason for the particular conditions and situations of the different residents. Such data would be useful, particularly for inexperienced workers, to learn the signs of potential problems or incidents that occur in daily situations. Accumulating data on residents with such concerns would be meaningful for assessing the residents and providing interventions when needed.

The responses of the participants after the second workshop were positive regarding the expected role of DRAW2, aiming at reducing the cost needed to collect knowledge about care applied in daily work, although there are several issues of its interface that need to be improved. However, because the number of participants was small, the results of this test case were limited to obtaining the general attitudes toward DRAW2. The available information, material, and scenarios applied during the workshops were also limited. The test use of DRAW2 during daily work should be considered in a future study to obtain more generalized insights.

4.2 Comparison Among Knowledge Representation Tools at Work

In this section, different types of knowledge representation tools applied in a work environment are compared. In addition to an unstructured representation (Zuzie Poetry and DRAW) and a semi-structured representation (DRAW2), a structured knowledge representation is taken into account. One example of structured knowledge representation is the Convincing Human Action Rationalized Model (CHARM) [19]. CHARM externalizes implicit knowledge for nursing guidelines in the form of a functional structure. Nishimura et al. [16] further formalized the process to externalize knowledge of care workers through a workshop.

Table 1 illustrates a qualitative comparison of three types of knowledge representation. An unstructured representation in the cases of Zuzie Poetry and DRAW enables users to represent their thoughts freely, which could lead to a broadening of their views. This representation style does not require specific knowledge or skills, but it may be difficult for some people to express their ideas without any guidelines or clues. In addition, the lack of formality in an unstructured representation affects its reusability in a negative manner, as was mentioned above, and a computational analysis is also less applicable. In contrast, a structured representation has its advantage in the formality of a representation, which assures a more rigorous representation of knowledge. Computational analysis is more applicable to this type of representation. Meanwhile, users of structured representation are required to have certain skills such as its notation, and some support by professionals. In addition, users are not allowed to add a new concept in a representation notation, which limits their expression.

Table 1. Comparison among knowledge representation types.

	Unstructured	Semi-structured	Structured
Freedom of representation	High	Middle	Low
Formality of representation	Low	Middle	High
Availability of computational analysis	Low	Middle	High
Ease of input for workers	Middle	High	Low
Examples	Zuzie Poetry [11], DRAW [12]	DRAW2	CHARM [19]

A semi-structured representation is characterized through the intermediate features between unstructured and structured representations. A semi-structured representation includes the basic scheme of the representation, which could guide the direction of the representation compared to an unstructured representation. In addition, some formalities and applicability of a computational analysis are assured. Meanwhile, a semi-structured representation allows users to add a different concept in its representation, which improves the freedom of representation.

These features of each representation type differentiate the use case. An unstructured representation would be more suitable to create innovative ideas without any assumptions. A structured representation is more applicable to creating general knowledge among workers in a workshop with the support of professional facilitators [16]. A semi-structured representation can contribute toward data accumulation in a daily work setting with a more specific focus, such as the usages described in the workshops of DRAW2.

The representation types introduced in this paper can be converted into one another. For example, represented knowledge with an unstructured representation can be a basic scheme of a semi-structured representation used to accumulate similar types of knowledge at work. Collected knowledge with a semi-structured representation can then be described as more structured knowledge through a workshop. The utilization of different representation types according to the objective is important.

4.3 Computational Utilization of Knowledge Representation

Although a computational analysis of a knowledge representation is not the major issue of this paper, collected knowledge representations have significant potential for more advanced utilization. A tag search for a semi-structured representation is a straightforward approach to obtaining intended knowledge for use. Different types of text mining methods will also be available for a semi-structured knowledge representation. Visual and voice data associated with texts and tags can be utilized through a text-based analysis, although the expected amount of data from the workplace would not be sufficient for machine learning.

The access to collected data is also an interesting topic. In addition to mobile devices, social robots, for example, can also introduce knowledge to care workers [20]. Different methods for a knowledge presentation should be taken into account in future research.

5 Conclusion

In this paper, three knowledge representation tools designed for use in the workplace were introduced. Specifically, DRAW2 as a system with a semi-structured representation for accumulating knowledge representations in daily work settings was illustrated. A test case of DRAW2 showed its ease of use in collecting representations with specific focuses in a work environment. A comparison among the different types of representations (unstructured, semi-structured, and structured) clarified their advantages, challenges, and expected usages. Users or those introducing a knowledge representation tool are required to consider which type of representation is suitable to their particular use cases.

As future research, an effective computational analysis will be conducted for each type of representation. An evaluation of the proposed tool through user tests in actual work environments is another future task.

Acknowledgment. I appreciate the sincere support for this study by Wakoen and Keiju Healthcare System. This work was supported by JSPS KAKENHI Grant Number JP 15K16174 and 16H02916.


References

1. Spohrer, J., Kwan, S.K.: Service Science, Management, Engineering, and Design (SSMED): an emerging discipline – outline & references. *Int. J. Inf. Syst. Serv. Sect.* **1**(3), 1–31 (2009)
2. Sundbo, J.: Management of innovation in services. *Serv. Ind. J.* **17**(3), 432–455 (1997)
3. Vargo, S.L., Lusch, R.F.: Evolving to a new dominant logic for marketing. *J. Mark.* **68**(1), 1–17 (2004)
4. Miwa, H., Fukuhara, T., Nishimura T.: Service process visualization in nursing care service using state transition model. In: 4th International Conference on Applied Human Factors and Ergonomics, CD-ROM (2012)
5. Sundbo, J.: Empowerment of employees in small and medium-sized service firms. *Empl. Relat.* **21**(2), 105–127 (1999)
6. Ackerman, M.S., Dachtera, J., Pipek, V., Wulf, V.: Sharing knowledge and expertise: the CSCW view of knowledge management. *Comput. Support. Coop. Work* **22**(4–6), 531–573 (2013)
7. Kesting, P., Ulhøi, J.P.: Employee-driven innovation: extending the license to foster innovation. *Manag. Decis.* **48**(1), 65–84 (2010)
8. Høyrup, S.: Employee-driven innovation and workplace learning: basic concepts, approaches and themes. *Transf.: Eur. Rev. Labour Res.* **16**(2), 143–154 (2010)
9. Fuglsang, L.: Bricolage as a way to make use of input from users. In: Sundbo, J., Toivonen, M. (eds.) *User-Based Innovation in Services*, pp. 25–44. Edward Elgar Publishing, Cheltenham (2011)
10. Telljohann, V.: Employee-driven innovation in the context of Italian industrial relations: the case of a public hospital. *Transf.: Eur. Rev. Labour Res.* **16**(2), 227–241 (2010)
11. Watanabe, K., Fujimitsu, S., Harada, Y., Niino, Y., Kobayakawa, M., Yamada, K., Sunaga, T., Sakamoto, Y., Nishimura, T., Motomura, Y.: Proposal of a design support tool for employees to represent services. In: *Proceedings of DESIGN 2014, Dubrovnik, Croatia* (2014)
12. Watanabe, K., Nishimura, T.: Employee-driven design activities for services: a case study in elderly-care. In: *AHFE 2015, Las Vegas, USA* (2015)
13. Schmidt, K., Bannon, L.: Taking CSCW seriously: supporting articulation work. *Comput. Support. Coop. Work* **1**(1), 7–40 (1992)
14. Watanabe, K., Fukuda, K., Nishimura, T.: A technology-assisted design methodology for employee-driven innovation in services. *Technol. Innov. Manag. Rev.* **5**(2), 6–14 (2015)
15. D’Emidio, T., Dorton, D., Duncan, E.: Service innovation in a digital world. *McKinsey Q.* **2014**(4), 55–62 (2014)
16. Nishimura, S., Fukuda, K., Nishimura, T.: Knowledge explication: current situation and future prospects. In: *IJCAI 2017 Workshop on: Cognition and Artificial Intelligence for Human-Centered Design, Melbourne, Australia*, pp. 1–7 (2017)
17. W3C: RDF 1.1 Concepts and Abstract Syntax. <https://www.w3.org/TR/rdf11-concepts/>. Accessed 23 Feb 2017

18. Watanabe, K., Fujimitsu, S., Harada, Y., Yamada, C.K., Sunaga, T., Kobayakawa, M., Niino, Y., Sakamoto, Y., Nishimura, T., Motomura, Y.: Development of a support system to represent and share work experience for nursing services. *J. Inf. Process. Soc. Jpn.* **56**(1), 137–147 (2015). (in Japanese)
19. Nishimura, S., Kitamura, Y., Sasajima, M., Williamson, A., Kinoshita, C., Hirao, A., Hattori, K., Mizoguchi, R.: CHARM as activity model to share knowledge and transmit procedural knowledge and its application to nursing guidelines integration. *J. Adv. Comput. Intell. Intell. Inform.* **17**(2), 208–220 (2013)
20. Jokinen, K., Nishimura, S., Watanabe, K., Nishimura, T.: Human-robot dialogues for explaining activities. In: *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology*, Singapore (2018)



Consideration of Application Cases of Structured Manual and Its Utilization

Satoshi Nishimura , Ken Fukuda, and Takuichi Nishimura

National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
satoshi.nishimura@aist.go.jp

Abstract. Medical and long-term care costs in Japan are increasing ahead of the rest of the world. Burdens of care worker are also increasing. Therefore, it is an urgent issue to improve the productivity and quality of care. In general, sharing workers' knowledge supports business operation. However, constructing manuals consistently is costly. Knowledge explication methodology was developed to systematize procedural knowledge by the employees themselves. The authors have applied the methodology to several domains including elderly care.

This study was conducted to clarify the application cases of knowledge explication and the resulting systematized knowledge: so-called structured manuals. First, the authors clarify the processes that occur related to knowledge explication. Second, the authors attempt to classify the application cases according to conventional classification and the contents of the structured manuals. The authors also describe avenues of their future work.

Keywords: Knowledge engineering · Service engineering · Service science

1 Introduction

Medical and long-term care costs are increasing because of the progress of Japan's aging society [1]. The demand for caregivers is also increasing [2]. Therefore, care process productivity should be raised by reducing staff education costs, procedural variation among staff, and risk. Some projects are using information technology to resolve the difficulties. Robotic Care Equipment Development and Introduction Project aimed at reducing physical burdens of staff using care robots¹. Several systems have been developed to reduce the recording time used for daily reports [3].

Construction of manuals for procedural knowledge of care is one solution to improve caregiver productivity. As described herein, we specifically examine manuals that support staff education, share standard procedural knowledge, and elucidate risks beforehand. Caregivers themselves can construct conventional text manuals.

Some difficulties are inherent in text manuals and conventional method proposed in knowledge engineering domain. The first is (a) that they not include implicit knowledge because text manuals are not systematized. A second is (b) that it is not easy to read sometimes because a text manual requires interpretation to understand its semantics. A

¹ Robotic Care Devices Portal: <http://robotcare.jp/?lang=en>.

knowledge engineering approach helps to overcome these difficulties, but some are inherent in the approach. The first is (c) that it is not easy for a domain expert to construct manuals because knowledge representation models, which are used for systematization, are too difficult for domain experts. A second is (d) that it is not easy to compare them because they are systematized for specific domains.

We proposed the methodology presented in Fig. 1 for this study of sharing procedural knowledge at elderly care sites [4]. The emphasis of this proposed methodology is that workers systematize site-specific procedural knowledge based on common procedural knowledge. The methodology can be applied to other domains such as higher education and local revitalization. We applied the methodology to some domains. As described in this paper, we clarify the characteristics of the knowledge explication method and each domain to which we applied our proposal.

2 Knowledge Explication

Figure 1 presents an overview of the proposed methodology: knowledge explication [4]. Procedural knowledge is produced according to the following steps. The first step is to systematize common procedural knowledge. Common procedural knowledge is knowledge that is included in textbooks and which is common among work sites in the same domain. The second step is explication of site-specific procedural knowledge by the workers themselves. Site-specific procedural knowledge is knowledge that often occurs at the site.

The following are the aims of our research.²

- (c) Caregivers can construct structured manuals easily and independently.
- (d) Structured manuals are easy to compare.

To achieve them, we clarified the perspective of knowledge construction and improved the visualization format of the knowledge representation model.

The conventional knowledge engineering approach is mainly used by researchers [5]. Our proposed approach is used by employees themselves. It is different from knowledge engineering approach. A knowledge discovery approach is also investigated. The benefit of such approaches is their lower cost than knowledge engineering approaches, but these approaches require big data or well-structured data such as Wikipedia. They are difficult to apply to the elderly care domain because of a shortage of related data.

Convincing Human Action Rationalized Model (CHARM) [6] is a goal-oriented model. An action is interpreted as a state change of a target object. A single action is realized using a sequence of detailed actions. In this context, a single action as a state change is interpreted as a goal of the sequence. Nishimura et al. conceptualized the reason a single action can be achieved by the sequence as a “way of action achievement.” The crucially important point is detachment of “what is achieved” and “how to achieve it.” Such a modeling perspective helps knowledge builders to explicate intermediate goals of actions, which are usually implicit in text. As described in this paper, we call the resulting knowledge a structured manual.

² The alphabetical characters c and d correspond to issues introduced in Sect. 1.

This structuring method provides commonality among service types, such as facility services, day care services, home care services, from a goal-oriented perspective. For instance, to achieve transfer in any service types, care takers almost always perform the following actions, determining how to transfer, preparing for transfer, transfer, checking a posture and physical conditions, doing post-processing, and recording the event. These actions can be achieved in different ways under different circumstances. The goal-oriented perspective helps users to consider which actions are common with other services.

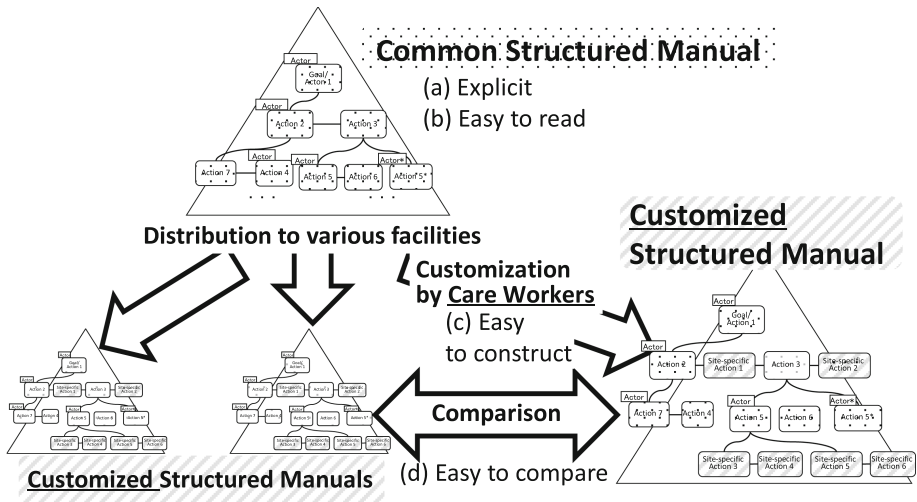


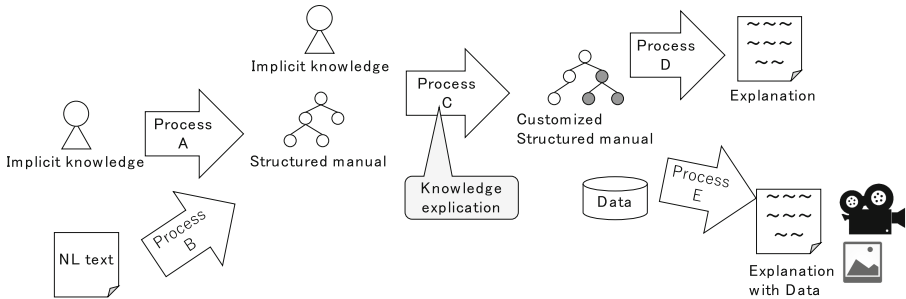
Fig. 1. Overview of knowledge explication.

3 Consideration of Process Related to Knowledge Explication

3.1 Overview

Figure 2 presents an overview of the consideration result. Five processes occur related to knowledge explication. We interpret the process as a state change of an object. In this case, the factor to distinguish the process is what type of object is observed before (and after) each process. For instance, process A shows the state change, which changes a representation form of the knowledge from implicit knowledge in the human brain to a structured manual as explicit knowledge. Process B changes a representation form of the knowledge from an explicit natural language text form to a structured manual as explicit knowledge. Process D changes a representation form of knowledge from a structured manual to an explanation for humans. Each process is a general concept. They can be specialized to more concrete concepts. In a knowledge explication context, processes C and E are important concepts that are specialized respectively by processes A and D. Process C changes a representation form of the knowledge from implicit knowledge in the human brain and a structured manual to a structured manual that contains more information. The operands before the state change of process C are described in greater detail than process A. Process

E changes a representation form of the knowledge from a structured manual and data related to the manual to explanation for human. Process E is also specialized from process D in the respect of the operands before the state change. In this context, we classify representation forms into implicit knowledge in the human brain, natural language text, and structured manuals. The criterion is their formalism.






	Type	Formalism	Description
Knowledge source	 Implicit knowledge	Low	This type of knowledge is usually stored in the human brain. It is difficult for others to access the knowledge. This type of knowledge is shared by oral instruction during OJT.
	 Natural Language Text	Medium	This type of knowledge is usually stored in the form of a text-manual. It is easier to access the knowledge rather than implicit knowledge. However, others need to interpret the knowledge to understand it. For the information systems, it is necessary to pre-process the text to utilize the content.
	 Structured manual	High	This type of knowledge is not usually used in practical service fields. It is easy to access the knowledge if the knowledge content is well-structured in a consistence manner. However, it is sometimes difficult to construct the knowledge in a such structured manner.

Fig. 2. Overview of knowledge-explication related processes.

3.2 Process A

Process A changes a representation form from implicit knowledge in a human brain to a structured manual. Such a process is crucially important for the use of human knowledge by information systems. Research in the knowledge engineering domain has presented some methods to realize process A such as interviewing [5]. The main actor of the process in the knowledge engineering domain is a knowledge engineer. The knowledge engineer is taken by the researcher and has know-how to explicate the knowledge from the domain expert and to systematize it.

As described earlier, we construct a common structured manual; then a domain expert systematizes their knowledge through customization of the common structured manual. We extracted the knowledge from a textbook to construct the common structured manual.

3.3 Process B

Process B changes a representation form from natural language text to a structured manual. Research in the knowledge engineering domain has revealed some methods to realize process B, such as protocol analysis [5]. The main actor of the process in such methods is also a knowledge engineer.

Stanford Open Information Extraction [7] can take an instance of process B. The system extracts the relation among important terms in natural language text and systematizes the information to triples. The systematization format of the output differs from the structured manual.

There is some space to consider automatic systematization from a natural language text in the form of CHARM. It is necessary to add information such as intermediate goals and risks, which are usually implicit in a natural language text.

3.4 Process C

Process C changes a representation form from implicit knowledge in the human brain and a structured manual to a structured manual that includes more information. We specialized the operand of the process and the specialization cases the specialization from process A into process C.

Process C occurs in taking process A in the knowledge engineering domain. It is not easy to systematize all the knowledge at once. The knowledge engineer usually receives feedback from the domain expert based on the systematized knowledge. The process can be interpreted as process C.

Knowledge explication can be interpreted as a kind of process C. The difference from the knowledge engineering approach is the actor who systematizes the knowledge [4]. The actor of the knowledge engineering approach is mainly the knowledge engineer. In contrast, the actor of the knowledge explication is mainly the domain expert. It might entail lower costs than those of the knowledge engineering approach to revise the customized knowledge.

3.5 Process D

Process D changes a representation form of the knowledge from a structured manual to explanation for humans. For instance, we expect to translate the form of structured manuals to a list form and display them to domain experts. If the manuals are well structured, then information extraction, such as extracting the actions which an elderly person takes in relation to care services, can be realized as an instance of process D. An interactive explanation of the structured manuals can also be considered. We present related details in Sect. 5.2.

We only experienced process D as displaying the structured manuals as they are. Those examples described above are to be explored further in future work. We are preparing schema for the structured manuals to realize computational support to translate and display them.

3.6 Process E

Process E changes a representation form of the knowledge from a structured manual and data related to the manual to explanation for human. We specialized the operand of process D from the structured manual alone into the structured manual and data in the service field. That causes the specialization of process D into process E.

We expect that the outcome of process E helps convince the person to whom an explanation is given. Data in the service field are more familiar to the workers. Such data can be interpreted as “evidence” when the relation among the data and the knowledge is explicit.

Ushiku et al. investigate to generate the caption from images [8]. The images can be interpreted as data. Therefore, the caption generation can be interpreted as a similar process to process E.

It is a new challenge for us to generate an explanation from the structured manuals and the data. The first step is how to produce a relation among the structured manuals and the data.

3.7 Discussion

We illustrate an abstract overview of the processes occurring relative to knowledge explication. The result of the consideration clarified five processes. Processes A, B, and C belong to the research field of knowledge representation and engineering. The emphasis of the processes is how to explicate and systematize the knowledge from domain experts. Processes D and E belong to the research field of information retrieval, user interface, and natural language generation. Such research specifically examines how to use the knowledge. Our research emphasis is on the first one: the research field of knowledge representation and engineering.

4 Application Area of Knowledge Explication and Structured Manuals

This section presents classifications of application areas to which we have applied knowledge explication method and structured manuals. First, we attempt to apply Japan Standard Industrial Classification (JSIC)³ to classify the application areas. Some areas to which we have applied our method are not business areas. Therefore, JSIC is unsuitable for classification in this research. Second, we attempt to classify the areas from the viewpoint of contents that we construct.

³ Japan Standard Industrial Classification (Rev. 13, October 2013) Structure and Explanatory Notes Index: http://www.soumu.go.jp/english/dgpp_ss/scido/sangyo/san13-3.htm.

4.1 Classification of Application Areas According to JSIC

Table 1 presents the application cases. “Domain” shows a short description of application areas. “Description of the application” shows what we intend to do in the application. It is related to how to use structured manuals that are constructed in the application case. The last column shows the category in JSIC correspondence to the application case. The domains of two cases are not business areas. Therefore, we cannot correspond to the category in JSIC to those cases.

The classification results show the following.

- We do not cover most industrial domains.
- Our proposal is suitable not only for service fields such as care and education, but also to other domains such as construction and manufacturing.
- The salient application aim is the support of human development.

Such findings indicate that classification of application areas according to JISC is insufficient for characterizing the knowledge explication method and structured manuals.

Table 1. Application cases and correspondence to JSIC

No.	Domain	Description of the application	Japan Standard Industry Classification
1	Elderly care	Manual construction and standardization of procedures	P: Medical, healthcare and welfare
2	Higher education	Support of reflection in active learning class	O: Education, learning support
3	Local revitalization support	Support of human development for making local seniors more active	–
4	Coimagination method	Systematic description of the procedure of coimagination method and making relations among the description and records	–
5	Construction	Support of construction control (use-case is human resource development.)	D: Construction
6	Waste treatment	Systematic description of plant operation	R: Services, N.E.C.
7	Autonomous vehicle	Systematic description of driving actions and making relation with law and case law	E: Manufacturing
8	Guitar rendition	Support of guitar rendition training	O: Education, learning support

4.2 Classification of Application Area According to Structured Manual Contents

The following sections describe our attempt at classification from the cases according to contents which we constructed and will construct in each application case. We employed the notion of “meta-function” and “base function” proposed in an earlier report of the literature [9]. Definitions of those terms are the following: “A function that enables others to work is called a meta-function.” and “Base functions are demanded by beneficiary directly or indirectly, and are performed in daily events in which customers usually participate.” Such a distinction of function (process in the context of this paper) layer helps to characterize the respective processes. As described herein, we use the terms “base actions” and “meta-actions” rather than “base functions” and “meta-functions” because our emphases are mainly human actions for services.

Elderly Care

We interpret the actions which elderly people live in daily life as base actions in the elderly care domain. Elderly people want to live daily life independently. It is usually difficult for elderly people in a care facility because of their difficulties related to aging. Caregivers support them to live their life in many ways. The goal of the caregiver’s actions is to make the daily life of elderly people possible. Therefore, we interpreted caregiver actions as meta-actions in contrast to the actions of the elderly people. In this domain, we specifically examine construction of structured manuals of meta-actions.

Higher Education

We collaborated with a teacher who has classes at a university. She held an active learning class [15], the subject of which was domestic science. The students had some activities such as presentation or data collection. Then, they reflected on their experiences in the class and shared information with other students and the teacher. The students were evaluated based on the results of the reflection and their activities.

We interpret the action by which students reflect on their activities in the class as a base action in a higher education domain. In this case, the students used learning materials to clarify the reflection contents. The teacher provides structured manuals as the learning materials. The teacher action makes the action of the students possible. Therefore, providing the learning materials is interpreted as a meta-action. In this case, the structured manual content is neither base nor meta-action.

Local Revitalization Support

We interpret the action by which local leaders of a certain community hold some activities for local revitalization as a base action in the case. The local leaders must manage their organization to continue their activities for local revitalization. Such management action is interpreted as a meta-action here.

We also interpret the management action as another base action in this case. Ideally, local leaders can manage their activities by themselves. However, local leaders usually have no experience of such management and activities. Under such circumstances, Tokyo Metropolitan Institute of Gerontology orchestrate supporters in local community, such as municipality and non-profit organization, to support local leaders [10]. Such support action by the supporters in the local community enables management action by

the local leaders. Therefore, the support action can be interpreted as a meta-action for the management action. We specifically examine meta-actions of two types to realize the base action in this case. The contents of the structured manuals are meta-actions of these two types, i.e., the management action by local leaders and the support action by supporters in the local community.

Coimagination Method (Dementia Prevention)

Coimagination method is a method for preventing dementia [11]. The method requires several people. Participants bring some pictures according to the theme of coimagination event. Then, each presenter explains personal experiences while pointing at various images. It should be a one minute explanation. The experience should have occurred in the recent past. The audience should listen to the presenters talk while using imagination: as though the listeners were the presenter. The audience learns a way to think of the presenter through the listening session. After the talk, a QA session is held. Through these sessions, Otake et al. reported that cognitive abilities of the participants were increased or maintained. Coimagination event participation might prevent the onset and development of dementia.

We interpreted an action by which participants improve or maintain their cognitive abilities as a base action. Meta-action is participation in the coimagination event. The participation action enables the improvement of their cognitive abilities. The meta-action decreases the risk of dementia through participation in coimagination event.

We can regard another meta-action similarly to case 3. Coimagination event organizers enable the participants to join in the event. Therefore, the organizing action is interpreted as a meta-action for participation action by participants. We specifically examine construction of structured manuals of meta-actions of two types in this case.

Construction

We interpret actions which reduce the crucial risks in construction area as base actions in this case. The meta-actions are observation and decision. Such observation and decision based on the observation enable reduction of the crucial risks. We specifically examine construction of the meta-actions in this case.

Waste Treatment

We interpret the operating action of the plant in this case. The operator uses a crane to move garbage to a combustion furnace. We specifically examined the plant operation, including using the crane and observing the trash stockyard, in this case. The resulting structure is used for understanding the operation to consider which work can be replaced by automation.

Autonomous Vehicle

We interpret the action/function which the driver moves from location A to location B by driving a car, as a base action. In this case, we do not consider the meta-action. Details of construction knowledge were presented in an earlier report of the literature [12].

Guitar Rendition

We interpret the rendition action as a base action in this case. Iino et al. aim to clarify guitar rendition by structuring knowledge and using it to educate novice guitar players [13]. They also do not specifically examine the meta-action in this stage.

4.3 Discussion

Table 2 presents a comparison among cases for which we applied our proposed method from the viewpoint of contents. We can identify two cases as having meta-action of two types. We designate the meta-action to enable realization of other meta-action as meta²-action here. Meta²-action is observed when there are supporters and supportees. In these cases, ideal and desired actions by a supportee are interpreted as base actions. The supportee acts to realize the ideal action. Some obstacles usually exist to prevent the

Table 2. Comparison among cases from the viewpoint of content

No.	Domain	Meta ² -action	Meta-action	Base action	What we construct
1	Elderly care	–	Caregiver supports elderly people	Elderly people live their daily life	Meta-action
2	Higher education	–	Teacher provides learning materials	Students reflect on their activities	– (learning materials which students use)
3	Local revitalization support	Supporters support the management of local leaders	Local leaders manage their organization	Local leaders provide activities for local revitalization	Meta-action, Meta ² -action (to be constructed)
4	Coimagination method	Organizers organize coimagination events	Participants participate in coimagination event	Participants improve their cognitive abilities	Base action, Meta-action
5	Construction	–	Employees observe and make decisions.	Employees reduce the crucial risks in construction area	Meta-action
6	Waste treatment	–	–	Employees operate the plant	Base action
7	Autonomous vehicle	–	–	Driver moves from location A to location B	Base action
8	Guitar rendition	–	–	Guitar players play guitar	Base action

realization of the ideal action. Then, the supporter helps the supportee. The action performed by the supporter is interpreted as meta²-action. The hierarchy constituted with meta², meta, and base action represents interaction among these actions, in other words.

5 Information System for Knowledge Explication

As described in Sect. 3, five processes are related to knowledge explication. In this section, we point out future work for system development to support these processes. We are particularly addressing processes A, B, C, and D.

5.1 Knowledge Explication Augmenter

We are developing support systems for the processes A, B, and C. We designate the system as kNoWledge eXplication augmentaR (kNeXaR). It has a simple taxonomy editor which adds concepts and their relations. The system also has an instance editor for procedural knowledge, so-called structured manuals, and allows a user to modify the structured manuals with a concept that is already defined in the taxonomy editor. The user merely fills the slot to produce the structured manuals. As future work, we will consider how to support making the structured manuals such as information recommendation or automated validation using logical formalism.

5.2 Knowledge Based Dialogue System

We are considering a system that does process D using a spoken dialogue system based on the structured manuals. Following are dialogue examples introduced by Jokinen et al. [14].

U: Hello Yuki!

N: Hi Kristiina! How can I help you?

U: Well I'd need some help to change a person's position.

N: Ok, let's change the person's position. How would you like to move the person?

U: I want to make the person sit.

N: Ok. Have you checked the person's physical condition?

U: Yes, the person is conscious, and lying on the bed.

N: Ok. Follow my instructions. When you are ready, please say ready.

U: I'm ready now.

N: Ok. First move the person sideways. When this is completed, please say continue.

U: I'm ready to continue.

N: Ok, lift the person up pushing sideways and supporting the head. Then check that the person sits properly and feels good.

U: Just a minute please.

N: Ok, take your time.

U: The person is sitting properly.

N: Ok. Does the person seem to be fine?

U: Yes, the person seems to be fine.

N: Ok, very good. Then we are done.

U denotes a user who talks with the dialogue robot; N denotes Nao robot⁴. The knowledge used in the dialogue is described as the structured manual. There are some issues to realize such dialogue. For instance, keyword extraction from a user's utterance, information querying from the structured manuals, and the way to generate the sentences used in the dialogue.

6 Conclusion

We clarified the processes related to knowledge explication method and the characters of each domain to which we applied our proposal. We consider the action as state change. From that viewpoint, we can classify five processes related to knowledge explication. We also tried to clarify which domains are suitable for application to our proposal. First, we used the Japan Standard Industrial Classification to classify the domains. However, the classification is developed for industry, but our application domains include no commercial domain such as local revitalization. Therefore, we consider the domains from the viewpoint of contents which we construct. As a result, we found that there are supporters and supportees related to focused action in some domains. Our comparison revealed interaction of their actions. We also pointed out future work to develop a support system for knowledge explication. How we develop the system and use it is a subject left for our future work.

Acknowledgments. This paper is based partly on results obtained from the “Future AI and Robot Technology Research and Development Project” commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and JSPS KAKENHI Grant Number JP16K16160. The authors extend their gratitude to Hanamaru-kai Medical Corporation, and Shinko Fukushima Social Welfare Corporation for cooperation in construction of structured manuals and their evaluation.

References

1. Ministry of Health, Labour and Welfare: Annual report of long-term care insurance on 2013 (2013). in Japanese. <http://www.mhlw.go.jp/topics/kaigo/osirase/jigyo/14/index.html>. Accessed 02 Mar 2018
2. Ministry of Health, Labour and Welfare: Estimation of demand and supply of caregivers towards 2025. (2015). (in Japanese) http://www.mhlw.go.jp/file/04-Houdouhappyou-12004000-Shakaiengokyoku-Shakai-Fukushikibanka/270624houdou.pdf_2.pdf. Accessed 02 Mar 2018
3. Nishimura, T., Fukuhara, T., Yamada, C., Hamasaki, M., Nakajima, M., Miwa, H., Watanabe, K., Motomura, Y.: Proposal of handover system for care-workers using community intelligence. In: 1st International Conference on Serviceology, pp. 135–142 (2013)
4. Nishimura, S., Ohtani, H., Hatakeyama, N., Hasebe, K., Fukuda, K., Kitamura, Y., Mizoguchi, R., Nishimura, T.: Employee driven approach to “knowledge explication”, in elderly care service. *Trans. Jpn. Soc. Artif. Intell.* **32**(4), C-G95_1–15 (2017). (in Japanese)

⁴ <https://www.ald.softbankrobotics.com/en/robots/nao>.

5. Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R.D., Shadbolt, N.R., Van de Velde, W., Wielinga, B.J.: *Knowledge Engineering and Management the CommonKADS Methodology*. MIT Press, Cambridge (2000)
6. Nishimura, S., Kitamura, Y., Sasajima, M., Williamson, A., Kinoshita, C., Hirao, A., Hattori, K., Mizoguchi, R.: CHARM as activity model to share knowledge and transmit procedural knowledge and its application to nursing guidelines integration. *J. Adv. Comput. Intell. Intell. Inform.* **17**(2), 208–220 (2013)
7. Angeli, G., Premkumar, M.J., Christopher D.: Manning: leveraging linguistic structure for open domain information extraction. In: *53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 344–354 (2015)
8. Ushiku, Y., Yamaguchi, M., Mukuta, Y., Harada, T.: Common subspace for model and similarity: phrase learning for caption generation from images, In: *IEEE International Conference on Computer Vision*, pp. 2668–2676 (2015)
9. Sumita, K., Kitamura, Y., Sasajima, M., Mizoguchi, R.: Are services functions? In: *Third International Conference on Exploring Services Science, Lecture Notes in Business Information Processing*, vol. 103, pp. 58–72 (2012)
10. Kawai, H., Nishimura, S., Nishimura, T., Yoshida, Y., Ejiri, A., Honshima, A., Yasunaga, M., Fujiwara, Y., Ohbuchi, S.: Extraction and structuring procedural knowledge for support of preventive care using knowledge explication. In: *12th Annual Meeting of Society for Applied Gerontology Japan*, p. 58 (2017). (in Japanese)
11. Otake, M., Kato, M., Takagi, T., Asama, H.: Development of coimagination method towards cognitive enhancement via image based interactive communication. In: *18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 835–840 (2009)
12. Nishimura, S., Iwata, A., Kurokawa, M., Maruta, S., Kaji, D., Niwa, S., Nishimura, T., Ehara, Y.: Autonomous vehicle system based on law and case law using qualitative representation. In: *30th International Workshop on Qualitative Reasoning*, pp. 1–7 (2017)
13. Iino, N., Nishimura, S., Fukuda, K., Watanabe, K., Jokinen, K., Nishimura, T.: Development and use of an activity model based on structured knowledge – a music teaching support system. In: *Fifth International Workshop on the Market of Data*, pp. 1–6 (2017)
14. Jokinen, K., Nishimura, S., Watanabe, K., Nishimura, T.: Human-robot dialogues for explaining activities. In: *International Workshop on Spoken Dialog System Technology*, pp. 1–10 (2018, forthcoming)
15. Nishimura, S., Fukuda, K., Nishimura, T., Dohi, M.: Proposal of knowledge sharing framework for active learning and its application. In: *17th European Conference on Knowledge Management*, pp. 684–691 (2016)

Author Index

- Abe, Akinori 51
Akagi, Kaya 83
- Bloomfield, Robin E. 335
Butler, Alastair 299
- Chapa-Martell, Mario Alberto 375
- Dai, Qin 355
Deguchi, Hiroshi 83
- Frana, Ilaria 253
Fukuda, Ken 401
- Goebel, Randy 35
- Hara, Yurie 282
Hayashi, Shintaro 188
Henderson, R. 231
Horn, Stephen Wright 299
Hunter, Anthony 336
- Imamura, Mitsuyoshi 97
Inoue, Naoya 355
Inui, Kentaro 355
Ishino, Yoko 112
Izumi, Yu 188
- Kikuchi, Takamasa 140
Kim, Mi-Young 35
Kinugawa, Kazutaka 339
Kiselyov, Oleg 241
Kunigami, Masaaki 140
- Lafourcade, Mathieu 214
Lai, Regine 266
Liang, Zilu 375
Liefke, Kristina 171
Lu, Yao 35
Luk, Zoe Pei-sui 266
- McCready, Elin 231
Mery, Bruno 214
- Mirzapour, Mehdi 214
Moot, Richard 214
Moulton, Keir 253
- Nakagawa, Kei 97
Nakamura, Hiroaki 314
Nishimura, Satoshi 375, 401
Nishimura, Takuichi 375, 401
- Ohuri, Kotaro 67
Ohsato, Takaya 83
Orita, Naho 282
- Ray, Oliver 20
Reisert, Paul 355
Retoré, Christian 214
- Sakai, Hiromu 282
Shibata, Masashi 157
Shimada, Eriko 67
Suge, Aiko 129
Suzuki, Ren 112
- Takahashi, Hiroshi 129
Takahashi, Masakazu 157
Takahashi, Shingo 67, 112
Takiguchi, Itsuki 51
Tang, Liping 200
Terano, Takao 140
Tsuruoka, Yoshimasa 339
- Umino, Kazunori 140
- Valvoda, Josef 20
- Watanabe, Kentaro 390
Winterstein, Grégoire 266
- Yamada, Hiroaki 67
Yamada, Takashi 140
Yamane, Shohei 67
Yoon, Sungjae 129
Yoshida, Kenichi 97
Yoshioka, Masaharu 5