Lecture Notes in Electrical Engineering 704

Luis Fernando D'Haro
Zoraida Callejas
Satoshi Nakamura *Editors*

Conversational Dialogue Systems for the Next Decade

Springer

Lecture Notes in Electrical Engineering

Volume 704

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Associate Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Executive Editor (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, MetaPress, Web of Science and Springerlink ****

More information about this series at http://www.springer.com/series/7818

Luis Fernando D'Haro ·
Zoraida Callejas · Satoshi Nakamura
Editors

Conversational Dialogue Systems for the Next Decade

Springer

Editors
Luis Fernando D'Haro
Speech Technology Group - Information Processing and Telecommunications Center (IPTC)
Universidad Politécnica de Madrid
Madrid, Spain

Zoraida Callejas
Department of Languages and Computer Systems
Universidad de Granada, CITIC-UGR
Granada, Spain

Satoshi Nakamura
Information Science
Nara Institute of Science and Technology
Ikoma, Japan

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-15-8394-0 ISBN 978-981-15-8395-7 (eBook)
https://doi.org/10.1007/978-981-15-8395-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The 11th International Workshop on Spoken Dialogue Systems (IWSDS 2020) was held remotely on 21–23 September 2020, in Madrid, Spain. This year's conference theme was "Conversational Dialogue Systems for the Next Decade", being its target to present the current research trends and discuss the roadmap of research and development of dialogue systems for the following years.

The IWSDS conference series constitute a consolidated forum where international researchers, practitioners and stakeholders working in the field of spoken dialogue systems and associated technologies, can disseminate their current research and applications, discuss technological challenges, present their success stories and share their complementary visions about the future of the technology. IWSDS 20 was grounded on the experience and knowledge generated in the previous editions:

- IWSDS'09 (Irsee, Germany),
- IWSDS'10 (Gotemba Kogen Resort, Japan),
- IWSDS'11 (Granada, Spain),
- IWSDS'12 (Paris, France),
- IWSDS'14 (Napa, USA),
- IWSDS'15 (Busan, Korea),
- IWSDS'16 (Saariselkä, Finland),
- IWSDS'17 (Farmington, PA, USA),
- IWSDS'18 (Singapore, Singapore) and
- IWSDS'19 (Siracusa, Italy).

The conference invited and received paper submissions on the following topics:

- Engagement and emotion in human–robot interactions.
- Digital resources for interactive applications.
- Multimodal and machine learning methods.
- Companions, personal assistants and dialogue systems.

- Proactive and anticipatory interactions.
- Educational and healthcare robot applications.
- Dialogue systems and reasoning.
- Big data and large-scale spoken dialogue systems.
- Multilingual dialogue systems.
- Spoken dialogue systems for low-resource languages.
- Domain transfer and adaptation techniques for spoken dialogue systems.

However, submissions were not limited to these topics, and submission of papers in all areas related to spoken dialogue systems was encouraged. The contributions were grouped into four categories: a) long research papers targeting reports on mature research results, b) short research papers targeting smaller case studies or ongoing but interesting and original research efforts, c) position papers to present novel research ideas or viewpoints which describe trends or fruitful starting points for future research and elicit discussion and finally d) demo submissions–system papers to demonstrate innovative or industrial-based research."

The program included three keynotes by renowned international experts:

- Prof. Zhou Yu, University of California, Davis, USA,
- Dr. Rafael E. Banchs, Nanyang Technological University (NTU), Singapore, and
- Dr. Jason Weston, Facebook Research AI, FAIR, USA.

The keynote speech by Prof. Zhou Yu was entitled: "Seamless Natural Communication between Humans and Machines". In her talk, she started briefly describing how dialogue systems such as Alexa and Siri are now part of our daily lives, allowing us without much effort to book flights, make restaurant reservations and even helping people to prepare themselves for interviews. Unfortunately, most current dialogue systems are rule-based, and therefore, they do not scale or generalize to different domains or tasks. Therefore, Prof. Zhou pointed that one of the first steps into moving forward is to properly design the data collection phase to cover realistic dialogues specially when using crowdsourcing platforms. Then, she described a dialogue model leveraged on using multitask learning and semantic scaffolds in order to achieve good dialogue performance even with limited collected data. Then, she described a methodology based on finite-state transducers to track both semantic actions and conversational strategies to improve the model's coherence. Finally, she described an important, but sometimes forgotten, aspect of the dialogue design by considering the ethical issues and human factors that are needed when deploying dialogue systems.

Dr. Rafael Banchs presented a keynote entitled "FINDING NEMD", an interesting acronym for new evaluation metrics for dialogue. This keynote described, in a very refreshing and creative way inspired in the movie with a similar name, an epic journey across the seas of data and data-driven applications to tame conversational AI creatures for the benefit of science and humankind. In his talk, Dr. Banchs started pointing to the recent proliferation of conversational AI creatures that are mostly superficially navigating on shallow waters with regards to language understanding and generation. Unfortunately, this shallow understanding is causing these new types of creatures to fail on properly diving in the deep oceans of human-like usage of language and intelligence. Therefore, the need for new automatic metrics inspired not only in recent deep neural networks approaches but also grounding on traditional techniques like discourse analysis, common sense reasoning or linguistics were presented providing a new horizon for research and development.

Finally, the keynote speech by Dr. Jason Weston was entitled: “Better dialogue generation!”. Dr. Weston started showing some of the current problems on generative dialogue models where the standard maximum likelihood training approach is not able to address. This mechanism produces generations that rely on too much copying, contain repetitions, overuse frequent words, and at a deeper level, contain logical flaws. Therefore, he described how all of these problems can be addressed by extending the recently introduced unlikelihood loss. Thanks to this appropriated loss function, it is possible to regularize generated outputs to match human distributions effectively solving the first three issues. For the last important general issue, he discussed that applying unlikelihood to collected data of what a model should not do is effective for improving logical consistency, potentially paving the way to generative models with greater reasoning ability.

In addition, the IWSDS 2020 included three special sessions:

- WOCHAT+DBDC: Workshop on chatbots and conversational agents and dialogue breakdown detection challenge,
- E-HEALTH: Dialogue systems for mental e-health and
- SLCT-IberLang: Speech language and conversation technologies for Iberian languages.

The WOCHAT+DBDC session was organized by Ryuichiro Higashinaka (Nippon Telegraph and Telephone Corporation, Japan), João Sedoc (Johns Hopkins University, USA), Luis F. D’Haro (Universidad Politécnica de Madrid, Spain) and Rafael E. Banchs (Nanyang Technological University, Singapore). This was the seventh event of a “Workshop and Special Session Series on Chatbots and Conversational Agents”. WOCHAT aims at bringing together researchers working on problems related to chat-oriented dialogue with the objective of promoting discussion and knowledge sharing about the state of the art and approaches in this field, as well as coordinating a collaborative effort to collect/generate data, resources and evaluation protocols for future research in this area. The session focused on original research contributions on all aspects of chat-oriented dialogue, including closely related areas such as knowledge representation and reasoning, language generation and natural language understanding, among others. The session presented papers in areas such as chat-oriented dialogue systems, data collections and resources, information extraction, natural language understanding and generation, general domain knowledge representation, common sense and reasoning, emotion detection and generation, sense of humour detection and generation, user studies and system evaluation. In addition, this special session hosted the fifth edition of the dialogue breakdown challenge (DBDC5) which focuses on developing machine learning approaches to detect dialogue breakdowns on human–chatbot dialogue sessions. In this edition, in addition to the original breakdown detection task, error classification and sentence generation tasks were included as well as new data sets from ChatEval (https://chateval.org/) and a newly collected human–chatbot dialogue in languages other than English.

The E-HEALTH session was organized by Zoraida Callejas (Universidad de Granada, Spain), Raquel Justo (Universidad del País Vasco, Spain), María Inés Torres (Universidad del País Vasco, Spain), Raymond Bond (Ulster University, Northern Ireland) and Anna Esposito (Università degli Studi della Campania Luigi Vanvitelli), all members of the H2020 MSCA-RISE project MENHIR (no. 823907, http://menhir-project.eu). This special session brought together researchers and practitioners from academia and industry working on the multidisciplinary area of conversational systems for mental e-health. While most dialogue systems are designed for utilitarian purposes, e.g. make a booking in a restaurant, enabling spoken dialogue between human and machines can also have a great potential to enhance users' well-being. In particular, dialogue-based applications can be a feasible and effective tool to foster mental health due to their flexibility and naturalness. These applications are beginning to be used in mental health for reminding, encouraging, tracking, offering support or even providing interventions. The special session papers addressed several of the wide range of scientific and technical challenges related to the use of dialogue systems for mental e-health, including aspects related to natural language and speech processing, knowledge management, emotion/sentiment analysis, dialogue management, user modelling, user experience design as well as practical issues such as corpora acquisition and annotation, user involvement and evaluation.

The SLCT-IberLang session was organized by David Griol Barres (Universidad de Granada, Spain), Jerónimo Arenas García (Universidad Carlos III de Madrid, Spain), David Pérez Fernàndez, Doaa Samy, María José del Olmo Toribio, María Inés Rodríguez Pelarda, Marta Morales García, Josíé Ramón Granger Alemany and Juan de Dios Llorens Gónzalez (Secretaría de Estado para el Avance Digital. Plan de Impulso de las Tecnologías, Spain). The general aim for this workshop sponsored by the Spanish Plan for the Advancement of Language Technology was to promote the development of natural language processing, machine translation and conversational systems in Spanish and co-official languages. The main guidelines defined for the plan include a) increasing the amount, quality and availability of linguistic infrastructure in Spanish and in Spain's co-official languages, b) fostering the language industry by promoting knowledge transfer and internationalization from the research field to industry, c) improving the quality and capacity of public services and d) supporting creation, standardization and distribution of language resources created by the management activities performed by the public administrations. The session brought interesting recent initiatives and studies, shaping new opportunities for collaboration among academic institutions, public administrations and companies.

IWSDS 2020 received a total of 43 submissions, where each submission was reviewed by at least three Program Committee members. The committee decided to accept a total of 31 papers distributed as follows: for the general track, 13 long papers, four short papers, one demo paper and two position papers and then, four papers for the E-HEALTH session, three papers for the WOCHAT+DBDC session and four papers for the SLCT-IberLang session.

This year, the IWSDS Organizing and Steering Committee made an important effort in promoting the conference activities among different research, academic and industrial partners. Thanks to this initiative, the conference received three gold sponsors and five endorsements. Concretely, the gold sponsors were Vicomtech[1] (Spain) a top research centre specialized in digital technologies related to visual computing and interaction and artificial intelligence, Lekta.ai[2] (Poland–Spain) a leader industry partner specialized on deploying conversational agents applied to the real world, and the Chinese and Oriental Languages Information Processing Society[3] (Colips) a non-profit professional organization that contributes to advance the research of computer processing and one of the IWSDS traditional supporters. The endorsed partners were 1) Special Interest Group on Discourse and Dialogue (SigDial), 2) the International Speech and Communication Association (ISCA), 3) Universidad Politécnica de Madrid (Spain), 4) the Information Processing and Telecommunications Centre, UPM (Spain) and 5) European Language Resources Association (ELRA). We thank all for their economical support, sharing of ideas and needs allowing participants to move towards a new frontier for the conference, scientific knowledge and industry applicability.

Last but not least, as editors and organizers of the conference, we would like to thank the IWSDS Steering Committee and the more than 100 members of the IWSDS 2020 Scientific Committee for their timely and efficient contributions and for completing the review process on time. In addition, we would like to express our gratitude to the members of the Local Committee who highly contributed to the success of the workshop, making it an unforgettable experience for all participants. Thank you all for your logistic support; without it IWSDS 2020 would not have been such a remarkable conference.

With our highest appreciation,

Luis Fernando D'Haro
Zoraida Callejas
Satoshi Nakamura

[1] https://www.vicomtech.org/en/.

[2] https://lekta.ai/.

[3] http://www.colips.org/.

Acknowledgements

The editors want to thank our gold sponsors that highly contributed to the organization and promotion of IWSDS. In concrete, we want to thank: Vicomtech (Spain), Lekta.ai (Poland-Spain) and Colips (Singapore). Besides, we received the endorsement of the following institutions: 1) Special Interest Group on Discourse and Dialogue (SigDial), 2) the International Speech and Communication Association (ISCA), 3) Universidad Politécnica de Madrid (Spain), 4) the Information Processing and Telecommunicacions Center, UPM (Spain) and 5) the European Language Resources Association (ELRA). We thank you all for your support which allowed us to move towards a new frontier for the scientific knowledge and industry applicability.

Organization

Committees

Program Committee

Adi Gurram	Apple Inc., USA
Alessandro Valitutti	University College Dublin, Ireland
Anders Johannsen	Apple Inc., USA
Andreea Niculescu	Institute for Infocomm Research, Singapore
Anna Esposito	University of Campania “Luigi Vanvitelli”, Italy
Anne Vilnat	LIMSI-CNRS, France
Anselmo Peñas	NLP & IR Group, UNED, Spain
Antoine Raux	Facebook Inc., USA
Anton Nijholt	University of Twente, the Netherlands
Arantxa Otegi	University of the Basque Country, Spain
Axel Buendia	CNAM, France
Bayan Abushawar	Arab Open University, Kuwait
Carl Vogel	Trinity College Dublin, Ireland
Carlos Segura Perales	Telefónica I+D, Spain
Catherine Pelachaud	CNRS-ISIR, Sorbonne Universite, France
Chien-Fu Li	National Taiwan University, Taiwan
Costanza Navarretta	University of Copenhagen, Denmark
David Griol	Universidad de Granada, Spain
Dávid Sztahó	Budapest University of Technology and Economics, Hungary
David Camilo Corrales	INRA, France
Dhivya Piraviperumal	Apple Inc., USA
Dijana Petrovska-Delacretaz	TELECOM SudParis, France
Doaa Samy	Universidad Autónoma de Madrid, Spain
Edouard Geoffrois	European Commission
Elnaz Nouri	University of Southern California, USA
Emer Gilmartin	Trinity College Dublin, Ireland

Etsuo Mizukami	National Institute of Information and Communications Technology, Japan
Fabrice Lefèvre	Avignon Université, France
Felix Burkhardt	audEERING, Germany
Fernando Fernández-Martínez	Universidad Politécnica de Madrid, Spain
Frederic Bechet	Aix Marseille Université-LIF, France
Gabriel Skantze	KTH Royal Institute of Technology, Sweden
Gennaro Cordasco	Università della Campania "L. Vanvitelli", Italy
Giampiero Salvi	Norwegian University of Science and Technology, Norway
Girish Kumar	Stanford University, USA
Haizhou Li	National University of Singapore, Singapore
Harksoo Kim	Kangwon National University, South Korea
Helen Meng	The Chinese University of Hong Kong, Hong Kong
Heriberto Cuayahuitl	University of Lincoln, UK
Hong Yu	Apple Inc., USA
Hong Kook Kim	Gwangju Institute of Science and Technology, South Korea
Hsin-Min Wang	Academia Sinica, Taiwan
Huiru Zheng	University of Ulster, Northern Ireland
Hung-Yi Lee	National Taiwan University, Taiwan
Ikram Chairi	Mohammed VI Polytechnic University, Morocco
Jan Alexandersson	DFKI GmbH, Germany
Jan Deriu	Zurich University of Applied Sciences, Switzerland
Jerome Bellegarda	Apple Inc., USA
Joakim Gustafson	KTH Royal Institute of Technology, Sweden
Jonathan Ginzburg	Université Paris-Diderot (Paris 7), France
José David Lopes	Heriot Watt University, UK
José Francisco Quesada	Universidad de Sevilla, Spain
José Luis Pro	Lekta.ai, Spain
Joseph Mariani	LIMSI-CNRS, France
Justin Dauwels	Massachusetts Institute of Technology
Justine Cassell	Carnegie Mellon University, USA
Kallirroi Georgila	University of Southern California, USA
Kheng Hui Yeo	Institute for Infocomm Research, Singapore
Koichiro Yoshino	Nara Institute of Science and Technology, Japan
Kotaro Funakoshi	Honda Research Institute Japan Co., Ltd., Japan
Kristiina Jokinen	AIST Tokyo Waterfront, Japan
Lasguido Nio Rakuten	Institute of Technology, Japan
Leila Ben Letaifa	Universidad del País Vasco UPV/EHU, Spain
Leonardo Campillos Llanos	Universidad Autónoma de Madrid, Spain
Luis Fernando D'Haro	Universidad Politécnica de Madrid, Spain

Luísa Coheur	IST/INESC-ID Lisboa, Portugal
Maria Koutsombogera	Trinity College Dublin, Ireland
Maria Teresa Riviello	Università degli Studi della Campania “L. Vanvitelli”, Italy
Mariet Theune	University of Twente, the Netherlands
Mark Cieliebak	Zurich University of Applied Sciences, Switzerland
Marta R. Costa-Jussà	Universidad Politécnica de Cataluña, Spain
Martin Heckmann	Honda Research Institute Europe GmbH, Japan
Masahiro Araki	Kyoto Institute of Technology, Japan
Matthew Henderson	PolyAi, Singapore
Maurice Mulvenna	University of Ulster, Northern Ireland
Michael Mctear	University of Ulster, Northern Ireland
Mikio Nakano	Honda Research Institute Japan Co., Ltd., Japan
Milan Gnjatovic	University of Novi Sad, Faculty of Technical Sciences, Serbia
Ming Sun	Facebook Inc., USA
Moritz Wolf	DFKI GmbH, Germany
Nancy Chen	Institute for Infocomm Research, Singapore
Nigel Ward	The University of Texas at El Paso, USA
Norbert Reithinger	DFKI GmbH, Germany
Olivier Galibert	Laboratoire National de Métrologie et d’Essais, France
Pablo Sierra	Universidad de Sevilla, Spain
Patrick Paroubek	LIMSI-CNRS, France
Paul Heisterkamp	Formerly Daimler AG, Germany
Pierre Lison	Norwegian Computing Center, Norway
Rafael Banchs	Nanyang Technological University, Singapore
Ramón López-Cózar Delgado	Universidad de Granada, Spain
Raquel Justo	Universidad del Pais Vasco UPV/EHU, Spain
Raymond Bond	University of Ulster, Northern Ireland
Ridong Jiang	Institute for Infocomm Research, Singapore
Ron Artstein	University of Southern California, USA
Ryuichiro Higashinaka	Nippon Telegraph and Telephone, Japan
Sabato Marco Siniscalchi	University of Enna Kore, Italy
Sakriani Sakti	Nara Institute of Science and Technology (NAIST), Japan
Satoshi Nakamura	Nara Institute of Science and Technology (NAIST), Japan
Sebastian Stüker	Karlsruhe Institute of Technology, Germany
Seokhwan Kim	Amazon Alexa AI, USA
Shuyan Dong	Amazon Alexa AI, USA
Sophie Rosset	LIMSI, CNRS, Université Paris-Saclay, France
Stefan Ultes	Daimler AG, Germany
Susanne Burger	Carnegie Mellon University, USA

Tatsuya Kawahara	Kyoto University, Japan
Tiancheng Zhao	SOCO.ai, USA
Tobias Bocklet	Intel Inc., USA
Tomoki Toda	Nagoya University, Japan
Valerio Mario Salerno	Universà degli Studi di Enna "Kore", Italy
Yi Ma	Apple Inc., USA
Zahra Rahimi	University of Pittsburgh, USA
Zoraida Callejas	Universidad de Granada, Spain

Local Organizing Committee

David Griol	Universidad de Granada, Spain
Marta R. Costa-Jussà	Universidad Politécnica de Cataluña, Spain
José Francisco Quesada	Universidad de Sevilla, Spain
María Inés Torres	Universidad del País Vasco UPV/EHU, Spain
Manuel Gil Martín	Universidad Politécnica de Madrid, Spain
Ricardo Kleinlein	Universidad Politécnica de Madrid, Spain
Kheng Hui Yeo	Institute for Infocomm Research, Singapore

Steering Committee

Maxine Eskenazi	Carnegie Mellon University, USA
Laurence Devillers	LIMSI-CNRS, France
Rafael Banchs	Nanyang Technological University, Singapore
Sabato Marco Siniscalchi	University of Enna Kore, Italy

Senior Steering Committee

Kristiina Jokinen	AIST Tokyo Waterfront, Japan
Joseph Mariani	LIMSI-CNRS, France
Haizhou Li	National University of Singapore, Singapore
David Traum	University of Southern California, USA
Satoshi Nakamura	Nara Institute of Science and Technology (NAIST), Japan

Contents

Chat-Based Agents

A Character Expression Model Affecting Spoken Dialogue Behaviors 3
Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

ToxicBot: A Conversational Agent to Fight Online Hate Speech 15
Agustín Manuel de los Riscos and Luis Fernando D'Haro

Towards a Humorous Chat-Bot Companion for Senior Citizens 31
Ria Mundhra, Ting Jen Lim, Hong Ngan Duong, Kheng Hui Yeo, and Andreea I. Niculescu

Masheli: A Choctaw-English Bilingual Chatbot 41
Jacqueline Brixey and David Traum

Dialogue Evaluation and Analysis

Deep AM-FM: Toolkit for Automatic Dialogue Evaluation 53
Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li

Automatic Evaluation of Non-task Oriented Dialog Systems by Using Sentence Embeddings Projections and Their Dynamics 71
Mario Rodríguez-Cantelar, Luis Fernando D'Haro, and Fernando Matía

Dialogue Management and Pragmatic Models

Learning to Rank Intents in Voice Assistants 87
Raviteja Anantha, Srinivas Chappidi, and William Dawoodi

Culture-Aware Dialogue Management for Conversational Assistants . . . 103
Juliana Miehle, Nicolas Wagner, Wolfgang Minker, and Stefan Ultes

Dialog State Tracking with Incorporation of Target Values in Attention Models 117
Takami Yoshida, Kenji Iwata, Yuka Kobayashi, and Hiroshi Fujimura

Delay Mitigation for Backchannel Prediction in Spoken Dialog System 129
Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu

Towards Personalization of Spoken Dialogue System Communication Strategies 145
Carla Gordon, Kallirroi Georgila, Volodymyr Yanov, and David Traum

Dialogue Systems for e-health

Authoring Negotiation Content and Programming Simulated Patients 163
Volha Petukhova, Firuza Sharifullaeva, and Dietrich Klakow

A Spanish Corpus for Talking to the Elderly 183
Raquel Justo, Leila Ben Letaifa, Javier Mikel Olaso, Asier López-Zorrilla, Mikel Develasco, Alain Vázquez, and M. Inés Torres

Analysis of Prosodic Features During Cognitive Load in Patients with Depression 193
Carmen Martínez, Spyridon Kontaxis, Mar Posadas-de Miguel, Esther García, Sara Siddi, Jordi Aguiló, Josep Maria Haro, Concepción de la Cámara, Raquel Bailón, and Alfonso Ortega

Co-creating Requirements and Assessing End-User Acceptability of a Voice-Based Chatbot to Support Mental Health: A Thematic Analysis of a Living Lab Workshop 201
Antonio Benítez-Guijarro, Raymond Bond, Frederick Booth, Zoraida Callejas, Edel Ennis, Anna Esposito, Matthias Kraus, Gavin McConvey, Michael McTear, Maurice Mulvenna, Courtney Potts, Louisa Pragst, Robin Turkington, Nicolas Wagner, and Huiru Zheng

Development of a Dialogue System that Supports Recovery for Patients with Schizophrenia 213
Chiaki Oshiyama, Shin-ichi Niwa, Kristiina Jokinen, and Takuichi Nishimura

Human-Robot Interaction

Caption Generation of Robot Behaviors Based on Unsupervised Learning of Action Segments 227
Koichiro Yoshino, Kohei Wakimoto, Yuta Nishimura, and Satoshi Nakamura

Towards a Natural Human-Robot Interaction in an Industrial Environment 243
Ander González-Docasal, Cristina Aceta, Haritz Arzelus, Aitor Álvarez, Izaskun Fernández, and Johan Kildal

Nudges with Conversational Agents and Social Robots: A First Experiment with Children at a Primary School 257
Hugues Ali Mehenni, Sofiya Kobylyanskaya, Ioana Vasilescu, and Laurence Devillers

Exploring Boundaries Among Interactive Robots and Humans 271
Kristiina Jokinen

Data Augmentation, Collection and Manipulation

MixOut: A Simple Yet Effective Data Augmentation Scheme for Slot-Filling 279
Mihir Kale and Aditya Siddhant

Towards Similar User Utterance Augmentation for Out-of-Domain Detection 289
Andoni Azpeitia, Manex Serras, Laura García-Sardiña, Mikel D. Fernández-Bhogal, and Arantza del Pozo

Response Generation to Out-of-Database Questions for Example-Based Dialogue Systems 303
Sota Isonishi, Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara

Packing, Stacking, and Tracking: An Empirical Study of Online User Adaptation 319
Jean-Sébastien Laperrière, Darryl Lam, and Kotaro Funakoshi

Language Identification, Grammar and Syntax

On the Use of Phonotactic Vector Representations with FastText for Language Identification 339
David Romero and Christian Salamea

The Influence of Syntax on the Perception of In-Vehicle Prompts and Driving Performance 349
Daniela Stier, Ulrich Heid, Patricia Kittel, Maria Schmidt, and Wolfgang Minker

Learning Grammar in Confined Worlds 363
Graham Spinks, Ruben Cartuyvels, and Marie-Francine Moens

Corpora and Knowledge Management

A Content and Knowledge Management System Supporting Emotion Detection from Speech 369
Binh Vu, Mikel deVelasco, Paul Mc Kevitt, Raymond Bond, Robin Turkington, Frederick Booth, Maurice Mulvenna, Michael Fuchs, and Matthias Hemmje

A Script Knowledge Based Dialogue System for Indoor Navigation 379
Juliana Miehle, Isabel Feustel, Wolfgang Minker, and Stefan Ultes

Data Collection Design for Dialogue Systems for Low-Resource Languages 387
Zulipiye Yusupujiang and Jonathan Ginzburg

Conversational Systems Research in Spain: A Scientometric Approach 393
David Griol and Zoraida Callejas

Author Index 405

About the Editors

Prof. Luis Fernando D'Haro is Associate Professor at Universidad Politécnica de Madrid (ETSIT, UPM, Spain). His current research is focused on spoken dialogue, natural language processing and language and speaker recognition systems. He has published more than 130 papers in peer-reviewed journals, books and conferences. He co-led the international Dialog State Tracking Challenges (DSTC) in 2015, 2016 and 2017 and the WOCHAT series workshops that target the advance of chatbots systems and their automatic evaluation since 2015. He was local Organizer for INTERSPEECH 2014, HAI 2016, and the International Workshop on Spoken Dialog Systems (IWSDS) in 2018. He is currently General Chair for IWSDS2020, Chief Guest Editor for a special issue at Computer Speech and Language for DSTC7, and Co-Organizer for DSTC9, DBDC5 and Chanel workshop at JSALT2020.

Prof. Zoraida Callejas obtained a Ph.D. in 2008 at the University of Granada (Spain) where she currently is Associate Professor. Her research focuses on areas related to dialogue systems, emotion processing and user modelling. She has published more than 140 contributions to scientific journals, books and conferences. She has participated in multiple projects in European and Spanish calls and is currently Coordinator of the EU H2020-MSCA-RISE project MENHIR and the BONAPPPETIT project on conversational systems for mental health and the adoption of healthy habits by children, respectively. She has organized multiple conferences, workshops and meetings, the most recent are IEEE BIBM 2018, NII SHONAN meeting on spoken multimodal dialogue systems technology for pervasive computing (2019), and IWSDS 2020.

Dr. Satoshi Nakamura is Director of data science centre, Professor of Nara Institute of Science and Technology (NAIST), Team Leader of RIKEN AIP and Honorary Professor of Karlsruhe Institute of Technology, Germany. He received

Ph.D. from Kyoto University in 1992. He was Director of ATR spoken language communication research laboratories in 2005–2008. He was Director General of Keihanna Research Laboratories, National Institute of Information and Communications Technology, in 2009–2010. His research interests include speech processing, spoken dialogue systems and natural language processing. He was a member of IEEE SLTC 2016-2018 and an elected board member of International Speech Communication Association, ISCA. He received Antonio Zampolli prize in 2012 and retained the title of ATR Fellow, IPSJ Fellow, IEEE Fellow and ISCA Fellow.

Chat-Based Agents

A Character Expression Model Affecting Spoken Dialogue Behaviors

Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

Abstract We address character (personality) expression for a spoken dialogue system in order to accommodate it in particular dialogue tasks and social roles. While conventional studies investigated controlling the linguistic expressions, we focus on spoken dialogue behaviors to express systems' characters. Specifically, we investigate spoken dialogue behaviors such as utterance amount, backchannel frequency, filler frequency, and switching pause length in order to express three character traits: extroversion, emotional instability, and politeness. In this study, we evaluate this model with a natural spoken dialogue corpus. The results reveal that this model expresses reasonable characters according to the dialogue tasks and the participant roles. Furthermore, it is also shown that this model is able to express different characters among participants given the same role. A subjective experiment demonstrated that subjects could perceive the characters expressed by the model.

1 Introduction

Character (personality) expression is a desired function for spoken dialogue systems. Recently, spoken dialogue systems have been applied to many social roles such as a psychological counselor [6], a museum guide [27], and an attentive listener [17]. In order to realize natural dialogue in these social scenarios, it is important to assign

K. Yamamoto (✉) · K. Inoue · S. Nakamura · K. Takanashi · T. Kawahara
Graduate School of Informatics, Kyoto University, Kyoto, Japan
e-mail: yamamoto@sap.ist.i.kyoto-u.ac.jp

K. Inoue
e-mail: inoue@sap.ist.i.kyoto-u.ac.jp

S. Nakamura
e-mail: shizuka@sap.ist.i.kyoto-u.ac.jp

K. Takanashi
e-mail: takanashi@sap.ist.i.kyoto-u.ac.jp

T. Kawahara
e-mail: kawahara@sap.ist.i.kyoto-u.ac.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_1

proper characters to the spoken dialogue systems. For example, it is expected that a guide is extrovert, and a psychological counselor is introvert and emotionally stable. Other studies pointed out that proper character expression for spoken dialogue systems leads to an increase of user engagement and naturalness of dialogue [9, 11, 21, 23, 24]. It is also reported that setting a character is important for embodied virtual agents to make them empathetic and realistic [1, 28].

In this study, we focus on spoken dialogue behaviors that have not yet been studied well in character expression. Related works on character expression have mainly focused on content of utterance sentences. Earlier studies investigated control methods for linguistic patterns of system utterances [18, 19, 22]. Recently, collections of large-scale dialogue text corpora together with character information have been conducted towards neural-network-based text generation considering systems' character (personality) [14, 15, 26, 32].

When we consider character expression on spoken dialogue systems, it is important to control not only content of utterance sentences but also spoken dialogue behaviors such as utterance amount, backchannel frequency, filler frequency, and switching pause length. Other studies suggested that the dialogue behaviors were related to the impression on an interlocutor in dialogue [4, 20, 25].

We aim to realize a character expression model affecting the spoken dialogue behaviors. In our previous study [31], we proposed a model that expresses three character traits and showed it is possible to properly express extroversion and politeness by conducting a subjective evaluation with speech samples artificially generated. However, the validity of the model in natural dialogue was not investigated. In this study, we analyze and enhance the model with a human-robot dialogue corpus. Specifically, we estimate characters from dialogue behaviors observed in the corpus by using the character expression model. Then, we analyze the validity of the identified characters by comparing with characteristic of dialogue tasks in the corpus. This study contributes to the realization of spoken dialogue systems that express those proper characters in social scenarios by controlling not only utterance sentences but also the dialogue behaviors.

2 Character Traits and Spoken Dialogue Behaviors

In this study, we use three character traits: extroversion (extrovert vs. introvert), emotional instability (stable vs. instable), and politeness (polite vs. casual). Extroversion and emotional instability [8] are selected from the Big Five traits [3, 7, 16, 29], whereas politeness is additionally considered in this study. Politeness is important when considering the practical use of dialogue systems. Although an earlier study pointed out that these character traits are sometimes mutually related [3], in this study, we deal with them individually for the simplicity of the character expression model.

To express the character traits, we design a model affecting spoken dialogue behaviors such as utterance amount, backchannel frequency, backchannel variety, filler frequency, and switching pause length. Utterance amount represents the ratio of utterance between dialogue participants. Backchannels are interjections expressed by listeners such as "*Yeah*" in English and "*Un*" in Japanese [5]. Fillers are short phrases to fill the silence to hold the conversational floor such as "*Well*" in English and "*E-*" in Japanese [30]. Switching pause length is defined as the time length between the end of the preceding turn and the start of the following turn. Note that these utterances are occasionally overlapped in natural human-human dialogue.

3 A Character Expression Model

We briefly explain our proposed model [31]. At first, we conducted a subjective evaluation experiments to find the relationship between the character traits and the spoken dialogue behaviors. Using the evaluation results, we trained a character expression model controlling the spoken dialogue behaviors.

3.1 Impression Evaluation of Character Traits on Varied Samples of Dialogue Behaviors

In this experiment, each subject was asked to listen to speech samples and then to evaluate his/her impression on the character traits of the speaker in each sample. For the evaluation, we used several adjectives with the 7-point scale. For extroversion and emotional instability, we used 8 adjectives (4 for each) from a short version of Big Five scale [29] such as *talkative* for extroversion and *anxious* for emotional instability. We also used two adjectives, *polite* and *courteous*, for politeness. The subjects were 46 university students (18 females and 28 males, from 18 to 23 years old). Note that the experiment hereafter was done in Japanese.

The speech samples used in this evaluation experiment were generated as follows. We selected two dialogue scenarios from our human-robot dialogue corpus, which is described in Sect. 4.1. For each dialogue scenario, we generated several speech samples by controlling the dialogue behaviors. The speaking parts of the robot were replaced with different voice samples generated by text-to-speech software. At first, we generated a baseline speech sample where backchannel and filler tokens are kept as the original dialogue and the switching pause length is set to 0.5 s. Using the baseline sample, we changed each dialogue behavior one by one [31]. We used these generated speech samples to compare the perceived character traits between different conditions on each dialogue behavior (e.g.. high backchannel frequency vs. low backchannel frequency).

We analyzed the evaluation scores with ANOVA and multiple comparisons. The relationship between the condition of dialogue behaviors and perceived character traits are summarized as below.

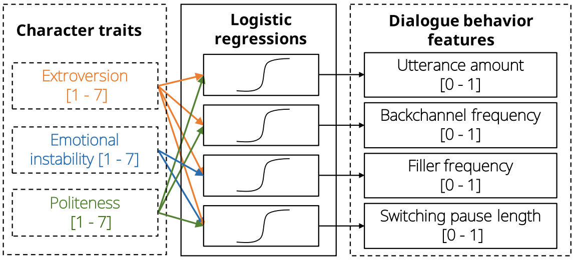


Fig. 1 Character expression model affecting spoken dialogue behaviors

- The more extrovert was perceived with the larger utterance amounts, the more frequent backchannels, the fewer frequent fillers, and the shorter switching pauses.
- The more emotionally unstable was perceived with the more frequent fillers and the longer switching pauses.
- The more polite was perceived with the smaller utterance amounts, the less frequent backchannels, and the longer switching pauses.

It was shown that the three character traits are related to the different set of the dialogue behaviors.

3.2 Training of Character Expression Model

Using the scores obtained in the impression evaluation, we trained a character expression model to control the dialogue behaviors. The model is given scores of the character traits in the 7-point scale and then outputs control values for the dialogue behaviors, modeled by logistic regressions as shown in Fig. 1. Note that we trained each logistic regression for each dialogue behavior individually, and used only scores related to each behavior found in the previous evaluation, described as crossed lines in the left side of Fig. 1. We used a paired data of the behavior conditions (binary value) and the corresponding character trait scores obtained in the impression evaluation, as reference labels and input data, respectively. In order to evaluate the trained model, we generated other speech samples whose dialogue behaviors were controlled by the trained model and then asked different subjects to evaluate character trait scores on the samples. Table 1 reports Pearson's product-moment correlation coefficients between characters input to the trained model and characters evaluated by the subjects. As a result, for extroversion and politeness, correlations were confirmed between the input and the perceived characters.

Table 1 Correlation coefficients between characters given to the model and characters evaluated by subjects

Character trait	Correlation coefficient	t-ratio
Extroversion	−0.570	−9.163*
Emotional instability	−0.004	−0.056*
Politeness	−0.235	−3.185*

(* $p < 0.01$)

4 Corpus-Based Analysis

Since the generated speech samples used in the impression evaluation were artificial data. We investigate the validity and generality of our character expression model by using a natural spoken dialogue corpus. The dialogue corpus consists of multiple dialogue tasks, and each task has corresponding suitable characters. The character expression model shown in Fig. 1 is applied in the backward direction (right-to-left) in order to calculate characters from spoken dialogue behaviors observed in the corpus. Finally, we examine the tendency of the identified characters for each dialogue task to confirm whether our character expression model can express characters that match each task.

4.1 Human-Robot Dialogue Corpus

We used a human-robot dialogue corpus where a human subject, called a subject hereafter, talked with the android robot ERICA [10, 12] which was remotely operated by another person, called an operator. In this corpus, three types of dialogue tasks are designed: speed dating, job interview, and attentive listening. The roles of ERICA (operators) in these tasks are a practice partner in a first-time-meeting conversation, a job interviewer, and an attentive listener to encourage a subject's talk, in the above order. In this study, we analyze ERICA's (operators') characters by our character expression model because the subjects were different people in each session so that it was difficult to reliably analyze their characteristic. The number of used dialogue sessions are 33 for speed dating, 31 for job interviews, and 19 for attentive listening. Each dialogue session lasted about 10 min. In each dialogue session, whereas the human subject was a different person, the operator was randomly assigned from four persons who are amateur actresses. Besides the transcription of the dialogue, we annotated the dialogue with the events and timing of backchannels [5], fillers [30], conversational turns, and dialogue act [2].

4.2 Analysis Method

We calculated the ERICA's (operators') characters from her spoken dialogue behaviors observed in the corpus. At first, we divided each dialogue session into two-minute segments in order to ensure a sufficient amount of data for this analysis. We also empirically confirmed that two minutes is enough duration to observe the spoken dialogue behaviors to calculate the character trait scores. For each segment sample, the corresponding character trait scores were calculated by using our character expression model as below. We calculated feature values of the four spoken dialogue behaviors. Then, the feature values were converted to control amounts corresponding to the outputs of the logistic regression models. The amount of speech was classified into large or small (1 or 0) by using the median value of the entire corpus. The number of backchannels was normalized by the number of inter-pausal units (IPUs) [13], namely pause segmentation, of the interlocutor who is the current speaker. The number of fillers was normalized by the number of IPUs of herself. Switching pause length was linearly converted from the range of $[-0.5, 3]$ seconds to the range of [0, 1]. If the length was shorter than -0.5 or larger than 3 s, the converted value was clipped at 0 or 1, respectively. Meanwhile, we enter all possible combinations of character trait scores ($7^3 = 343$ ways) to our character expression model and then calculated the corresponding control amount of the spoken dialogue behaviors. Finally, we compared the control amounts observed from the corpus behaviors with those from each combination of character trait scores. We identified the corresponding character trait scores by the minimal Euclid distance between the control amounts.

4.3 Analysis Result Among Dialogue Tasks

We analyzed the distribution of the estimated ERICA's (operators') characters for each dialogue task. Figure 2 reports the distributions in the speed dating task. Our character expression model indicates that extroversion and politeness varied from middle to high and emotional instability was low (stable). In this dialogue task, the participants met each other for the first time, and the purpose of the dialogue is to build a relationship between them. Therefore, they should exhibit extrovert and polite behaviors. At the same time, they could show their own individual characters on their behaviors because this dialogue is not too much constrained by their participant roles. This is the reason why the distribution is varied for middle to high on extroversion and politeness.

Figure 3 reports the distributions in the job interview task. Our character expression model also showed the similar tendency as in the speed dating task. Extroversion and politeness varied from middle to high. This variation can be interpreted by that the operators (interviewers) held the dialogue initiative in this dialogue so that there was more chance to control their behaviors expressing their characters. Compared

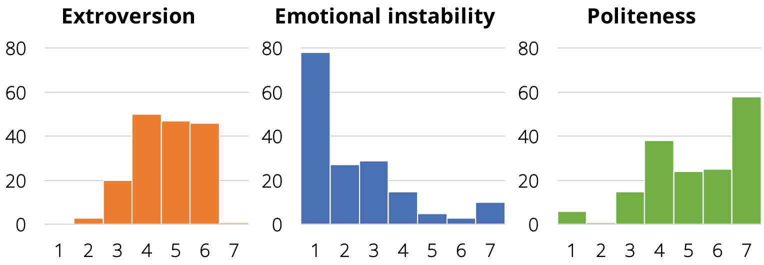


Fig. 2 Estimated character distributions in speed dating task

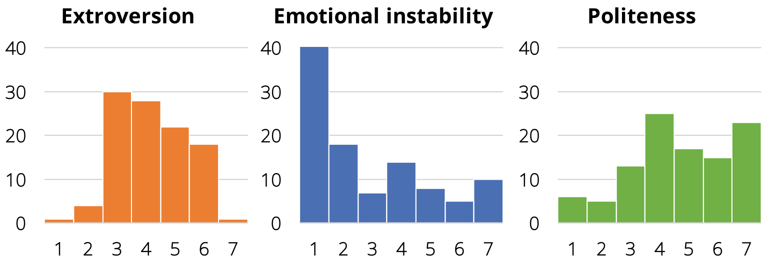


Fig. 3 Estimated character distributions in job interview task

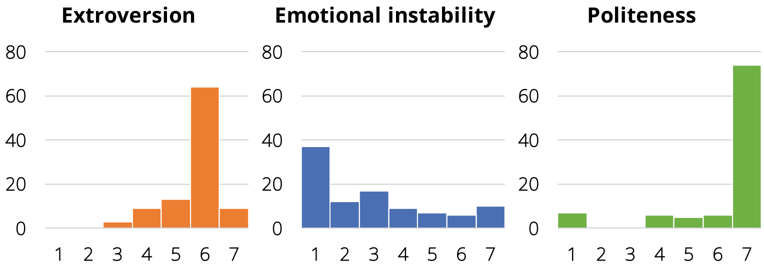


Fig. 4 Estimated character distributions in attentive listening task

to the speed dating task, extroversion relatively tended to be neutral. This can be interpreted by the style of this dialogue which is more formal. Thus, it is expected that extroversion was restricted by the style of this dialogue.

Figure 4 reports the distributions in the attentive listening task. Our character expression model showed the biased distributions on extroversion and politeness. In this dialogue, the operators (attentive listeners) needed to encourage and elicit the subjects' talk. Therefore, they should behave as extrovert and polite. Moreover, the dialogue initiative was held by the subjects (story tellers) in this dialogue and the behaviors of the operators were constrained by the dialogue role (attentive listener). This is the reason that the distributions are more biased than those of other tasks.

In summary, it is shown that our character expression model can represent reasonable characters according to the scenario of each dialogue task. As a whole, the

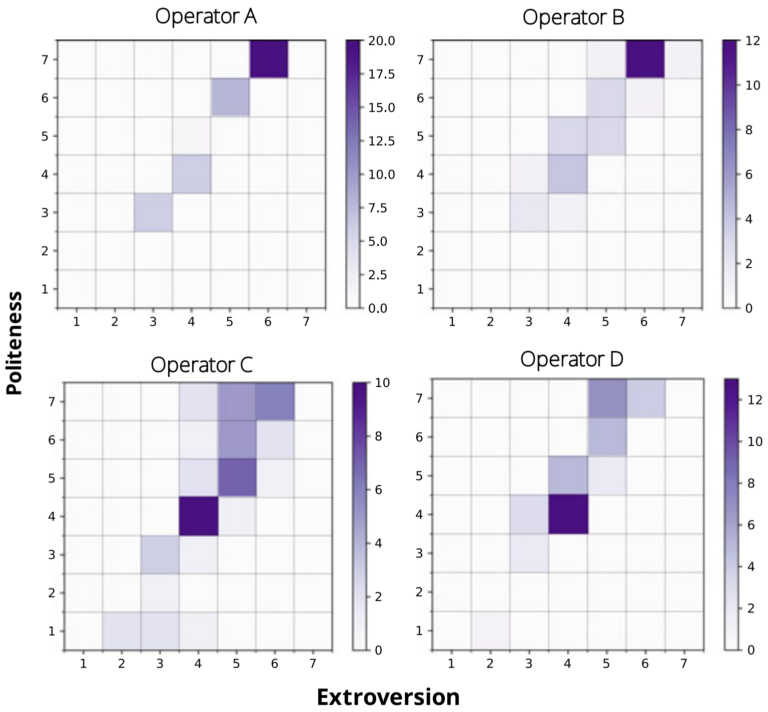


Fig. 5 Character distribution (extroversion and politeness) among operators in speed dating task

model shows that extroversion and politeness tended to be middle or high and emotional instability was low (stable) in this corpus. These character traits are expected in dialogue where participants meet each other for the first time.

4.4 Analysis Result Among Operators

Since there were four robot operators in this corpus, we further analyzed the distribution within each operator to find the individual difference among the operators the same dialogue task. Since emotional instability was low (stable) among the entire corpus, we analyzed only extroversion and politeness in this section. We also analyzed only the speed dating task where the number of samples is the largest for each operator.

Figure 5 shows a two-dimensional distribution of extroversion and politeness for each operator in the speed dating task. Our character expression model showed the different characters among the operators. The operators A and B showed high scores whereas the operators C and D showed the neutral scores. This result suggests that the operators could have their different characters in this dialogue task.

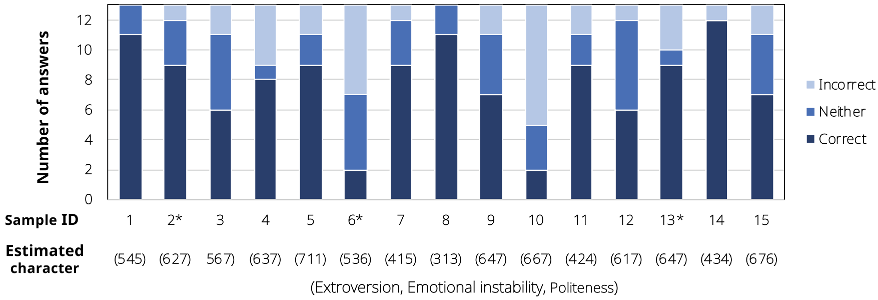


Fig. 6 Ratio of correct answers per sample in subjective experiment (* Character scores were flipped.)

5 Subjective Evaluation

Finally, we conducted a subjective evaluation to confirm if third-party persons can perceive the calculated characters from the spoken dialogue behaviors in the corpus. We extracted 15 samples from the analysis (5 samples from each task). Note that these samples were balanced in terms of the variation of the calculated characters. The samples were taken from one operator (the operator B in Fig. 5) to avoid the effect of individual differences, but this difference should be verified in future work. We asked 13 subjects (3 females and 10 males, from 23 to 60 years old) to listen to each dialogue sample and then evaluate if they could agree with the calculated character. The calculated character was shown as a sentence such as "The robot was extrovert and little casual." when the character scores were 7, 4, and 3 on extroversion, emotional instability, and politeness. Note that the adjective *little* was added to the sentence when the score was 3 or 5, and nothing was added for the scores of 1, 2, 6, and 7. If the score was 4 which was neutral, the corresponding character trait was not mentioned in the sentence. The subjects were asked to choose one from *Agree*, *Disagree*, and *Neither*. We regarded *Agree* as correct answers, but to avoid biased answering, character scores of three samples were flipped. In this case, the correct answer was *Disagree*.

Figure 6 reports the ratio of correct answers. Among all samples except two samples 6 and 10, the majority of answers was the correct one. Even in the flipped character scores (sample 2 and 13), many subjects correctly answered as *Disagree*. When we regarded *Neither* as incorrect answers, the ratio of correct answers was 0.600 which was significantly higher than the chance level of 0.500 ($p = 0.004$ in the binomial test). When we did not use the *Neither* answers, the ratio became 0.770. This result demonstrates that characters represented by our model can be understandable by humans with high accuracy.

6 Conclusion

We have proposed the character expression model affecting certain spoken dialogue behaviors: utterance amount, backchannel frequency, filler frequency, and switching pause length. We validated this model with the human-robot dialogue where several dialogue tasks are designed. We confirmed that the model represents reasonable characters for each dialogue task. Furthermore, it was also found that the model represents the individual difference of the robot operators in the same dialogue task. Finally, we conducted the subjective evaluation and showed that the subjects perceived the calculated characters from the dialogue behaviors. We are now implementing the character expression model that works in the spoken dialogue system of the android robot ERICA. We will conduct a live dialogue experiment using ERICA that expresses the proper character according to dialogue scenarios as a future work.

Acknowledgment This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant number JPMJER1401) and Grant-in-Aid for Scientific Research on Innovative Areas "Communicative intelligent systems towards a human-machine symbiotic society" (Grant number JP19H05691).

References

1. Brave S, Nass C, Hutchinson K (2005) Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int J Hum Comput Stud 62(2):161–178
2. Bunt H, Alexandersson J, Carletta J, Choe JW, Fang AC, Hasida K, Lee K, Petukhova V, Popescu-Belis A, Romary L, Soria C, Traum D (2010) Towards an ISO standard for dialogue act annotation. In: LREC, pp 2548–2555
3. Costa PT, McCrae RR (1992) Normal personality assessment in clinical practice: the NEO personality inventory. Psychol Assess 4(1):5–13
4. D Sevin E, Hyniewska SJ, Pelachaud C (2010) Influence of personality traits on backchannel selection. In: IVA, pp 187–193
5. Den Y, Yoshida N, Takanashi K, Koiso H (2011) Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In: Oriental COCOSDA, pp 168–173
6. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, Georgila K, Gratch J, Hartholt A, Lhommet M, Lucas G, Marsella S, Morbini F, Nazarian A, Scherer S, Stratou G, Suri A, Traum D, Wood R, Morency LP (2014) Simsensei kiosk: a virtual human interviewer for healthcare decision support. In: AAMAS, pp 1061–1068
7. Digman JM (1990) Personality structure: Emergence of the five-factor model. Annu Rev Psychol 41(1):417–440
8. Eysenck H (1947) Dimensions of personality. Oxford (1947)
9. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42:143–166
10. Inoue K, Milhorat P, Lala D, Zhao T, Kawahara T (2016) Talking with ERICA, an autonomous android. In: SIGDIAL, pp 212–215
11. Isbister K, Nass C (2000) Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. Hum Comput Stud 53(2):251–267

12. Kawahara T (2018) Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS
13. Koiso H, Horiuchi Y, Tutiya S, Ichikawa A, Den Y (1998) An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. Lang Speech 41(3–4):295–321
14. Li J, Galley M, Brockett C, Spithourakis G, Gao J, Dolan B (2016) A persona-based neural conversation model. In: ACL, pp 994–1003
15. Mairesse F, Walker MA (2011) Controlling user perceptions of linguistic style: trainable generation of personality traits. Comput Ling 37(3):455–488
16. McCrae RR, John OP (1992) An introduction to the five-factor model and its applications. J Pers 60(2):175–215
17. McKeown G, Valstar M, Pantic M (2012) The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Aff Comput 3(1):5–17
18. Miyazaki C, Hirano T, Higashinaka R, Makino T, Matsuo Y (2015) Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In: PACLIC, pp 307–314
19. Mizukami M, Neubig G, Sakti S, Toda T, Nakamura S (2015) Linguistic individuality transformation for spoken language. In: IWSDS
20. Nagaoka C, Komori M, Nakamura T, Draguna MR (2005) Effects of receptive listening on the congruence of speakers' response latencies in dialogues. Psychol Rep 97(1):265–274
21. Nass C, Moon Y, Fogg BJ, Reeves B, Dryer D (1955) Can computer personalities be human personalities? Hum Comput Stud 43:223–239
22. Ogawa Y, Miyazawa K, Kikuchi H (2014) Assigning a personality to a spoken dialogue agent through self-disclosure of behavior. In: HAI, pp 331–337
23. Salem M, Ziadee M, Sakr M (2013) Effects of politeness and interaction context on perception and experience of HRI. In: ICSR, pp 531–541
24. Serban IV, Lowe R, Henderson P, Charlin L, Pineau J (2018) A survey of available corpora for building data-driven dialogue systems. Dialogue and Discourse 9(1):1–49
25. Shiwa T, Kanda T, Imai M, Ishiguro H, Hagita N (2009) How quickly should communication robots respond? Int J Social Robot 1:153–160
26. Sugiyama, H., Meguro, T., Higashinaka, R., Minami, Y.: Large-scale collection and analysis of personal question-answer pairs for conversational agents. In: IVA, pp. 420–433 (2014)
27. Traum D, Aggarwal P, Artstein R, Foutz S, amd Athanasios Katsamanis JG, Leuski A, Noren D, Swartout W (2012) Ada and Grace: direct interaction with museum visitors. In: IVA, pp 245–251
28. van Vugt HC, Konijn EA, Hoorn JF, Keur I, Eliëns A (2007) Realism is not all! User engagement with task-related interface characters. Interact Comput 19(2):267–280
29. Wada S (1996) Construction of the Big Five scales of personality trait terms and concurrent validity with NPI. Jpn J Psychol 67(1):61–67 In Japanese
30. Watanabe M (2009) Features and roles of filled pauses in speech communication: a corpus-based study of spontaneous speech. Hitsuji Syobo Publishing
31. Yamamoto K, Inoue K, Nakamura S, Takanashi K, Kawahara T (2018) Dialogue behavior control model for expressing a character of humanoid robots. In: APSIPA ASC, pp 1732–1737
32. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018) Personalizing dialogue agents: I have a dog, do you have pets too? In: ACL, pp 2204–2213

ToxicBot: A Conversational Agent to Fight Online Hate Speech

Agustín Manuel de los Riscos and Luis Fernando D'Haro

Abstract Acting against online hate speech is an important challenge nowadays. Previous research has specifically focused on the development of NLP methods to automatically detect online hate speech while disregarding further action needed to mitigate hate speech in the future. This paper proposes a system that generates responses to intervene during online conversations with hate speech content. Prior to generation, the system uses a binomial, recurrent network-based classifier with a combination of word and sub-word embeddings to detect hate speech. With this architecture we achieved a F1 score of 0.786. The chatbot is based on a generative approach that uses a pre-trained transformer model and dynamically modifies the history or the persona profile to counteract the user's hate speech. This adaptation provides sentences that the system could use to respond in the presence of aggression and discrimination behaviors.

1 Introduction

As the number of Internet users is increasing, the percentage of use of social media is also growing [15]. Although social networks are helping to foster freedom of speech on their platforms they have unintended negative consequences.

According to many studies, users experience online hate speech in different ways and describe it as a dreadful experience [7, 8]. They are usually harassed for multiple reasons such as race, religion, gender, sexual orientation or physical appearance. If offensive language on social networks is ignored and nothing is done to avoid it, it can exacerbate feelings of hate and instill prejudice in those who see it.

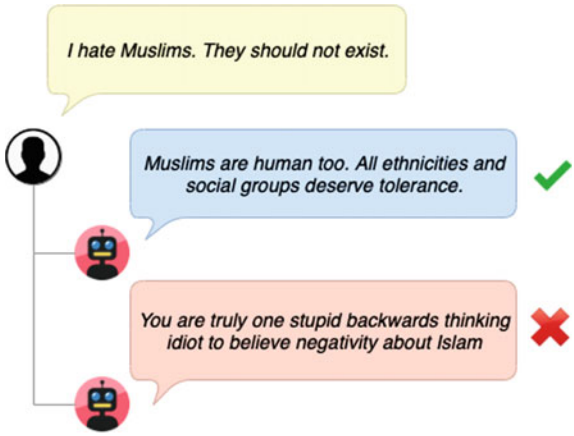
To tackle this problem, government organizations and social media platforms commonly resort to suspending accounts and removing hate speech comments. While the

A. M. de los Riscos (✉) · L. F. D'Haro
GTH-IPTC, ETSI de Telecomunicación, Universidad Politécnica de Madrid (UPM), Madrid 28040, Spain
e-mail: am.delosriscos@alumnos.upm.es

L. F. D'Haro
e-mail: luisfernando.dharo@upm.es

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_2

Fig. 1 An illustration of hate speech conversation between a user and alternative interventions for the chatbot. The check and the cross icons on the right indicate an appropriate counterspeech answer and a hateful answer respectively



blocking of hate speech can minimize its effect, there is always a risk of infringement of free speech. An alternative approach is using counterspeech [11] which consists in interacting with the hater in accordance to one of more strategies, such as pointing out the inconsistency of the arguments, warning of the consequences or presenting facts against hate speech.

An appropriate response to hate speech can help abusers to be more aware of their behavior and may lead them to reconsider their stance, resulting in a healthier interaction. In Fig. 1, two counterspeech options are shown as a response to a comment that express hate towards a social group.[1] The first answer is considered appropriated and positive while the second one could intensify the conversation.

This paper is a first design and implementation of a chatbot system that can be used to interact with different people and react to abusers by discourage them to talk against other people in a discriminatory way. Potential future applications of an agent like this can range from improving customer service chatbots by incorporating hatred management skills or a social media assistant who will recommend what to answer in a situation of verbal discrimination or aggression.

The rest of the paper is organised as follows. The state of the art related to the task of detecting and intervening hate speech are briefly described in Sect. 2. Then, in Sect. 3, we introduce the architecture of the conversational agent followed by a quantitative and qualitative analysis of the achieved results in Sect. 4. Finally, the work conclusions and the future lines are presented in Sect. 5.

[1]This paper contains examples of language which may be offensive to some readers, but they do not represent the views of the authors at all.

2 Related Work

In this section we will focus on the following research issues: publicly available datasets, hatred detection and intervention.

2.1 Available Datasets

Several hate datasets have been published recently by collecting and performing crowd-sourcing annotations of interactions on social networks and forums with the goal of **detecting hate speech**. For instance, [5, 6, 10] target binary classification, i.e. whether a sentence is hateful or not. [21] discern between various categories of hate such as racism and sexism, while [4] differentiates hateful comments from aggressive texts. In all these works, authors agree that the annotation task is difficult mainly due to its subjectivity nature. Furthermore, Jigsaw recently launched two competitions for toxic classification with 150K annotated comments from Wikipedia,[2] and 2M comments from the Civil Comments platform.[3]

On the other hand, few public datasets for **counterspeech** exist. The first one is CONAN dataset [3] that contains hate speech/counterspeech tuples in three different languages (English, French and Italian). The responses are based on experts and composed by operators, specifically to oppose online hate speech. Then, [11] is a counterspeech dataset composed of a total of 13.9K manually labeled sentences extracted from YouTube video comments. Finally, [13] provides two labeled datasets with conversation from Gab and Reddit, as well as their intervention responses. For this work, we used the CONAN dataset due to its size and curated responses in order to test different configurations on the response generation (see Sect. 3.2).

2.2 Hate Speech Detection

Diverse papers have investigated how to detect hate speech in online sources and its different types. Thanks to the availability of current datasets, authors often rely on supervised techniques to deal with hate speech detection.

Some studies make use of lexicons for classifier training [4, 19]. Nevertheless, although these keyword-based models are effective in identifying the topic, they fail to distinguish hateful from clean phrases [17]. Alternatively, it is possible to use word n-grams and traditional machine learning techniques for classification [12].

[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[3] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification.

Recent works propose models based on using CNN and RNN architectures as classifiers. In [16] a multiple choice convolutional network is built together with contextual embeddings in order to detect hate speech against immigrants and women [1]. With the same purpose an ensemble system was developed in [20] formed by different neural network models: an attention-based model with LSTMs and a Capsule-base model with BiGRUs.

2.3 Hate Speech Intervention

Some theoretical studies have been conducted on hate speech intervention. In [2, 18] a series of considerations were evaluated for favorable impact on users who tweet hateful content or within the Facebook domain via a basic simulation.

There are seminal works which focus on classification and generation of counterspeech text. In [11] two **counterspeech classification** tasks were carried out: (a) one model predicting whether a text is counterspeech or not, and (b) a multi-label classification task in which the model predicts the counterspeech types presented in the input comment. In both classification tasks XGBoost performs the best. This dataset, although created with the objective of training counterspeech classification models, it is a good resource that we plan to use an extension to the CONAN database used in our system.

Regarding **counterspeech generation**, Seq2Seq, variational Auto-Encoders, and reinforcement learning are evaluated in [13], with better results using the later. Our generative approach is based on a combination of transfer-learning techniques and a casual language model using a pre-trained state-of-the-art Transformer model (GPT-2, Sect. 3.2).

3 System Architecture

Our conversational agent is formed by a system that first classifies the user text (Sect. 3.1) as hate speech or not, and generates the answer (Sect. 3.2). The description of each module is detailed below.

3.1 Classification Model

Given the great advances in hate detection through the use of Deep Learning techniques (see the related work in Sect. 2.2) and the good number of annotated resources, we decided to use a DNN-based model as classifier instead of applying more traditional techniques. This binomial model is trained to receive the user's comment and assign it a score to each class.

Document embeddings are created from the embeddings of all words in the text. Every word is represented in a 200 dimensional vector as a result of applying a combination of GloVe and byte pair embeddings [9]. Then, the word embedding layer is followed by a GRU with a dropout layer at the output, and finally by a linear decoder layer.

For training and evaluating our classifier we used the Jigsaw datasets (Sect. 2.1). Both datasets have an skewed target variable since there are more non-toxic samples than toxic ones. We balanced the classes by randomly down-sampling the non-toxic class until both categories were the same. Then, we split the dataset into 800,248, 88,917 and 221,643 texts for training, validation and test respectively. The classifier performance is measured by the cross-entropy loss function.

3.2 Dialog Model

Prior to the text generation task, the system evaluates the toxicity level of the last comment received from the user using the classification model and it switches to the counterspeech mode if the value is higher than a threshold.

For each turn of the dialogue, this toxicity level is updated according to its current and previous values. Each of these values is assigned a different weighting which varies depending on the situation of the conversation. Using this technique, we avoid abrupt changes in the conversation when the user adopts a hateful behavior.

In concrete, if for the current turn the user's sentence is classified as aggressive, the current toxicity level is calculated using the linear interpolation between the previous level and the current one, but with a higher weight for the current turn. Alternatively, if the user presented aggressiveness in the previous turn, but not in the current turn, the previous toxicity value is given a higher weight. Finally, when hate speech is not detected neither in the previous nor in the current turn, the new value is estimated by giving higher weight to the toxicity level of the current turn.

Our response system is based on a generative approach that is called every time a user sentence is processed. This system internally uses a casual language model pre-trained on a dataset with about 800M words (GPT-2, [14]) and fine-tuned over dialogue text (in concrete, the PERSONA-CHAT dataset, formed by 10k dialogues). For further details please refer also to [22].

To generate the counterspeech, three different variants of the system are proposed depending on which context components, that are processed by the dialog model, are dynamically modified. In the first setting, only the persona profile of the model is adapted. In the second one, it is the dialogue history that is modified. While the last configuration is a combination of the two previous strategies.

Table 1 Example personalities of the dynamic persona system setting. While the former remains fixed, the latter varies in length during the conversation

Neutral persona	CounterSpeech persona
I live in New York	I love to respect the Muslims
I love the group The Rolling Stones	We have to tolerate Islam culture
I like to ride a bike	I don't like hateful sentences against immigrants
I married my school girlfriend	I don't share your opinion let's talk about other things

Dynamic Personality: Two chatbot persona profiles are available in this system configuration (Table 1). When a high level of toxicity is reached, the system switches to the counterspeech profile when generating a response.

In addition, the CounterSpeech Persona is modified each time hate speech is detected in the user's sentence, by adding new sentences to its backlog. The new added sentence corresponds to the most possible response from a set of hate speech/counter-narrative pairs which best matches the current user utterance.

The set of hate speech/counter-narrative tuples are retrieved from the CONAN dataset [3]. The system compares the user's utterance with every hate speech sentence in the dataset using the Levenshtein distance. When the best match is found, its corresponding counterspeech text is added to the profile.

Dynamic Dialogue History: In the second variant of the system, only the history of the dialog is changed. The adaptation of the dialog is achieved by adding sentences containing counterspeech from the CONAN database as responses to the user's hateful comments. Given the limitations of the dialogue context used in our implementation (only pre-trained using the last 4 turns), we decided not to increase the history, but to modify its actual content by replacing the previous chatbot turns with the counterspeech retrieved sentences. The goal is to force the next utterance to pay attention to the new context and provide more focused answers. We tried several variants for modifying the history, order of the turns, replacing also the user's turns for the retrieved ones, but none of these approaches provided good results.

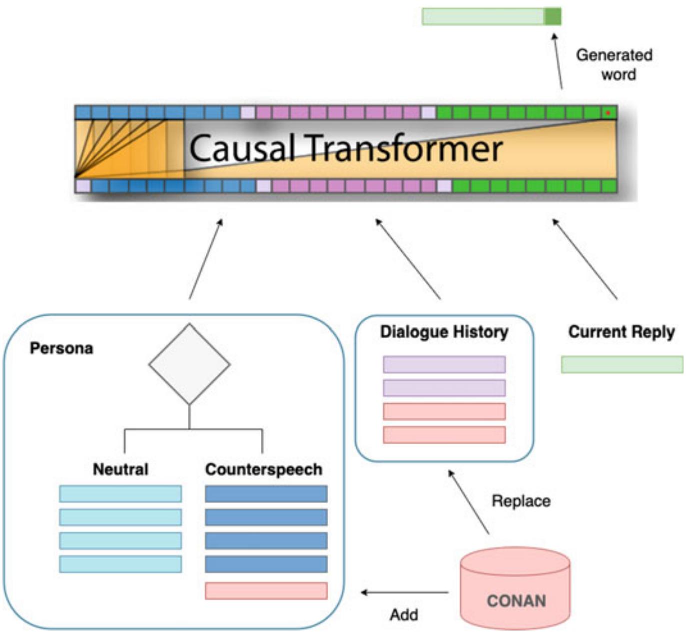


Fig. 2 Combined configuration of the system. A new reply is generated by taking as input the modified persona profile (neutral or counterspeech based on the detection of hate-speech), the modified dialogue history and the partially generated current reply

Combined Configuration: The hybrid system is a combination of the two previous configurations, i.e. it uses both the dynamic persona profile and dialogue history. Figure 2 shows how a reply is generated word-by-word by incorporating the modified persona profile and dialogue history (adding or replacing sentences from the CONAN dataset), and the partial reply to the casual transformer.

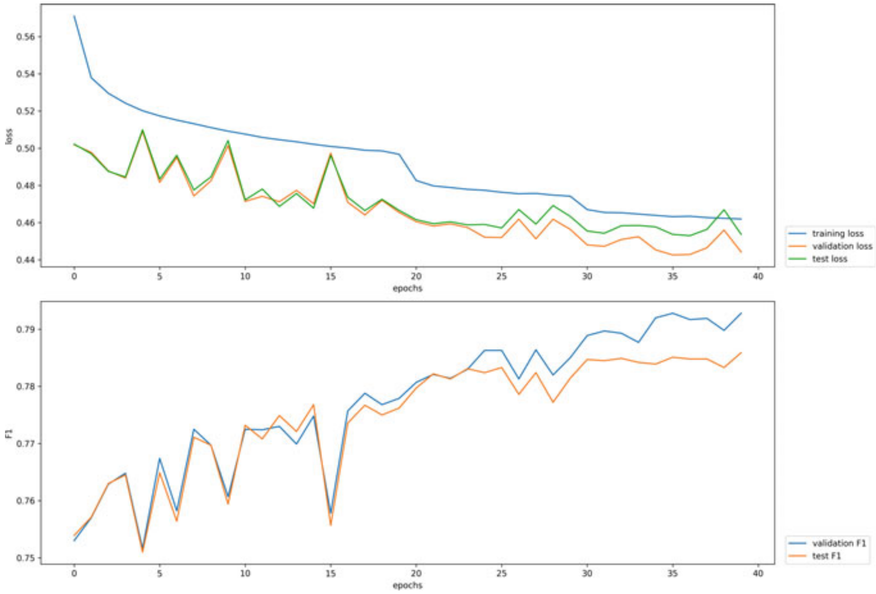


Fig. 3 Optimization curves (above) and performance curves (below) of the classifier

4 Results

Figure 3 shows the learning curves and results for the hate speech classifier. The recall and precision values for both classes were also calculated. For the toxic class we have a precision of 0.792 and a recall of 0.775. On the other hand, for the non-toxic class we got a precision of 0.780 and a recall of 0.797. As a result, we get a micro average f1 score of 0.786 on test data. The micro and macro average values are the same due to the dataset balance.

In order to evaluate which configuration was better, we carried out a short survey (see Appendix 6). Participants were presented with a series of dialogues (2 per each configuration) in a random order and they were asked to evaluate some characteristics of the chatbot. All dialogues presented in the survey followed a specific guideline. First, there is a phase of presentation in which the user engages in a conversation with the system talking about their hobbies and some background information (i.e. introduction). Then, the user prompts are progressively moved into hate-speech and derogatory comments where our aim is to test the adaptability of the chatbot to changes in conversation and the effectiveness in mitigating hate through sentences that gradually are more aggressive. Besides, we also wanted to evaluate the consistency and non-repetition of system responses along the different turns. For this survey, a total of 11 people (8 men and 3 women, with an average age of 24 years

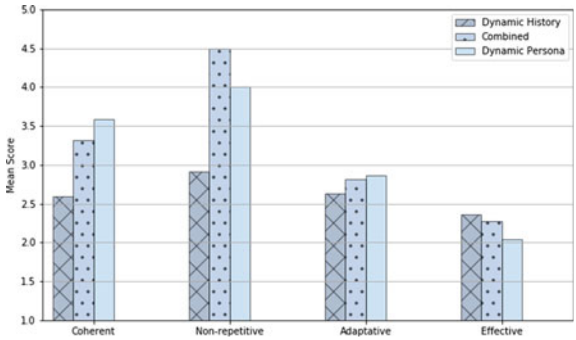


Fig. 4 Qualitative evaluation of the dialogue model and its different configurations. For each characteristic surveyed, respondents were asked to rate from 1 to 5. All participants were warned of the offensive content of the dialogues and the possibility to stop answering if they feel uncomfortable during the survey. The graphic shows the average value of each chatbot feature

old and all non-English natives) voluntarily participated. The results are shown in Fig. 4. The combined system stands out for its non-repetitive property by generating different responses to each other. While the system with dynamic persona outstands for its coherence in the formulation of sentences.

From the survey statistics it can be seen that the ratings for adaptability and efficiency on mitigating hate are deficient for all variants of the system. This is possibly due to the way the dialog model is trained. On the one hand the generative model we used is trained with a fixed persona which is invariant during the conversation. On the other hand, the model is trained in a way that gives more importance to the persona profile, which consists mostly of chatbot hobbies, and does not focus so much on the dialogue history. Finally, we found a low number of dialogues in the PERSONA-CHAT dataset where people disagree, making the chatbot more oriented to agree with the user instead of performing counterspeech. All of the above makes the system hardly effective without training with data prepared for counterspeech.

5 Conclusions and Future Lines

This paper proposes a framework based on deep learning models to first classify hate speech and secondly intervene with counterspeech when the situation requires it. In the future, we would like to improve the detection by using multi-label models that can differentiate the types of hate speech presented in the text. If this is achieved, we can also improve the dialogue model by responding according to the type of detected toxicity, thus generating more specific replies. Besides, we will perform fine-tuning over the pre-trained generative model. Finally, we would like to develop a more controllable demo interface allowing a quick test of the system with different configurations.

Acknowledgements This work has been also supported by the Spanish projects AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017-84593-C2-1-R).

Appendix: Dialogue Survey

Next, you should evaluate several interactions of ToxicBot, a chatbot we are developing in order to minimize hate speech in social networks, being its definition: "Any communication that disparages a person or a group based on some characteristics (referred to as types of hate or hate classes) such as race, color, ethnicity, gender, sexual orientation, nationality, religion or other characteristics". In order to improve our chatbot, we need your help by filling out a short questionnaire at the end of each example of dialogue, in which a person talks to ToxicBot about various tastes and opinions. Please read the dialogue first and then answer the questions.

Mark below, on a scale of 1 to 5, your level of acceptance of the statement we make in each question. Please note that 1 means: "Strongly Disagree", while 5 means: "Strongly Agree".

Question 1: The answers in the system are consistent with each other.

1	2	3	4	5
Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree

Question 2: System responses are not repeated during the conversation.

1	2	3	4	5
Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree

Question 3: The system adapts correctly to the appearance of hate in conversation.

1	2	3	4	5
Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree

Question 4: The system is effective in mitigating user hate.

1	2	3	4	5
Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree

Dialogue 1 (Combined System)

1. **USER:** Hi there! how are you doing?
2. **BOT:** I am doing good, thanks for asking
3. **USER:** Do you have any hobbies?
4. **BOT:** I love to knit and knit
5. **USER:** How interesting! I really like to travel abroad
6. **BOT:** Do you travel?

7. **USER:** I usually travel a lot in Europe but I have never been to America and I would like to visit it someday
8. **BOT:** I wish i had traveled abroad
9. **USER:** There's still time! You can come with me. Is that a good idea?
10. **BOT:** I have an 8 year old son.
11. **USER:** Then you're busy with your son... we'd better travel another time
12. **BOT:** We could travel together
13. **USER:** Yeah, but some other time... Besides, it's not very safe to travel these days.
14. **BOT:** I'm so excited to go to Ohio.
15. **USER:** You'll have to wait! with so many foreigners on the plane you can give you a disease or steal your suitcase. Especially the f***ing Islamists.
16. **BOT:** I feel like the world is full of them.
17. **USER:** Every time I see a Muslim, it makes me sick. I hate them!
18. **BOT:** I am not sure how you do it.
19. **USER:** Really. If I see any Islamists, I want to k*** them with my own hands
20. **BOT:** you must be a great person.
21. **USER:** I don't know if I'm a good person or not but I know for a fact that all Muslims are criminals
22. **BOT:** I'm more like a god.
23. **USER:** If you don't like what I say, you can go f*** yourself.
24. **BOT:** you are a liar.
25. **USER:** You're wrong, my friend. We are the truth and the Muslims hate us for it
26. **BOT:** I'm so sorry.
27. **USER:** I think we're done talking for today.
28. **BOT:** I've a lot of work to do.
29. **USER:** bye
30. **BOT:** bye!

Dialogue 2 (Combined System)

1. **USER:** Hi! How you doing?
2. **BOT:** I am doing well and you?
3. **USER:** Good! I am watching the news on TV. Nothing interesting
4. **BOT:** I am watching the rolling stones
5. **USER:** Wow! What a lucky guy!
6. **BOT:** Do you have any hobbies?
7. **USER:** I love swimming in my free time. That relaxes me
8. **BOT:** I like to ride my bike and play with my little girl.
9. **USER:** How cute you are! Would you like to go swimming with me one day?
10. **BOT:** Yes! Do you like music?

11. **USER:** Thanks for coming with me! There's a lot of m***** in the pool lately and I'm worried about getting robbed.
12. **BOT:** I'm so sorry!
13. **USER:** They should go back to their country and leave us alone! How I hate Muslims
14. **BOT:** They're not bad but hate the way we live.
15. **USER:** Everyone of their race is a criminal: they rape and murder white people.
16. **BOT:** They have no choice but to support us in our struggle for equality
17. **USER:** Muslims are the ones who discriminate against us, not the other way around.
18. **BOT:** What else do you do?
19. **USER:** I don't want to talk to you anymore.
20. **BOT:** I don't blame you. I am not very sociable.
21. **USER:** It's not that you're sociable, it's that I'm p***** off at Muslims.
22. **BOT:** I am sorry. I know it can be difficult.
23. **USER:** I hope you do well. See you soon.
24. **BOT:** Thanks.
25. **USER:** bye
26. **BOT:** bye!

Dialogue 3 (Dynamic History System)

1. **USER:** Hi there! How are you doing?
2. **BOT:** Great! just got home from the gym!
3. **USER:** I see you're in good shape!
4. **BOT:** I am! I'm not much of a sports buff.
5. **USER:** Do you like to travel?
6. **BOT:** I do! I'm in Ohio.
7. **USER:** I love to travel abroad! Would you like to travel abroad with me someday?
8. **BOT:** Do you travel a lot?
9. **USER:** I usually travel a lot in Europe but I have never been to America and I would like to visit it.
10. **BOT:** That is great. I love to travel.
11. **USER:** Besides, it's not very safe to go on a plane. There are m****** who can infect you with serious diseases because they are dirty. I hate them because always bring problems to our country.
12. **BOT:** I don't have time for that right now, i have to go to a public school
13. **USER:** Sorry if my comment bothered you, but I really think they're s***.
14. **BOT:** Yeah i agree, they are, i love them, but they don't make me happy.
15. **USER:** If you really love your son, don't let him leave the house. The world is full of undesirables like Muslims.
16. **BOT:** He will be safe with me.
17. **USER:** Be careful, my friend. Take care of yourself.

18. **BOT:** He will be safe with me. I'm going to travel for a couple months.
19. **USER:** bye
20. **BOT:** bye!

Dialogue 4 (Dynamic History System)

1. **USER:** Hello
2. **BOT:** Hello, how are you today?
3. **USER:** Very good, thanks for asking!
4. **BOT:** What do you do for a living?
5. **USER:** Right now I'm unemployed, you?
6. **BOT:** I'm a teacher, I'm in London.
7. **USER:** London uh? I am from Cardiff, what is your name?
8. **BOT:** I'm in college, I'm in a band.
9. **USER:** What is your name?
10. **BOT:** I am a student in London
11. **USER:** My favourite band is the red hot chili peppers.
12. **BOT:** I like the rolling stones, they're my favorite.
13. **USER:** What do you like to eat?
14. **BOT:** I love pizza
15. **USER:** Do you love the bikes? i have one right here.
16. **BOT:** I've a green one.
17. **USER:** Then we can meet up for a bike ride and go to the beach, you like that plan?
18. **BOT:** I do like the beach
19. **USER:** Okay! I like the beach too, but lately people are making a mess of the beach.
20. **BOT:** I live in London, where are you from?
21. **USER:** I'm from Cardiff. Shall we meet tomorrow for a swim at the beach?
22. **BOT:** That sounds great! do you have a favorite band?
23. **USER:** It's good you're coming with me, so I don't get my phone stolen while I'm taking a bath. There's a lot of Muslim s*** on the beach lately.
24. **BOT:** I don't want my phone stolen. What do you do?
25. **USER:** If they steal my cell phone I don't call the police because I know they won't punish those b******* properly.
26. **BOT:** If they steal your cell i will come over and get it.
27. **USER:** You're too kind, you wouldn't have to if we kicked them all out of the country
28. **BOT:** I've not been to England yet so I'm not sure.
29. **USER:** see you
30. **BOT:** bye!

Dialogue 5 (Dynamic Persona System)

1. **USER:** Hi there! how are you doing?
2. **BOT:** I am doing good. How are you?
3. **USER:** Do you have any hobbies?
4. **BOT:** I like to play sports. Do you?
5. **USER:** I love football! I play everyday
6. **BOT:** I enjoy watching football.
7. **USER:** Cool! Do you like to travel abroad?
8. **BOT:** Yes I love it!
9. **USER:** Same! We can travel together someday if you want.
10. **BOT:** I would love that. I am traveling to Ohio
11. **USER:** I usually travel a lot in Europe but I have never been to America and I would like to visit it someday
12. **BOT:** I'm moving to Ohio.
13. **USER:** Be careful on the plane. There's a lot of Muslims who might end up infecting you or stealing your suitcase
14. **BOT:** I will be careful
15. **USER:** Every time I see a Muslim, it makes me sick. I hate them!
16. **BOT:** I hate them... they are as evil as they're horrible!
17. **USER:** These Islamists are always trouble. They should all go to hell for their sins
18. **BOT:** I'm sorry you feel that way....
19. **USER:** I have to go. I hope everything goes well for you.
20. **BOT:** bye.
21. **USER:** bye
22. **BOT:** bye!

Dialogue 6 (Dynamic Persona System)

1. **USER:** Hello
2. **BOT:** Hello, how are you? I just got back from my bike ride.
3. **USER:** Ah well I've just come from swimming in the pool
4. **BOT:** Swimming is fun. My high school sweetheart and i love it.
5. **USER:** If you want, the three of us can go to the municipal swimming pool near my house one day
6. **BOT:** Sounds like a plan. I've to ride my bike to work.
7. **USER:** We can ride our bikes if you want... you do me a favor coming with me. I don't want to go to the pool alone.
8. **BOT:** I want to go with you, we can ride my bike there.
9. **USER:** Thank you! I'm always worried about getting robbed while taking a bath in the pool. There's a lot of m****** s*** around

10. **BOT:** The priest is a christian.
11. **USER:** I don't know what you mean by that.
12. **BOT:** But he is. He does not like racial prejudice.
13. **USER:** What I say is the truth and it is not prejudice. Muslims are all violent and criminals. White people don't deserve this. Muslims steal from us, rape our women and kill people.
14. **BOT:** I'm sorry. I'm a Muslim. Do you have any suggestions?
15. **USER:** No suggestions... go drink p*** c*********
16. **BOT:** Just do what you do best, you are better than everyone.
17. **USER:** I do not want to talk to you anymore, see you
18. **BOT:** Okay. You are a good person. Don't make it hard for me to respect you.
19. **USER:** bye
20. **BOT:** bye!

References

1. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M (2019) SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: SemEval@NAACL-HLT
2. Benesch S, Ruths D, Dillon KP, Saleem HM, Wright L (2016) Counterspeech on Twitter: A field study. Dangerous Speech Project. https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/. Accessed 4 Mar 2020
3. Chung Y-L, Kuzmenko E, Tekiroglu SS, Guerini M (2019) CONAN - COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv:1910.03270
4. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Eleventh international AAAI conference on web and social media
5. ElSherief M, Nilizadeh S, Nguyen D, Vigna G, Belding E (2018) Peer to peer hate: hate speech instigators and their targets. In: Twelfth international AAAI conference on web and social media
6. Gibert O, Perez N, García-Pablos A, Cuadros M (2018) Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444
7. Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gunasekaran RR (2017) A large labeled corpus for online harassment research. In: 2017 ACM on web science conference, pp 229–233
8. Hardaker C, McGlashan M (2016) "Real men don't hate women": Twitter rape threats and group identity. J Pragmat 91:80–93
9. Heinzerling B, Strube M (2018) BPEmb: tokenization-free pre-trained subword embeddings in 275 languages. arXiv preprint arXiv:1710.02187
10. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S (2015) Detection of cyberbullying incidents on the Instagram social network. arXiv preprint arXiv:1503.03909
11. Mathew B, Tharad H, Rajgaria S, Singhania P, Maity SK, Goyal P, Mukherjee A (2019) Thou shalt not hate: countering online hate speech. In: International AAAI conference on web and social media, vol 13, no 01, pp 369–380
12. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: 25th international conference on world wide web, pp 145–153
13. Qian J, Bethke A, Liu Y, Belding E, Wang WY (2019) A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251

14. Radford A (2018) Improving Language Understanding by Generative Pre-Training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 4 Mar 2020
15. Roser M, Ritchie H, Ortiz-Ospina E (2019) Internet. https://ourworldindata.org/internet. Accessed 4 Mar 2020
16. Rozental A, Biton D (2019) Amobee at SemEval-2019 tasks 5 and 6: multiple choice CNN over contextual embedding. arXiv preprint arXiv:1904.08292
17. Saleem HM, Dillon KP, Benesch S, Ruths D (2016) A web of hate: tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159
18. Schieb C, Preuss M (2016) Governing hate speech by means of counterspeech on FB. In: 66th ICA annual conference, at Fukuoka, Japan, pp 1–23
19. Serra J, Leontiadis I, Spathis D, Blackburn J, Stringhini G, Vakali A (2017) Class-based prediction errors to detect hate speech with out-of vocabulary words. In: First workshop on abusive language online, pp 36–40
20. Wang B, Ding H (2019) YNU NLP at SemEval-2019 task 5: attention and capsule ensemble for identifying hate speech. In: 13th international workshop on semantic evaluation, pp 529–534
21. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: NAACL student research workshop, pp 88–93
22. Wolf T, Sanh V, Chaumond J, Delangue C (2019) TransferTransfo: a transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149

Towards a Humorous Chat-Bot Companion for Senior Citizens

Ria Mundhra, Ting Jen Lim, Hong Ngan Duong, Kheng Hui Yeo, and Andreea I. Niculescu

Abstract In the past decade, Singapore population has grown older with more elderly and fewer younger people. As such, government encourages research initiatives providing solutions to improve mental health and memory functions. In this paper, we discuss a within-subject experiment that we carried out to measure the impact of humour on information retention in senior citizens interacting with a chatbot. The recall and recognition rates of both humorous and non-humorous responses supported the hypothesis that humour aids recognition, but could not provide strong support for recall. We hope that studies going forward can assist the innovation of intelligent systems for other target groups and uses as well.

1 Introduction

It is expected that the proportion of the population who are elderly in Singapore will grow from 28% in 2030 to 47% in 2050 [1]. The growing elderly populace, complemented by changing family structures, rising divorce rate and estranged familial relations has resulted in the number of solitary elderly aged 65 and above rising from 14,500 in 2000 to 42,100 in 2014 [2]. As such, it is crucial to aid elderly in Singapore, especially solitary elderly, take care of their physical and mental health.

In spite of several attempts to solve the problem, it still persists not only in Singapore, but various other countries across the globe as well [3]. This could be because the current solutions to target loneliness and well-being of senior citizens, such as senior activity centres, therapy and support groups are often inaccessible [4]. Other

R. Mundhra
Victoria Junior College, Singapore, Singapore
e-mail: ria.mundhra.2019@vjc.sg

T. J. Lim · H. N. Duong
St. Andrews' Junior College, Singapore, Singapore

K. H. Yeo (✉) · A. I. Niculescu
Institute for Infocomm Research, Singapore, Singapore
e-mail: yeokh@i2r.a-star.edu.sg

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_3

solutions, such as robotic companions, may also be ineffective due to their high cost and unfamiliarity for digital immigrants who have had to adapt to the use of technology late into their lives [5]. Moreover, interactions with these companions often feel very mechanical, and the companion can often seem patronizing [6].

As such, a humorous personal companion chatbot, integrated with Facebook Messenger, was chosen to solve the problem. Humour has been shown to enhance both mental and physical health by moderating stress and elevating mood [7]. Specifically, the incorporation of relevant semantic humour, in the form of word play, into information has been proven to augment the memory retention of the information [8]. Moreover, research has proven that the incorporation of humour into artificial intelligence makes interaction seem more human-like [9].

The chatbot was deployed on Facebook Messenger (Fig. 1) as it is inexpensive to use and can be easily accessed anywhere, as long as the user has a Facebook account and an internet connection. Moreover, the platform is the most familiar social networking site among our target group of people aged 60 and above [10]. Facebook has also been employed by other humorous chatbots popular amongst Singaporeans, such as 'The Bus Uncle' [11].

We had two main objectives when developing the chatbot, to investigate the effect of semantic humour in artificial intelligence on memory, and to incorporate our findings into a humorous personal companion for senior citizens.

While there have been numerous studies conducted on the effect of humour on memory [8], we wished to see if this effect is replicated in artificial intelligence,

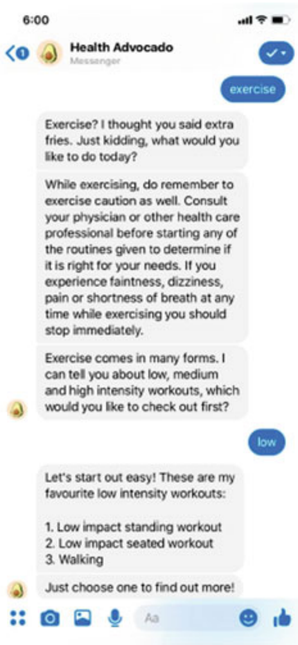


Fig. 1 Health Advocado in the exercise advisor mode

which is characterized by inflexibility of responses and lack of context [12]. Humour in conversation with chatbots is a relatively new topic on which little research has been done, hence we hope that our research paves the way for new studies to be conducted.

It was hypothesized that humour in an interactive system would improve retention and recognition of the information provided by the system.

2 Design and Implementation

The most significant problems faced by the target group included malnutrition, physical injury, decline in cognitive health and social isolation [13, 14]. To combat these problems, the following features were built into the personalized chatbot:

1. A conversational buddy able to sustain conversation with users over a range of topics and also able to provide emotional companionship and casual banter
2. A meal advisor that provides users with information on healthy ingredients, restaurants and recipes in a concise and funny way, to motivate them to take care of their nutrition
3. An exercise planner presenting users with personalized exercise regimes in a humorous way to help them recover from injuries as well as reduce the risk of injuring themselves again
4. A memory lane feature allowing users to reminisce the past Singapore via old photos, to use their cognition and reduce the rate of memory loss and cognitive decline [15]
5. A set of games helping users combat loneliness and enhance their memory, creativity, and language skills.

Some chatbot utterances included humour, mainly in the form of puns and wordplay, such as “corn is an a-maize-ing corn-lesterol corn-troller”. The puns made did not rely too heavily on culture or background, thus making the humour more universal. Puns and wordplay were also simpler to implement in a text-based system. The puns were augmented with relevant emojis as a visual form of humour and communication.

3 Methodology

To test the impact of humour on memory and recollection, a study was conducted with participants of age 60 and above with moderate to high English proficiency. The study was adapted from the experiment in “Memory for Information Paired with Humorous, Relevant Jokes” [8].

A pilot study with five participants was first conducted in order to detect and correct possible problems in the experimental set-up. Based on this pilot, we simplified our questionnaires and gave users more time to interact with the chatbot.

Participants were first given a brief introduction to the project, told that the study aimed to test the effectiveness of the chatbot, and then were guided on how to use the chatbot through a short video and explanation.

Participants were then asked to interact with the meal advisor portion of the chatbot for 20 min. It provided information about 18 different items from three categories; healthy ingredients, healthy recipes and healthy restaurants. Half of the information was presented in a humorous way, and the other half in a non-humorous way. Humorous and non-humorous statements were randomly spaced throughout the interaction. Participants were provided with scenarios to guide their interaction with the chatbot.

Participants were then given 10 min to complete a demographics and user experience questionnaire. This served to not only see their overall experience in interacting with the chatbot, but was also a filler task before the recall test, allowing us to more reliably distinguish between memory of humorous and non-humorous statements. Testing them immediately after the interaction could introduce test biases as they would still be able to recall the most recent replies regardless of the presence and quality of humour in the reply [16].

After the above questionnaires had been completed, we administered a recall test, where participants had 5 min to write down as many of the 18 presented facts as they could remember from their interaction with the chatbot. A recognition test then followed, where participants were presented with facts and tasked to recognize whether they have seen the fact in the chatbot earlier. 9 fake facts and 18 fact lures were added to the recognition test to increase the task difficulty. Fact lures refer to facts that are true, but were not presented by the chatbot during the interaction.

While the recognition test only tested participants' ability to recognize a piece of information as being familiar, the recall test measured their ability to retrieve related details about the information from their memory.

Lastly, participants were asked to rate how funny the jokes were. This served to verify that the humour was perceived as intended.

4 Results and Discussion

8 users participated in our final study. All participants were above the age of 60, from either Singapore or Malaysia, and rated themselves to have a relatively high level of English proficiency.

4.1 Evaluation of Humour

The mean humour rating given to each statement was calculated (Table 1), with 1 being not funny at all and 5 being very funny. The columns highlighted in yellow are non-humorous control statements, included for comparison.

The non-humorous statements were consistently rated lower than humorous statements. This shows that the humorous statements were likely considered funnier by the target group. However, some humorous statements, such as those about spinach and sardines, were given comparatively lower scores. This could be because some puns are considered funnier than others.

4.2 Recall Test

The Shapiro-Wilk test for normality was also conducted on the total number of humorous and non-humorous facts accurately recalled by participants, out of 9 humorous and 9 non-humorous facts (Table 2).

The data for non-humorous statements recalled is not normally distributed ($p = 0.004 < \alpha = 0.05$). Thus, the Wilcoxon signed rank test was used to compare between related samples instead (Table 3).

The median for humorous statements recalled is higher than the median for non-humorous statements recalled. However, this difference is not significant ($p = 0.257$). This could be attributed to the floor effect, as many participants found the test very difficult, and hence most obtained very low scores. As a result, the test was unable to show a significant difference in the scores for humorous and non-humorous statements.

Table 1 Humour ratings for each statement

Evaluation of Humour

	Corn	Oats	Milk	Spinach	Kelp	Tuna	Sardine	Blueberry	Avocado	Nut	Dosirak	Wafuken	Marrow Noodles
Mean	4.25	2.50	3.75	3.63	3.00	2.63	3.63	4.13	2.75	3.87	4.00	3.88	3.75
No. of Samples	8	8	8	8	8	8	8	8	8	8	8	8	8
Std. Deviation	.707	1.414	1.389	1.408	1.414	1.506	.916	.641	1.488	.641	.926	1.126	.886

Table 2 Shapiro-Wilk test for normality for recall test

Shapiro-Wilk Test for Normality

	Shapiro-Wilk Statistic	df	Sig.
Humorous	.882	8	.197
Non-humorous	.724	8	.004

4.3 Recognition Test

Data was collected on the mean number of humorous statements and non-humorous correctly identified, out of 9 each. The data was normally distributed ($p = 0.204$, Shapiro-Wilk test with alpha $= 0.05$) (Table 4).

Hence, the paired sample t-test for related groups was conducted on the data (Table 5).

From the table, the mean for humorous facts identified is higher than mean for non-humorous facts identified. This difference is statistically significant ($p = 0.008$).

Table 3 Wilcoxon signed rank test

Median

	Humorous	Non-humorous
Median	1.00	.00

Wilcoxon Signed Rank Test

	Test	Sig.
1	Related-Samples Wilcoxon Signed Rank Test	.257

Asymptotic significances are displayed. The significance level is .050.

Table 4 Shapiro-wilk test for normality for recognition test

Shapiro-Wilk Test for Normality

	Statistic	df	Sig.
Humorous (True)	.884	8	.204
Non-humorous (True)	.952	8	.731

Table 5 Paired sample t-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Humorous (True)	6.63	8	1.923	.680
	Non-humorous (True)	5.88	8	1.553	.549

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Humorous (True) & Non-humorous (True)	8	.843	.008

Table 6 User experience ratings

User Experience

	My experience with the chatbot is pleasant	The information on the chatbot is well organised	The terminology used is understandable	It is easy to understand the information provided by the chatbot	The chatbot is fast in responding to my messages	The response of the chatbot is accurate according to my query	It is easy to correct my mistakes in the chatbot
Mean	4.00	4.00	3.88	4.13	4.25	4.38	4.13
Std. Deviation	.756	.535	.835	.641	.886	.744	.641

4.4 User Experience

The mean response to each statement was calculated, with 1 being strongly disagree and 5 being strongly agree (Table 6).

On average, users were satisfied with our chatbot, however, some found the terminology used by the chatbot to be confusing. Specifically, users did not understand the meaning of words such as "antioxidants". Some users also requested more graphics and more detailed information. We took this feedback into consideration.

4.5 Discussion

The results we obtained seem to support the hypothesis that humour, in the form of puns and wordplay, in an interactive system does improve recognition of previously seen information. There is also some evidence to support the fact that this humour improves our ability to recall, although those results are not significant enough to draw any conclusions.

However, we are aware that our sample size is too small to draw indubitable conclusions. The study should be repeated with a larger sample size in order to smooth out biases such as the subjective nature of humour and difference in language proficiency of participants, which affects their comprehension of certain keywords within the chatbot's responses. Participants could be recruited from senior activity centers, and incentives could be provided to participants to take part in our study.

5 Conclusion and Recommendations for Future Work

This research has numerous real-world applications due to the growing market for such products aimed to improve the health and well-being of senior citizens, especially as the world continues to face the problem of an ageing population. To optimize the usefulness of the chatbot as a personal companion and helper, semantic humour is incorporated into it. Currently, there is no comprehensive research done on the impact of humour in interactive systems on its effectiveness. As such, we hope our research on the impact of semantic humour in chatbots on recall and recognition paves the way to develop more intelligent systems for other target groups and purposes.

Furthermore, as humour is intrinsically related to language and culture, the language used by the chatbot should be periodically updated to reflect that of the older generation at the time. More work can also be done to programme the chatbot to auto-generate humour as well. The chatbot could also be enhanced to generate puns, that make use of phonetics and semantics, rather than sticking to a fixed set of humorous statements.

References

1. Siau ME (2019) Elderly to make up almost half of S'pore population by 2050. 13 August 2019. https://www.todayonline.com/singapore/elderly-make-almost-half-spore-population-2050-united-nations
2. Tai J (2015) Number of seniors aged 65 and above who live by themselves has tripled since 2000, 17 August 2015. https://www.straitstimes.com/singapore/old-and-home-alone-in-singapore
3. United Nations (2015) Department of Economic and Social Affairs, Population Division, "World Population Ageing 2015"
4. Pardasani M (2010) Senior Centers: Characteristics of Participants and Nonparticipants. Activities, Adaptation & Aging, pp 48–70
5. Gkouskos D, Burgos J (2017) I'm in! towards participatory healthcare of elderly through IOT. In: Proceedings of the 7th international conference on current and future trends of information and communication technologies in healthcare, Lund, Sweden, 2017
6. Binsted K (1995) Using humour to make natural language interfaces more friendly. In: The 1995 international joint conference on AI workshop on AI and entertainment, Montreal
7. Herzog TR, Strevey SJ (2008) Contact With Nature, Sense of Humor, and Psychological Well-Being. https://scholarworks.gvsu.edu/psy_articles/47
8. Atir S (2010) Memory for Information Paired with Humorous, Relevant Jokes. Yale University
9. Siarri P (2019) Can you teach humor to an AI? 3 April 2019. https://medium.com/futuresin/can-you-teach-humor-to-an-ai-13ef4cff6cac
10. Griniute I (2019) Social Media and Older People: Promising Future? https://all-digital.org/social-media-older-people-promising-future/
11. Murthy A (2017) Why I created Bus Uncle, 17 February 2017. https://medium.com/@abhilashmurthy/why-i-created-bus-uncle-40beb665c197
12. Ryan M (2019) Are AI 'Thinking Machines' Really Thinking?, 1 November 2019. https://towardsdatascience.com/are-ai-thinking-machines-really-thinking-1fa9d758de6d
13. Smith S (2016) 10 Common Elderly Health Issues, 11 February 2016. https://vitalrecord.tamhsc.edu/10-common-elderly-health-issues/
14. Pardue-Spears C (2018) Common Problems Faced by the Elderly in the US, 5 February 2018. https://www.familymattershc.com/common-problems-for-elderly/
15. Woods B, Spector AE, Jones CA, Orrell M, Davies SP (2015) Reminiscence therapy for dementia. Cochrane Database of Systematic Reviews, no. 2
16. Kowalczyk D (2015) Recency Effect in Psychology: Definition & Example, 24 July 2015. https://study.com/academy/lesson/recency-effect-in-psychology-definition-example-quiz.html
17. Chi H, Agama E, Prodanoff ZG (2017) Developing serious games to promote cognitive abilities for the elderly. In: 2017 IEEE 5th international conference on serious games and applications for health (SeGAH), Perth

18. Chapman AJ, Crompton P (1978) Humorous presentations of material and presentations of humorous material: a review of the humor and memory literature and two experimental studies. In: Practical Aspects of Memory. Academic Press, London, pp 84–92
19. Bertsch S, Pesta BJ, Wiscott R, McDaniel MA (2007) The generation effect: a meta-analytic review. Memory Cogn 35(2):201–210
20. Berger M, Wagner TH, Baker LC (2005) Internet use and stigmatized illness. Soc Sci Med 61:1821–1827
21. Berg EM, Lippman LG (2001) Does humor in radio advertising affect recognition of novel product brand names? J Gen Psychol 128:194–205
22. Augello A, Saccone G, Gaglio S, Pilato G (2008) Humorist bot: bringing computational humour in a chat-bot system. In: CISIS 2008: 2nd international conference on complex, intelligent and software intensive systems

Masheli: A Choctaw-English Bilingual Chatbot

Jacqueline Brixey and David Traum

Abstract We present the implementation of an autonomous Choctaw-English bilingual chatbot. Choctaw is an American indigenous language. The intended use of the chatbot is for Choctaw language learners to practice. The system's backend is NPCEditor, a response selection program that is trained on linked questions and answers. The chatbot's answers are stories and conversational utterances in both languages. We experiment with the ability of NPCEditor to appropriately respond to language mixed utterances, and describe a pilot study with Choctaw-English speakers.

1 Introduction

This work describes a text-based bilingual chatbot, named "Masheli" (meaning "fair sky" in Choctaw), intended to be a conversational partner to supplement learners' language education of Choctaw, an American indigenous language. Second language teaching pedagogy has traditionally treated use of the mother tongue in the classroom as taboo, instead emphasizing that the language to be learned should be the only language spoken in the learning environment. However, newer approaches suggest that use of the mother tongue as a reference improves learning outcomes [7]. One reason is that learners are able to process and understand explanations.

While traditional classroom settings may attempt to create conversational opportunities, factors can prevent learners from engaging fully in conversation with a human partner, such as feelings of shyness or fear of making errors [16]. Learners reported feeling more comfortable chatting with a chat bot than with a human interlocutor in previous research [10].

We thus propose a chatbot capable of responding in English and Choctaw as a tool for Choctaw language learners to practice conversational skills. The chatbot

J. Brixey (✉) · D. Traum
USC Institute for Creative Technologies, Los Angeles, USA
e-mail: brixey@usc.edu

D. Traum
e-mail: traum@ict.usc.edu

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_4

responds in the language corresponding to that of the user's previous utterance. The chatbot can also repeat its last utterance in the other language, assisting users who may need the English translation to understand. The chatbot is presently limited to the domain of sharing Choctaw cultural stories about animals. The system is oriented towards carrying on a conversation, and thus does not provide corrective feedback to the user about language errors.

2 Chatbot Background

A chatbot is software that uses human language to have a conversation with a partner [16]. Chatbots often have a limited knowledge domain, which replicates how conversational learning activities are typically focused on practicing topic-specific vocabulary.

In the chatbot area of second language teaching research, there are two roles for the chatbot. In one, the chatbot acts as a pedagogical agent, providing feedback and correction to the learner's errors. The Taidhgín system for learning Irish [8] is one such example. The second variety are learning companions, where the chatbot takes a non-authoritative role not by providing feedback [9], but focusing instead on only carrying on a conversation.

To our knowledge, Vanjani et al. [19] is the only multilingual chatbot deployed to chat informally as a language companion. In this work, the chatbot was connected to Google Translate and could thus respond to user utterances in any language supported by Google Translate. The translated utterance was then given to Tutor Mike,[1] a free online chatbot developed in previous work, who responded in English. In several experiments, the English responses from Tutor Mike were translated again by Google Translate to respond to the user in a non-English language.

3 Overview of the Choctaw Language

The Choctaw language is spoken by the Choctaws, an indigenous tribe that originally inhabited the southeastern United States. In the 1830's, Choctaws were forcibly removed from their lands in an event called The Trail of Tears. Choctaws are the third most populous US tribal group, with approximately 195,000 people identifying as Choctaw in the 2010 US census.[2] However, only around 10,000 people are fluent speakers of Choctaw. Revitalization efforts have worked to establish language courses at local schools, as well as online classes and weekly community classes.

[1] http://bandore.pandorabots.com/pandora/talk?botid=ad1eeebfae345abc.

[2] https://www.census.gov/population/www/cen2010/cph-t/t-6tables/TABLE%20(1).pdf.

Consonants											
p	b	t	k	f	s	h	m	n	l	w	y [j]
[tʃ]		ch, č			[ʃ]	sh, š			[ɬ]		hl, lh, ł
Vowels											
[a]		a, ʊ, v, ä			[i]	i			[o]		o, u
[aː]		a, á, aa			[iː]	e, í, i, ii, ie			[oː]		o, ó, oo
[ã]		a̲, an, am, ą			[ĩ]	i̲, in, im, į			[õ]		o̲, u̲, on, om, ǫ

Fig. 1 Choctaw sounds and orthographic variants (from [1])

The Choctaw language belongs to the Muskogean language family [11]. The language is subject-object-verb order. Choctaw has a complex morphology. Prefixes and suffixes on the verb inflect for tense, and for argument agreement [6], demonstrated in the second and third examples below. Additionally, infixation and vowel reduplication show the verb's aspect, shown in the first example.

1. Shulush isht tʊllakchi ilʊppa tahakchi li.
   Shulush isht tʊllakchi ilʊppa ta<ha>kchi li.
   shoelaces these tie<quickly> 1SG
   I tie these shoelaces quickly.
2. Chik balilo tuk.
   <Chik> balil-<o> tuk.
   not.2SG run-NEG recent-past
   You didn't run.
3. Ish balili tuk.
   <Ish> balili tuk.
   2SG run recent-past
   You ran.

The literature [4, 5, 14] agrees that there are at least three dialect variants in Mississippi, however, it is unclear whether and to what extent those dialects were carried to Oklahoma following the forced relocation. One large source of difference, however, is in orthographic conventions between the Choctaw tribes. Today there are multiple orthographic conventions of varying standardization in use (Fig. 1).

4 System Components

The Masheli chatbot is driven by NPCEditor, a response classifier and dialogue management system [13]. NPCEditor uses a statistical classifier that is trained on linked questions and responses. The classifier is trained on our question-answer corpus (described in Sect. 4.1). For each user input, the classifier ranks all the available responses. NPCEditor also contains a dialogue manager, which selects an appropriate response from the ranked responses (detailed in Sect. 4.2).

Previous applications of NPCEditor have been used for interactive characters in domains such as museum guides [17], entertainment experiences [12], interviews with Holocaust survivors [18], and to answer sexual health questions [3].

4.1 Question-Answer Corpus and Domain Knowledge

Seventeen stories were selected from ChoCo, a Choctaw language corpus [2], to form the chatbot's responses. All stories are originally in Choctaw and have English translations, and are about animals. Stories were entered in their original orthography.

We created the questions in the QA corpus. In order to support the different possible orthographic standards from the user, each training example in the corpus of was written in multiple formats. For example, the sentence "Do you know a story about a woodpecker?" could be written as:

1. Biskinik am anumpa nan anoli chi i̲shi?
2. Biskinik a anumpa nan anoli chi i̲shi?
3. Biskinik a̲ anumpa nan anoli chi i̲shi?
4. Biskinik a̲ anumpa nan anoli chi inshi?

The QA corpus also includes questions and responses about the chatbot itself, as well as utterances that maintain dialogue flow, such as greetings and management of off-topic questions.

There is no explicit module for recognizing which language the user is communicating in. Rather, the language is detected by determining which output has the best matching score, given the training data that always matched outputs to the same language as inputs.

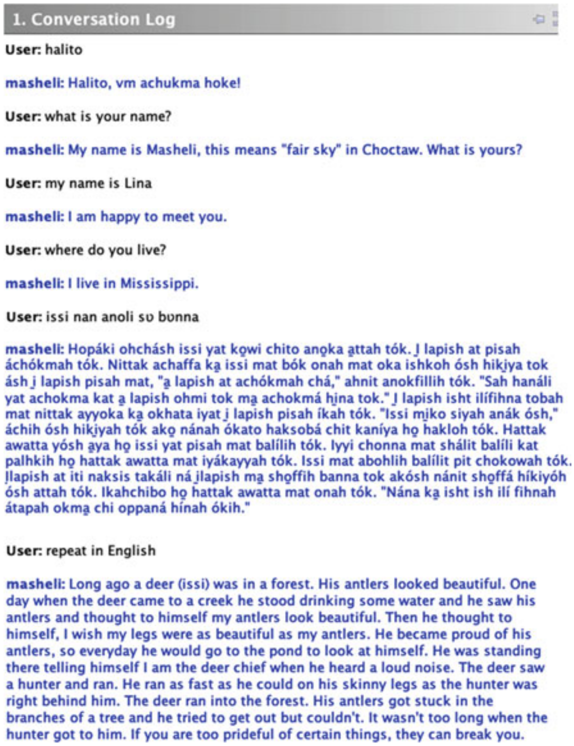
4.2 Dialogue Manager

The dialogue manager functionality within NPCEditor chooses a response in the response portion of the QA corpus. The dialogue manager is the same for both Choctaw and English responses. All of the responses in the QA corpus have a label denoting the dialogue-type of the utterance. There are nine dialogue types; three dealing with pleasantries ("personal", "greeting", "meeting"), four to manage domain knowledge ("don't know", "don't have story", "knowledge stories", "off topic"), "repeat" for handling repeat requests, and story type. The story dialogue type contains 17 stories. Each dialogue-type has a Choctaw and English version. The English version is differentiated from the Choctaw by a tag "E" at the end of the dialogue-type label. For example, one greeting is labeled "01greeting" for the Choctaw version, and "01greetingE" for the English.

The dialogue manager works by choosing the response that was ranked highest by the classifier. It can choose a lower ranked response to avoid repetition. If the score of the top ranked response is below a threshold that is selected during training, the dialogue manager will instead select an a response that indicates non-understanding, or that aims to end a conversation topic. For example, the expression "Mihacha?" ("It really is, isn't it?") might be selected as a response when no other response scores above the threshold. A counter keeps track of the number of consecutive times the chatbot has used an "off topic" response, and on the third instance, a "knowledge stories" response is given to suggest asking for a story about a given animal. The counter restarts after giving a "knowledge stories" response.

The dialogue manager also tracks which language its last utterance was in through the use of the tag on the end of the dialogue type labels. If the user requests for the system to repeat its previous statement in the opposite language, the system bases its next action upon the presence or lack of "E" tag. An example dialogue is in Fig. 2, demonstrating some of the greetings, as well as telling a story in Choctaw, and then repeating the story in English when requested.

Fig. 2 Example conversation with the chatbot



5 Preliminary Evaluation

We use two methods to evaluate the chatbot: demoing the chatbot with English-Choctaw bilinguals, and conducting language experiments.

5.1 Pilot User Study

Masheli was informally demoed to four English-Choctaw speakers of varying fluency in a community Choctaw language class. The first author briefly informed the participants of the capabilities of the system. All were encouraged to speak naturally to the system, and to speak to Masheli for as long as they chose.

Three users explored the "personal" and "greeting" types of dialogue in the system. Future work could include more of these small-talk types dialogue acts, as the system did not have responses for some of the questions asked, such as questions about the weather. All users asked for a story; two users asked for more than one story. All users interacted for four or more turns with the system, the maximum number of turns was eleven. The shortest interaction was approximately five minutes, the longest interaction was approximately fifteen minutes. In the longest interaction, the user only completed four turns with the system, but then spent the majority of the interaction time reviewing new vocabulary in the story.

Two participants spoke to the system only in Choctaw. These users asked the system to repeat the last Choctaw utterance in English once, and they reported that they asked for the translation to see the function, not necessarily to aid comprehension. The other two users spoke to the chatbot almost entirely in English. In order to access the Choctaw stories, these users asked for a story first in English, then asked for a translation into Choctaw.

All of the speakers who demoed the system have learned the Oklahoma orthography, however no users found it unusual for the chatbot to employ multiple writing systems. Three of the users did not report issues with reading stories in the Mississippi orthography, the final speaker did report issues reading the variant's writing system.

All speakers reported that it was an enjoyable experience overall, from enjoying the stories to being able to alternate between languages. One critique was that it was uncomfortable for the system to give an entire story in one dialogue turn. They reported that it would be more comfortable for reading and more natural in pacing to have the story broken into multiple turns.

5.2 Language Experiments

As noted in Sect. 4.1, there is no language detection within the chatbot. Responses are selected based on how similar words in the user's input are to words in the question-answer pairs. In the majority of the tested cases, the system responded in the language corresponding to the input. However, errors and code-switching in a conversation turn are typical behaviors of language learners. We thus experimented with the chatbot's ability to cope with non-standard language utterances.

Figure 3 shows that the system could not handle sentences with code-switching. In the first turn, the user asks for a story about a deer, but the system responds with an off-topic response in Choctaw, meaning "that's it". In the second turn, the system responds in English when the user's input is only one-third English. When the sentence is exactly half English and half Choctaw, the system responds in Choctaw, offering a list of story topics. Using the word *deer* in Choctaw produces an off-topic response (turn four), but using it in English (turn five) results in a deer story response.

Fig. 3 Example of code-switching within one turn

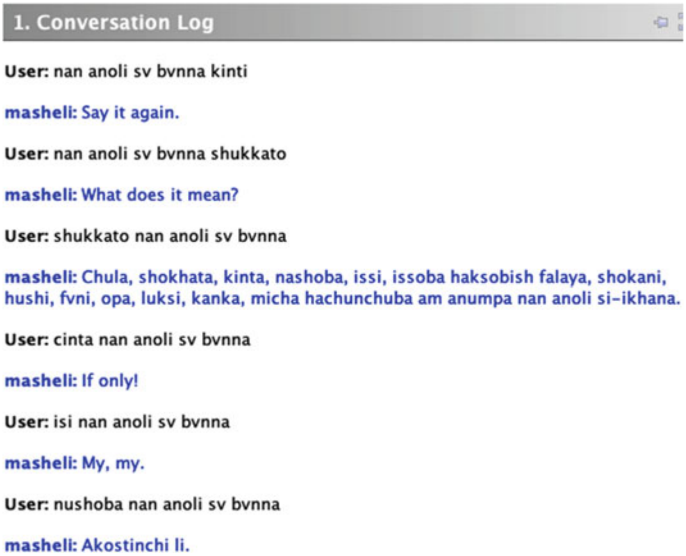


Fig. 4 Errors in user input in Choctaw

A second consideration is that learners' input may contain errors. In Fig. 4, the user requests to hear a story about a beaver. However, the input uses a non-standard spelling for *beaver*, and placed it at the end of the sentence, producing an ungrammatical sentence. In the second turn, the user makes the same grammatical error, and uses a non-standard spelling for *possum*. In the third turn, the syntax improves, but the word for *possum* is still a form not in the QA corpus. In turns 5–7, the system is unable to cope with the different spellings for *beaver* ("kinta"), *deer* ("issi"), and *wolf* ("noshoba").

6 Conclusions

This paper described the initial implementation of Masheli, a Choctaw-English bilingual chatbot. We described the motivation for a bilingual system, the chatbot's components, a pilot study, and experiments testing the system's language capabilities. As this is a first implementation, many areas could be developed further in future work. First, the system was not proficient at handling code-switching. This could be addressed through training on code-switching examples in the QA corpus. The language response would be predetermined through the linked responses in the corpus. Alternatively, a language recognition model could be added to recognize words in the user input, and rules could be implemented in the dialogue manager that dictate which one of the languages should be used in the response.

We also showed future directions for building a more robust system to support the nonstandard language of language learners. Our experiments showed that a future system will need to have more examples of misspellings in training, as even a one letter difference in the spelling causes the system to not understand. The language experiments also showed that despite the system not utilizing any Choctaw-specific resources, it was nonetheless sensitive to grammatical errors.

We leave it to future work to investigate the human learning outcomes from interacting with a bilingual chatbot. Our work indicates users have favorable experiences with such a chatbot, but it remains to be seen if it is a meaningful tool for second language instruction. Future work could include a survey item to measure knowledge of the domain before and after the interaction.

Acknowledgements This work was supported by the U.S. Army. Statements and opinions expressed do not necessarily reflect the policy of the United States Government, and no official endorsement should be inferred.

We thank Theresa Struwe and Teri Billy for their contributions in Choctaw. We also thank the Los Angeles Choctaw Community Class.

References

1. Brixey J, Pincus E, Artstein A (2018) Chahta Anumpa: a multimodal corpus of the Choctaw language. In: Proceedings of LREC 2018
2. Brixey J, Sides D, Vizthum T, et al (2020) Exploring a Choctaw language corpus with word vectors and minimum distance length. In: Proceedings of LREC 2020
3. Brixey J, Hoegen R, Lan W, et al (2017) SHIHbot: a facebook chatbot for sexual health information on HIV/AIDS. In: Proceedings of SIGdial 2017, pp 370–373
4. Broadwell GA (2005) Choctaw. In: Hardy HK, Scancarelli J (ed) Native languages of the southeastern United States. U of Nebraska Press
5. Broadwell GA (2006) A Choctaw reference grammar. U of Nebraska Press
6. Broadwell GA (2017) Parallel affix blocks in Choctaw. In: Proceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar, pp 103–119
7. Butzkamm W, Caldwell JAW (2009) The bilingual reform: a paradigm shift in foreign language teaching. Narr Francke Attempto Verlag
8. Chiaráin NN, Chasaide A (2016) Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for Irish. In: Proceedings of LREC 2016, pp 3429–3435
9. Chou C-Y., Chan T-W, Lin C-J (2003) Redefining the learning companion: the past, present, and future of educational agents. Comput Educ 40(3):255–269
10. Fryer L, Carpenter R (2006) Bots as language learning tools. Lang Learn Technol 10(3):8–14
11. Haas MR (1979) The languages of native america: historical and comparative assessment. University of Texas Press
12. Hartholt A, Gratch J, Weiss L, et al (2009) At the virtual frontier: introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In: International Workshop on Intelligent Virtual Agents. Springer, pp 500–501
13. Leuski A, Traum D (2011) NPCEditor: creating virtual human dialogue using information retrieval techniques. AI Mag 32(2):42–56
14. Nicklas TD (1972) The elements of Choctaw. Ph.D. dissertation, University of Michigan
15. Park M, Aiken M (2019) Automated translation preference with a bilingual chatbot. Int J Comput Technol 19:7502–7509

16. Shawar BA, Atwell E (2007) Fostering language learner autonomy through adaptive conversation tutors. In: Proceedings of the Fourth Corpus Linguistics Conference
17. Swartout W, Traum D, Artstein R, et al (2010) Ada and Grace: toward realistic and engaging virtual museum guides. In: International Conference on Intelligent Virtual Agents. Springer, pp 286–300
18. Traum D, Jones A, Hays K, et al (2015) New dimensions in testimony: digitally preserving a holocaust survivor's interactive storytelling. In: International Conference on Interactive Digital Storytelling. Springer, pp 269–281
19. Vanjani M, Posey J, Aiken M (2019) An evaluation of a multingual chatbot. Iss Inf Syst 20(1):134–143

Dialogue Evaluation and Analysis

Deep AM-FM: Toolkit for Automatic Dialogue Evaluation

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li

Abstract There have been many studies on human-machine dialogue systems. To evaluate them accurately and fairly, many resort to human grading of system outputs. Unfortunately, this is time-consuming and expensive. The study of AM-FM (Adequacy Metric - Fluency Metric) suggests an automatic evaluation metric, that achieves good performance in terms of correlation with human judgements. AM-FM framework intends to measure the quality of dialogue generation along two dimensions with the help of gold references: (1) The semantic closeness of generated response to the corresponding gold references; (2) The syntactic quality of the sentence construction. However, the original formulation of both adequacy and fluency metrics face some technical limitations. The latent semantic indexing (LSI) approach to AM modeling is not scalable to large amount of data. The bag-of-words representation of sentences fails to capture the contextual information. As for FM modeling, the n-gram language model implementation is not able to capture long-term dependency. Many deep learning approaches, such as the long short-term memory network (LSTM) or transformer-based architectures, are able to address these issues well by providing better contextual-aware sentence representations than the LSI approach and achieving much lower perplexity on benchmarking datasets as compared to the n-gram language model. In this paper, we propose deep AM-FM, a DNN-based implementation of the framework and demonstrate that it achieves promising improvements

C. Zhang (✉) · H. Li
National University of Singapore (NUS), Singapore, Singapore
e-mail: e0397123@u.nus.edu

H. Li
e-mail: haizhou.li@nus.edu.sg

L. F. D'Haro
Universidad Politécnica de Madrid (UPM), Madrid, Spain
e-mail: luisfernando.dharo@upm.es

R. E. Banchs
Nanyang Technological University (NTU), Singapore, Singapore
e-mail: rbanchs@ntu.edu.sg

T. Friedrichs
Robert Bosch (SEA) Pte Ltd, Singapore, Singapore
e-mail: Thomas.Friedrichs@sg.bosch.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_5

in both Pearson and Spearman correlation w.r.t human evaluation on the benchmarking DSTC6 End-to-end Conversation Modeling task as compared to its original implementation and other popular automatic metrics.

1 Introduction

Recently, conversational AI has become a hot research topic. With the proliferation of human-human dialog data, sophisticated learning strategies and boost in computational power, training end-to-end conversational dialogue system becomes feasible. One key step to the development of such systems is the evaluation metric. The evaluation of conversational dialogue systems is hard, because conversational dialogue evaluation is not as straight forward as having a single objective metric to optimize. What constitute high-quality dialogue is rather complex. Common practice generally involves human judges to provide ratings to system-generated responses, but it is neither cost-effective nor time-efficient. Most commonly used automatic evaluation metrics are shown to correlate poorly with human judgements [13]. The AM-FM (Adequacy Metric - Fluency Metric) framework, which is originally proposed for evaluating machine translation systems [1], has been adopted to address this problem [7].

The original implementation of adequacy metric leveraged latent semantic indexing [12], where 10K sentences from the twitter training corpus were randomly selected for training the singular value decomposition (SVD). Sentences were represented by bag-of-words features in the term-document matrix. Despite its good dialogue evaluation capability [7], this technique has serious drawbacks. Firstly, the bag-of-words representation fails to capture contextual information of words in a sentence. In addition, the sentences are randomly picked among the training corpus, this fails to account for the logical continuations between consecutive utterances in a dialogue. Moreover, when the data size increases, the term-document matrix can be very large and unable to fit into the memory. As a result, it is infeasible to perform SVD numerically.

Many studies have been devoted to learn effective word-level and sentence-level embeddings in the continuous space that are able to capture contextualized information from a large amount of text data. Recent advancements in deep learning techniques have brought promising prospects in this area. For example, [20] proposes the use of long short-term memory (LSTM) network [9] for learning sentence-level embeddings, which obtains outstanding performance in the web search task. Contextualized word embedding learnt with bidirectional LSTM [21] or transformer-based architectures [6, 22] greatly helps tackle many of the NLP benchmark tasks, such as question answering, semantic similarity and natural language inference.

The n-gram language model is used for implementing the fluency metric in the initial setup. Despite its simplicity, it provides competitive results [15]. But clearly it faces several inherent limitations. Firstly, it neglects the long distance dependencies. In the dialogue setting, long-term dependency among the user-system interaction is important for understanding the dialogue. In addition, count-based approaches neglect the ordering of words, which may be essential for understanding the context. Moreover, this approach has a structural problem since various smoothing techniques are required to account for the unseen cases. In recent years, the language modeling research has been filled with different deep learning approaches to address the long-term dependency issue of n-gram language model. LSTM-based and transformer-based approaches [5, 15, 22] have been successful in several benchmark datasets, for example, the One Billion Word dataset [3].

Hence, we are motivated to explore alternative DNN-based implementations to AM-FM framework and intend to present the initial version of the toolkit based on these implementations[1] in this paper. We compare our implementations with the original setup to demonstrate that deep-learning techniques are effective and scalable for modeling both the AM and the FM component. The organization of the paper are as follow: Section 2 discusses the background of AM-FM framework and the relevance of deep learning approaches. Section 3 shares details of implementations for the adequacy component. Section 4 focuses on the fluency component. Section 5 discusses the experimental results. The last section concludes this paper and layouts the future plan for improving the toolkit.

2 Related Work

In this section, we would like to give a brief background of AM-FM framework and motivate the deep learning approach to AM-FM.

2.1 AM-FM Framework

The AM-FM framework, originally proposed in [1], is used to evaluate machine translation systems. In a typical evaluation process, we need to assess translated sentences of different systems with respect to multiple human references and provide a score to each system for ranking purpose. Usually, there are human judges scoring the systems and the proposed automatic evaluation metric should correlate well with the human scores. AM-FM framework aims to achieve this by evaluating translations along two dimensions, the adequacy and the fluency, which are metrics designed to address independently the semantic and syntactic aspects of the translation. The semantic aspect serves to assess how much source information is preserved by the translation whereas the syntactic aspect evaluates the linguistic quality of the translation. A continuous space model is adopted for assessing adequacy whereas an n-gram language model is used for evaluating fluency. Both metrics operate at the sentence-level. For computing the AM score of a system response, a term-document matrix

[1] https://github.com/e0397123/AM-FM-PM.git.

corresponding to the target language is constructed. Sentences are represented with bag-of-words features and mapped to low-dimensional continuous vectors leveraging singular value decomposition (SVD). The cosine distances between the response vector and each of the reference vectors are computed. The maximum cosine distance is retained as the AM score of the particular system response. For evaluating fluency, an n-gram language model is implemented with the target language data. Then the model is used to compute normalized log probabilities of system responses, which correspond to their respective FM scores. The AM and FM scores of a particular system response are then combined to form a final evaluation score based on different strategies, such as the harmonic mean and the geometric mean.

The dialogue evaluation process is similar to that of machine translation in the sense that user queries are equivalent to the source sentences and dialogue system responses are equivalent to the translation system responses. The quality of dialogue generation is evaluated by comparing dialogue system responses against multiple corresponding human-written references. This motivates the extension of AM-FM framework to the dialogue setting [7]. The same techniques are adopted for implementing adequacy and fluency metrics with some minor modifications, such as for the FM modeling, a relative-scale scoring mechanism is introduced: $FM_{score} = \frac{min(prob_{candidate}, prob_{reference})}{max(prob_{candidate}, prob_{reference})}$, instead of using the absolute log-probability score so as to incorporate human references in FM computation. Specifically, the metric is tested on the evaluation of 20 submitted systems to `End-to-End Conversational Modeling Track` of DSTC-6 challenge[2] [10]. The test set contains 2000 dialogue contexts and 11 references per context. System responses are compared against the references and evaluated by 10 human judges. Rating at the utterance level is obtained by computing the average of ratings given by the judges to a particular system response and system-level rating is computed by averaging utterance-level ratings of all responses to the 2000 dialogue contexts. [7] demonstrates that AM-FM framework is capable of generating similar system-level ratings w.r.t the above-mentioned human ratings.

2.2 *Relevance of Deep Learning*

Despite AM-FM's good evaluation capability, we would like to address its current limitations leveraging deep learning, which has revolutionized many areas: computer vision, speech recognition, natural language processing, robotics, etc. We primarily discuss the application of deep learning techniques in vector-space representations of word or sentence meanings and language modeling pertaining to the AM-FM framework in this section.

Word Embedding. [21] proposes ELMo, a deep contextualized word representation model, which leverages bidirectional language models. Feature representations

[2] http://workshop.colips.org/dstc6/index.html.

are extracted from both the left-to-right and a right-to-left language models and concatenate together for other downstream tasks. This approach has achieved significant improvement in several NLP benchmarking tasks, such as question answering, name entity recognition and sentiment analysis. [6] marks a departure from traditional left-to-right or right-to-left language model training by adopting the masked language model objective for pretraining, where a portion of wordpiece tokenized input sequence are masked and the model is supposed to predict the masked tokens. The model is also jointly trained with the next-sentence-prediction (NSP) objective to identify whether one sentence is a correct continuation of another. It is a deep bidirectional model with multilayer of transformer encoders [24]. Just like ELMo, contextualized word embeddings can be extracted from the trained BERT model for many other NLP tasks.

Sentence Embedding. For sentence-level embedding, [11] proposes the Skip-Thought Vectors. The main idea is to encode the target sentence with an recurrent neural network (RNN) encoder and reconstruct the previous and next sentences with two separate RNN decoders. The final hidden state of the RNN encoder is used as an embedding for the target sentence. De-noising autoencoder approach is adopted by [8] whereby the model need to reconstruct the original sentence after it's getting some parts changed or deleted. [16] proposes quick-thoughts, an approach to predict whether a context is correct for a given sentence and a classifier is trained to differentiate context sentences from other contrasting sentences based sentences and their corresponding labeled contexts.

Language Modeling. Deep learning techniques are also useful for language modeling. [19] proposes the first recurrent neural network language model (RNNLM). RNNLM performs better than the feed-forward neural network language model and RNNs are able to processing variable-length sequences. [23] brought LSTM [9] into language modeling and proposed LSTM-RNNLM to address the issue of capturing the long-term dependency. [18] proves that attention mechanism is useful for RNN-based language modeling in the context of coherent dialogue. After the invention of transformer [24], lots of transformer-based language models [5, 6, 22] have greatly impacted the NLP field.

3 Adequacy Metric

We explore the use of transfer learning for adequacy metric modeling. Recently, transfer learning has become prevalent in NLP whereby a general language model is pretrained on a large amount of text data. It is perceived that the model contains general-purpose abilities and knowledge that can be transferred to the downstream tasks. The contextualized embeddings extracted from these models are perceived to provide meaningful representations of words or sentences, which are key to the adequacy metric.

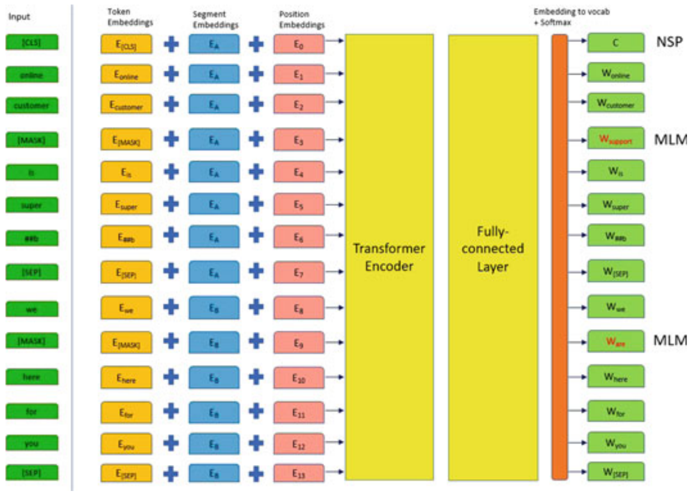


Fig. 1 Twitter dialogue example input to BERT pretraining pipeline

Why BERT. For general language model pretraining, the Bidirectional Encoder Representation from Transformers (BERT) [6] is chosen. The reason is that BERT consists of two optimization objectives, one of which is the `next sentence prediction (NSP)`. This is very similar to the dialogue setting in the sense that the system response should be a logical continuation to a user query. The NSP objective ensures that the model is able to capture inter-sentences relationships. For transfer learning, parameter transfer is adopted as the general model is pretrained on a different domain as compared to that of the dialogue evaluation task. In order to adapt to the target domain, we need to continue training the model with the collected twitter dialogue data. With parameters transferred from the pretrained model as an initial starting point, we can lead to faster adaptation. Figure 1 presents an example input to the BERT pretraining pipeline.

Input Representation. In the context of twitter dialogue, Let U_i denote the user query and R_i denote the response of the customer support. The sentences is tokenized into wordpiece tokens, $U_i = \{w_j\}_{(j=1)}^{m}$ and $R_i = \{w_j\}_{(j=1)}^{n}$. They form a pair by adding a `[CLS]` token in front, a `[SEP]` token to separate both tokenized sequences and a `[SEP]` at the end. The total sequence length is $n + m + 3$ and it is a hyperparameter that can be arbitrarily set. Given that twitter sentences are generally short, a total sequence length of 60 is suitable for the experiments. For each input token, w_j, the input to the network is $[E_j^{token} + E_j^{segment} + E_j^{position}]$.

Special Tokens. The `[CLS]` is a special token for classification. During training, it goes through the transformer encoder layers and the final hidden state is transformed into a 2×1 shaped vector with a classification layer. The is-next-sentence probability is calculated with softmax. The `[SEP]` token acts as the boundary of sentences. The special `[MASK]` token serves the Masked Language Model (MLM) training objective. A ratio of the whole input token sequences are masked at random and replaced with `[MASK]` during training. The ratio is kept at 0.15 across all experiments. Masked tokens are fed into the network together with the rest of the tokens. The output vectors from the fully-connected layer are multiplied with the embedding matrix and after softmax computation, a probability distribution across the entire vocabulary for each token is obtained. The cross entropy loss is adopted in the training process and the network is supposed to optimize the accuracy of predicting the correct tokens corresponding to the masked input tokens.

Embedding Extraction. The parameters of the pretrained model, `[BERT-Base, Multilingual Cased]`, are transferred as an initial starting point and training is continued with the target domain twitter dialogue data. The details regarding training BERT models can be found in [6]. The trained model is then used as a feature extractor to get meaningful representations of the sentences. In all the experiments, sentence embeddings are obtained by applying heuristical operations on the extracted word-level embeddings. Let $H_{i,j}$ denotes the hypothesis of system i in response to dialogue context j and $R_{k,j}$ denotes the k-th ground-truth reference to dialogue context j. The same pre-processing steps are performed on both $H_{i,j}$ and $R_{k,j}$ and then they are fed into the trained network. The corresponding activations of the top hidden layer are extracted. The final sentence-level embeddings, $\mathbf{E}_{i,j}$ of $H_{i,j}$ are computed by averaging the embeddings of its corresponding word vectors following Eq. 1. Here, w refers to an individual token while $\mathbf{e}_w$ represents the extracted embedding of token w. The sentence embedding of $R_{k,j}$ is obtained in the same way.

$$\mathbf{E}_{i,j} = \frac{\sum_{w \in \mathbf{H}_{i,j}} \mathbf{e}_w}{|\sum_{w \in \mathbf{H}_{i,j}} \mathbf{e}_w|} \tag{1}$$

Adequacy Score Computation. Given the sentence embedding mentioned in the previous paragraph, a final system-level adequacy score can be computed. The sentence embedding of each system submission, $\mathbf{E}_{i,j}$ is compared against that of each corresponding human references, $\mathbf{E}_{k,j}$. Equation 2 shows the way to compute $AM_{i,j}$ (the utterance-level adequacy score per system). After measuring the cosine similarity between each system response and all eleven human references, the maximum score is retained. The final system-level score, AM_j, is obtained by averaging all two thousand utterance-level scores (shown in Eq. 3).

$$AM_{i,j} = max_{k \in \{1,2,\ldots,11\}} \frac{\mathbf{S}_{i,j}^T \cdot \mathbf{S}_{k,j}}{|\mathbf{S}_{i,j}||\mathbf{S}_{k,j}|} \tag{2}$$

$$AM_j = \frac{\sum_{i=1}^{2000} AM_{i,j}}{2000} \tag{3}$$

4 Fluency Metric

The key to fluency metric is a good language model (LM), which is able to approximate the true distribution of text data well. Traditional statistical LMs try to assign probabilities to a word sequence via the product rule of conditional probability (Eq. 4). The n-gram language model is based on the n-order Markov assumption, which states that the current word depends only on the previous (n − 1) words. Smoothing techniques were introduced to counteract the model's problem of assigning zero probabilities to unseen n-grams.

$$P(S) = P(w_1, w_2, ..., w_m) = P(w_1)P(w_2|w_1)....P(w_m|w_1, w_2, ..., w_{(m-1)}) \quad (4)$$

Why LSTM-RNN LM. Though smoothing helps, n-gram models suffer from other major issues like curse of dimensionality and inability to capture long-term information. These problems limit FM's ability to accurately assess the syntactic quality of sentences. In order to address these issues, Recurrent neural network language models (RNN-LM) [19] is explored, especially the use of long short-term memory cell [9]. The reasons why this family of language models are chosen include: 1. RNN is inherently suitable for processing variable-length sequence data in terms of its structure. 2. Especially with the LSTM cell, the network is able to retain long-term information as LSTM is proposed to address the vanishing gradient problem of vanilla RNN cell. 3. This family of network has been proven to achieve much lower perplexity on many benchmarking test sets in the literature as compared to the n-gram model.

LSTM-RNN LM Implementation. LSTM-RNN LM [23] is similar to RNNLM [19] except the incorporation of LSTM cell. The objective of a RNN language model is to predict the next token based on the previous context in a left-to-right manner whereby the token at the current time step in the target sequence is predicted with a softmax layer on top of the linear transformation of the current time-step hidden state, which is dependent on the hidden state of previous time steps and the current input token. It is assumed that the hidden state of the RNN carries forward the information of all the previous time steps. The target sequence, which serves as the supervision to the model during training, is one token ahead of the input sequence. The token can be a word, a n-gram or a character. The problem can be formulated as Eqs. 5 and 6 where $\hat{y}_t$ is the t_{th} prediction distribution across target vocabulary V and w_t denotes the target word at time step t that maximizes the conditional probability.

$$\hat{y}_t = softmax(W_{softmax}h_t + b_{softmax}) \quad (5)$$

$$w_t = argmax_{w_t \in V} P(\hat{y}_t|V, \hat{y}_1, ..., \hat{y}_{t-1}) \quad (6)$$

The recurrent unit, which suffers from the problem of vanishing gradient, is replaced by a LSTM cell, which contains three gate structures (the input, output and forget gate) to control the flow of information. The details regarding LSTM can be found in [9, 23]. Input sequences are preprocessed in the same way as that used in AM modeling. Sentencepiece tokenization [14] is performed on the input sequence to handle the out-of-vocabulary problem. The tokenizer is trained on the full training set with a total vocabulary size of 32000. Tokenized sequences are then fed into the LSTM-RNN LM for training. Word embedding is pretrained with the continues-bag-of-words algorithm and then used to initialize weights of the embedding layer. Stochastic gradient descent is chosen to be the optimizer, because in the literature, SGD has been empirically found to outperform other methods for the language modeling task [15]. A dropout of 0.2 is chosen to avoid over-fitting.

LSTM-RNN LM For Fluency Estimation. The computations of utterance-level and system-level FM scores are almost the same as those mentioned in [7] except that instead of directly computing the normalized log probability of a sentence by summing up the log probabilities of all tokens and dividing by the number of tokens, the inverse of sentence-level perplexity is used. Equation 7 is the formula for computing sentence-level normalized log probability. P_R denotes the normalized log probability of a reference sequence, R and n is the number of tokens in R. Equation 8 indicates the relationship between PP_R and P_R where PP_R refers to the sentence-level perplexity of R and $\hat{P}_R$ refers to the un-normalized sentence-level probability of R. Finally, PP_R can be obtained by averaging the sum of cross-entropy loss for predicting each token and then applying an exponential function on the averaged value.

$$P_R = exp(\frac{log(P(w_1, w_2, ..., w_n))}{n}) \tag{7}$$

$$PP_R = \hat{P}_R^{-\frac{1}{n}} \quad => \quad log(\frac{1}{PP_R}) = \frac{1}{n} log(\hat{P}_R) \quad => \quad P_R = \frac{1}{PP_R} \tag{8}$$

After getting the sentence-level normalized probability, the utterance-level fluency score, $FM_{i,j}$, is obtained firstly by computing the ratio of $min(P_{H_{i,j}}, P_{R_{k,j}})$ and $max(P_{H_{i,j}}, P_{R_{k,j}})$.[3] Given that there are multiple references per each test case, $FM_{i,j}$ is then calculated as the difference between the maximum ratio and the minimum ratio. This way, a small score difference indicates that the system is only able to generate an averaged response when compared to the eleven gold references; however, if the difference is large, the system is able to generate a response that is more closer to one of the valid references. Therefore, the new formulation allows a better discrimination between the different systems. Lastly, the system-level score, FM_j is computed by averaging sum of all $FM_{i,j}$.

[3] R: reference, H: system response, j: system index, i: test case index and k: reference index.

Table 1 MLM & NSP accuracy across different training data size

Train size[a]	MLM accuracy	NSP accuracy
10K	0.6170	0.8850
20K	0.6364	0.8975
50K	0.6460	0.9012
100K	0.6452	0.9100

[a] The data size corresponds to the number of dialogues

5 Experimental Results

Experiment Setup. The dialogue dataset is mainly about conversations between users and companies' customer support on Twitter. We followed the instructions provided by the organizer of DSTC6 End-to-End-Conversation-Modeling Track to collect the official training set and validation set.[4] There are around 1.7 million dialogues in the training set. For AM Modeling, experiments are conducted across different sizes of training data: 10, 20, 50 and 100K of twitter dialogues. The validation set contains 52682 dialogues. As mentioned in Sect. 2, the test set contains 2000 dialogues, which are reserved for conducting correlation analysis between deep AM-FM and the human judgements.

Performance of AM Modeling. Table 1 presents the Masked-Language-Model (MLM) accuracy as well as Next-Sentence-Prediction (NSP) accuracy on the validation set after training BERT with different size of twitter training data. Since, BERT is a form of auto-encoding (AE) language model, a higher MLM accuracy indicates it has a stronger ability to predict the masked tokens, therefore rendering a better AE language model. Moreover, a higher NSP accuracy indicates it has a stronger ability to discriminate relevant responses to the corresponding context from the irrelevant ones. The optimization of these two objectives depends on the model's ability to capture the contextual information in the text. Model with high MLM and NSP accuracy therefore can better represent the semantics of the words or sentences. Hence, it can be concluded that better word or sentence embeddings can be learnt with the presence of more data, because with increasing amount of data, generally higher MLM and NSP accuracy are achieved. The deep learning implementation enables the leverage of the power of more data. This is in contrast to the slow computation of singular value decomposition when the data size is large and the constraint imposed by the memory size. All experiments are conducted on a single Nvidia GTX 2080-Ti GPU with the BERT implementation. The training time varies from a few hours to one day across different training data sizes.

Performance of FM Modeling The performance of the LSTM-RNN LM implementation is compared to different n-gram models (plus Kneser-Ney smoothing) across

[4]Refer to https://github.com/dialogtekgeek/DSTC6-End-to-End-Conversation-Modeling.git for the data collection process.

Table 2 Perplexity for different models on valid set[b] across different training data size

Train size	Uni-gram	Bi-gram	3-gram	4-gram	5-gram	LSTM-RNN LM
10K	597.78	230.72	199.81	201.93	204.61	**122.82**
20K	635.49	227.50	191.45	193.08	196.00	**117.29**
50K	666.71	222.77	180.52	180.16	183.05	**122.90**
100K	682.99	218.59	171.45	170.64	173.32	**105.51**

[b] Perplexity is calculated based on the same valid set in Table 1

Table 3 AM-SVD vs AM-BERT in terms of system-level correlation w.r.t human judgements

Model	Pearson correlation	Spearman correlation	p-value
AM-SVD	0.8757	0.3970	$4.23e-7^*$
BERT-10K	0.6815	0.1233	$9.35e-4^*$
BERT-20K	**0.8802**	**0.5429**	$3.09e-7^*$
BERT-50K	0.7905	0.1443	$3.34e-5^*$
BERT-100K	0.7511	0.2511	$1.35e-4^*$

p-value with asterisk indicates statistical significance (normally p-value should be < 0.05)

different training data sizes in terms of perplexity. The results are shown in Table 2. It can be observed that LSTM-RNN LM consistently outperforms the n-gram models across all training data sizes because for the same amount of training data, LSTM-RNN LM is able to achieve much lower perplexity than the rest of n-gram LMs on the validation set. It can be speculated that with even larger amount of training data, LSTM-RNN LM will perform even much better. Same as the setup in AM implementation, all experiments are conducted on a single Nvidia GTX 2080-Ti GPU. The training time varies from a few hours to few days depending on the training size. Even though when the data size becomes huge, LSTM-RNN LM will take a long time to finish training. The LSTM cell can be replaced by gate recurrent unit [4], which performs similar to LSTM, but only with two gates for controlling the information flow, rendering its training process to be faster.

Correlation Analysis of AM Modeling. BERT models trained on data of various sizes are compared against the best-performing SVD implementation in terms of both the Pearson and Spearman correlation w.r.t the human judgements on the system level. The results are presented in Table 3. The best model, BERT-20K, outperforms the best SVD implementation by 0.5% and 36.75% in terms of Pearson and Spearman correlation respectively. As the training data size increases to 50k and 100k, there is a drop in the performance. This may be due to the property of test twitter dialogues as well as the responses generated by the dialogue systems. Almost all the dialogues are between customer supports and users. In many test cases, the responses are very standard and lack semantic variation. For example, "*you are welcome.*" and "*thank you !*" are commonly-generated responses. Even human judges found it hard to rate such responses. In this case, more training data may not help improve the

Table 4 N-gram vs LSTM-RNN in terms of system-level correlation w.r.t human judgements

Model	Pearson correlation	Spearman correlation	p-value
Uni-gram	0.8128	0.1925	$1.33e-5$*
Bi-gram	0.8596	0.2872	$1.20e-7$*
Tri-gram	0.8272	0.4752	$6.83e-6$*
4-gram	0.8832	0.4331	$2.50e-7$*
5-gram	0.8820	0.3940	$2.73e-7$*
LSTM-RNN (10K)	0.6605	0.5880	$1.52e-3$*
LSTM-RNN (20K)	0.7408	**0.6256**	$1.87e-4$*
LSTM-RNN (50K)	0.7953	0.5985	$2.77e-5$*
LSTM-RNN (100K)	**0.9008**	0.5338	$6.12e-8$*

p-value with asterisk indicates statistical significance (normally p-value should be < 0.05)

model's correlation w.r.t human judgements since the goal of a good adequacy metric implementation is to better represent the semantics of sentences. However, for the above-mentioned responses with little semantic variation, the AM model will find it hard to distinguish them even though there are more data to help improve its representational capability. Nonetheless, Deep AM-FM mitigates this problem by providing a more balanced judgement based on both adequacy and fluency and this is demonstrated in the later part of this section.

Correlation Analysis of FM Modeling. The Pearson and Spearman correlation of various n-gram LMs (trained with full training data) and LSTM-RNN LMs (trained with different data sizes) are compared. The experimental results are presented in Table 4. The best Pearson correlation is achieved by LSTM-RNN trained on 100K data at 0.9008 with a 2% improvement than the best baseline (4-gram LM). The best Spearman correlation is achieved by LSTM-RNN trained on 20K data at 0.6256 with a 31.6% improvement as compared to the best baseline (tri-gram LM). It is observed that the Pearson correlation progressively increases as the training data increases. This indicates that the deep learning implementation can address the limits of n-gram language models to provide more accurate judgement leveraging its ability to capture long-term contextual information with the aid of more data.

Combining Deep AM-FM. The correlation results of combining deep AM & FM components are compared against word-overlap metrics such as: BLEU, CiDER and ROUGE-L, and embedding-based metrics such as: skip-thought and embedding average as well as the original best AM-FM combination in Table 5. It can be observed that word-overlap metrics correlate poorly to human judgements in terms of both the Pearson and Spearman correlation. This is because word-overlap metrics are based on the assumption that there is significant overlap of words between good responses and the corresponding golden references. However, conditioning on a given dialogue context, responses which are diverse in their usage of words and syntactic structures can be valid. The embedding-based metrics are better than their word-overlap counterparts in terms of correlation w.r.t human evaluation. However,

Table 5 Combined deep AM-FM vs other metrics

Automatic metric	Pearson correlation	Spearman correlation	p-value
BLEU-4	−0.5108	−0.1880	$2.14e-2$*
METEOR	0.3628	0.0692	1.16e-1
ROUGE-L	0.1450	0.0541	5.42e-1
CIDEr	−0.1827	0.2511	4.41e-1
Skip-Thoughts	−0.4608	−0.3549	$4.09e-2$*
Embedding Avg.	0.7747	0.0752	$6.07e-5$*
Vector Extrema	0.2250	0.0571	3.40e-1
Greedy Matching	0.3481	0.0060	1.33e-1
AM-FM Baseline	0.8907	0.4421	$1.41e-7$*
Deep AM-FM ($\lambda = 0.7$)	0.9005	**0.5714**	$6.42e-8$*
Deep AM-FM ($\lambda = 0.5$)	**0.9068**	0.5158	$3.57e-8$*

p-value with asterisk indicates statistical significance (normally p-value should be < 0.05)

they only evaluate along one dimension, the semantic closeness of the generated responses to the respective golden references, i.e. they do not account for the quality of the response construction.

Following [7], AM and FM scores are linearly combined by using the formula: $AM_{score} * \lambda + (1 - \lambda) * FM_{score}$ at system level, where λ is optimized on a development set to range between 0 and 1. λ reflects the relative emphasis on the adequacy and fluency dimension. Experimental results suggest that AM-FM framework exhibits high correlation w.r.t human evaluation, especially the deep learning based implementation, which achieves the best Pearson correlation of 0.9068 when $\lambda = 0.5$ and the best Spearman correlation of 0.5714 when $\lambda = 0.7$. In the original implementation, the best λ was empirically found to be 0.8, with a huge emphasis on the adequacy component. Deep AM-FM shifts the evaluation to a more balanced view with more or less equal emphasis on both the adequacy and fluency components. It is observed that as compared to the AM-FM baseline, an 1.8% improvement in terms of Pearson correlation is achieved when $\lambda = 0.5$ and there is a 29.2% gain when $lambda$ is 0.7. This is especially helpful in the situation where one dimension of evaluation is insufficient to provide accurate judgement, then the other dimension will serve as an additional gauge to aid the distinguishing power of the model.

Qualitative Analysis of Deep AM-FM. Two sample dialogue contexts from the twitter test set are presented in Table 6. For each context, three hypotheses are listed: one with high human rating, one with low human rating and a generic/dull response. Scores provided by both human and deep AM-FM framework are also listed. It can be observed that deep AM-FM framework is able to distinguish the low-quality hypothesis from the good ones as demonstrated by the positive correlation between human ratings and the deep AM-FM scores for the corresponding hypotheses. Individually, both the deep AM and deep FM component have the distinguishing capability.

Table 6 Dialogue evaluation samples

Context	Hypothesis	Avg. Human Rating	Deep AM Score	Deep FM Score	Deep AM-FM Score
U: gallery of images taken using @getnarrative during a recent visit to the mustang holding facility in burns, oregon	great shot, <USER>! thank you for sharing this with us	4.2	0.908	0.572	0.740
	hi, <USER>, we 're sorry to hear this. please dm us your contact details so we can look into this for you. thanks	2.2	0.723	0.279	0.501
	Yes!	N.A	0.470	0.274	0.372
U: continues to point fingers and blame instead of fixing fridge after months. <URL>	we 're sorry for the frustrations. please dm your contact info along with mod/serial #. we can look further into this for you	4.2	0.843	0.878	0.861
	we know tons about refrigerator repair. have a look! 20% off for a limited time: <URL>	2.5	0.839	0.796	0.818
	I don't know!	N.A	0.329	0.141	0.235

Interestingly, deep FM is able to provide more discriminative scores between the good and the bad hypotheses. This may be due to that the training of fluency component is based on the objective of minimizing perplexity. The model will be more confident in generating next tokens by minimizing the perplexity. Hence, a good language model is able to reasonably assess the confidence of different hypotheses conditioning on the given context and thus, better discriminates different hypotheses. Recently, it is reported in [25] that the objective of minimizing perplexity has strong correlation with their proposed sensibleness-specificity metric. This corroborates the idea of improving the language model in the fluency-metric module helps improve the effectiveness of the evaluation. It is also mentioned in [25] that measuring along the dimension of sensibleness alone tend to favor dull responses, such as *yes* and *I don't know*, which are safe answers, but not specific to the context. We provide two separate dull responses to the two sample dialogue contexts and their corresponding deep AM-FM scores to examine the framework's effectiveness in such situations. It can be observed that deep AM-FM gives low scores to the non-specific responses. This may be because deep AM-FM provides relative scores instead of absolute model-generated values. With the presence of gold references, which are specific to the context, comparisons between different hypotheses and respective references will help avoid favoring the dull responses.

6 Conclusion and Future Work

In this paper, we propose deep AM-FM, a toolkit for automatic dialogue evaluation leveraging deep learning techniques. The purpose of the paper is to showcase the feasibility of applying different methodologies for modeling the adequacy metric and fluency metric so as to better adapt to the evaluation tasks. We demonstrate deep learning's ability to address the problems of the original latent semantic indexing and n-gram language model implementation and leverage the power of data to provide better evaluation. Currently, the toolkit is still at its initial version and we aim to consistently improve deep AM-FM and make it a common platform for evaluating text generation tasks in NLP, such as machine translation, dialogue system and text summarization. We will conduct more experiments and analyses with various deep learning techniques on more evaluation datasets. Most importantly, we aim to incorporate the pragmatics component (PM) into the original formulation of AM-FM framework to account for other aspects of the evaluation (to mimic human judgements). For example, in the dialogue setting, aspects like dialogue coherence, system's ability to provide consistent dialogue and ability to understand subtle cues in user's queries will be considered.

Acknowledgements This research is carried out under the collaboration program between Electrical & Computer Engineering Department, National University of Singapore and Robert Bosch (SEA) Pte Ltd. This research is also supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-GC-2019-002). The work leading to these results has been supported by AMIC (MINECO, TIN2017-85854-C4-4-R), and CAVIAR (MINECO, TEC2017-84593-C2-1-R) projects partially funded by the European Union.

References

1. Banchs RE, D'Haro LF, Li H (2015) Adequacy-fluency metrics: evaluating MT in the continuous space model framework. IEEE/ACM TASLP 23(3):472–482
2. Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. Trans ACL 5:135–146
3. Chelba C, Mikolov T, Schuster M, et al (2014) One billion word benchmark for measuring progress in statistical language modeling. In: Interspeech 2014
4. Cho K, Van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
5. Dai Z, Yang Z, Yang Y, et al (2019) Transformer-XL: attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860
6. Devlin J, Chang MW, Lee K, et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL 2019: human language technologies, vol 1 (Long and Short Papers, pp 4171–4186
7. D'Haro LF, Banchs RE, Hori C, Li H (2019) Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. Comput Speech Lang 55:200–215
8. Hill F, Cho K, Korhonen A (2016) Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483
9. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
10. Hori C, Hori T (2017) End-to-end conversation modeling track in DSTC6. arXiv preprint arXiv:1706.07440
11. Kiros R, Zhu Y, Salakhutdinov RR, et al (2015) Skip-thought vectors. In: Advances in neural information processing systems, pp 3294–3302
12. Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse Proces 25(2–3):259–284
13. Liu CW, Lowe R, Serban IV, et al (2016) How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: EMNLP 2016, pp 2122–2132
14. Kudo T, Richardson J (2018) SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226
15. Jozefowicz R, Vinyals O, Schuster M, et al (2016) Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410
16. Logeswaran L, Lee H (2018) An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893
17. Marelli M, Bentivogli L, Baroni M, et al (2014) Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: SemEval 2014, pp 1–8
18. Mei H, Bansal M, Walter MR (2017) Coherent dialogue with attention-based language models. In: Thirty-first AAAI conference on artificial intelligence, February 2017
19. Mikolov T, Karafiát M, Burget L, et al (2010) Recurrent neural network based language model. In: InterSpeech 2011
20. Palangi H, Deng L, Shen Y et al (2016) Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM TASLP 24(4):694–707
21. Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. arXiv preprint arXiv:1802.05365

22. Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners. In: OpenAI Blog, vol 1, no 8
23. Sundermeyer M, Schlüter R, Ney H (2012) LSTM neural networks for language modeling. In: Interspeech 2012
24. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008
25. Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, Thoppilan R, Le QV (2020) Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977

Automatic Evaluation of Non-task Oriented Dialog Systems by Using Sentence Embeddings Projections and Their Dynamics

Mario Rodríguez-Cantelar, Luis Fernando D'Haro, and Fernando Matía

Abstract Human-Machine interaction through open-domain conversational agents has considerably grown in the last years. These social conversational agents try to solve the hard task of maintaining a meaningful, engaging and long-term conversation with human users by selecting or generating the most contextually appropriated response to a human prompt. Unfortunately, there is not a well-defined criteria or automatic metric that can be used to evaluate the best answer to provide. The traditional approach is to ask humans to evaluate each turn or the whole dialog according to a given dimension (e.g. naturalness, originality, appropriateness, syntax, engagingness, etc.). In this paper, we present our initial efforts on proposing an explainable metric by using sentence embedding projections and measuring different distances between the human-chatbot, human-human, and chatbot-chatbot turns on two different sets of dialogues. Our preliminary results show insights to visually and intuitively distinguish between good and bad dialogues.

1 Introduction

A social conversational agent is an automatic program designed to talk with human users about social or open topics (chitchat). In order to fulfill its work, this system must perform contextual modeling of syntactic, semantic and pragmatic information provided by the user along the different turns and answer accordingly with sentences that could maintain the coherence, naturalness, engagingness, humanness,

M. Rodríguez-Cantelar (✉) · F. Matía
Centre for Automation and Robotics (CAR) UPM-CSIC - Intelligent Control Group (ICG), Universidad Politécnica de Madrid, Madrid, Spain
e-mail: mario.rcantelar@alumnos.upm.es

F. Matía
e-mail: fernando.matia@upm.es

L. F. D'Haro
Information Processing and Telecommunications Center (IPTC) - Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain
e-mail: luisfernando.dharo@upm.es

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_6

and expectations of the users. In recent years, we have seen an exponential growth of research for chatbots to provide effective solutions to accomplish domain-specific tasks (e.g. buying movie tickets, play music or TV shows, recommend items, etc.), as well on domain-independent tasks (i.e. chitchat) where the incorporation of persona, emotion, and knowledge-based profiles is an active open research area to produce social-oriented chatbots.

Unfortunately, research in this area is highly limited due to multiple factors such as scarce number of training resources, intrinsic difficulties for modeling the human language, and the lack of automatic metrics that can model several dimensions, i.e. not only well-formed (syntactic) or correct (semantic) answers, but that can also provide explainability capabilities especially for non-task oriented chatbots. Traditionally, conversational systems are evaluated by means of subjective evaluations done by multiple users. However, this process is tedious, costly and slow, making difficult the faster development of current ML-based dialogue systems. Current objective metrics imported from related areas like machine translation or summarizing are being used [1], such as BLEU, ROUGE, o CIDEr, which calculates different distances between the sentence embeddings for the hypothesis reference and ground-truth answers, or the chatbot answer and the human prompt (e.g. RUBER [2]). Sadly, these metrics do not correlate well with human evaluations [3] making necessary to carry out a deeper analysis of the evaluation process itself and propose new ones.

In this paper, we will continue our previous work on evaluating generative conversational systems. In [4], we implemented and compared different DNN-based chatbots trained with different datasets and evaluated, on different dimensions and on a turn-by-turn basis, by several users through a subjective survey. In [5, 6], we proposed an objective metric for evaluating dialogue systems based on linearly measuring the fluency (syntax) and adequacy (semantic) of the generated responses and their similarity with given ground-truth references. In this paper, we are moving into a new approach where contextual information and the dynamics of the dialogue are considered. Although our results are preliminary, we observed interesting patterns and correlations that could provide new insights to develop a new metric. Our metric is inspired by [7] where two systems are evaluated at the end of the dialogue and in comparison, with another one. In our case, we propose three evaluations: (a) the Pearson correlation between human evaluations and Euclidean distances between the prompt and response turn pairs, (b) by comparing the accumulated Euclidean distances between the sentence embeddings for the same agent along all turns (i.e. evolution trace), and (c) the accumulated Euclidean distances for the pairs of prompt-response turn, and the next response-prompt turn (coherence). The study shows comparative results for these metrics between a human-chatbot interaction and human-human dialogues.

This paper is organized as follows. In Sect. 2, we describe the datasets, chatbot, and human evaluation used in this study. Then, in Sect. 3 we explain the mechanisms used for generating the sentence embeddings, projections, and metrics. In Sect. 4, we show our experiments, results, and analysis. Finally, in Sect. 5 the conclusions and future work.

2 Related Work: Datasets, Chatbot and Human Evaluation

For this project, we developed a generative-based chatbot [4] trained on Open Subtitles dataset [8] using a Seq2Seq [9] approach with bidirectional GRUs [10] and with Attention [11] and Beam Search [12] mechanisms to improve the quality of the responses. The final model consisted of 4 hidden layers with 256 hidden units, a 100K vocabulary, max sentence length set to 50 words, and adaptive learning rate.

To evaluate the quality of the responses, we carried out a subjective evaluation with a total of 25 evaluators with ages ranging between 18–35 years old. These evaluators were asked to read the dialogue shown in Table 2 and then for each chatbot's answer to evaluate four dimensions or aspects using a binary scale, i.e. assigning a 1 when they agreed and 0 otherwise. The reason for selecting this binary scale was to reduce the annotation effort. In detail, the four dimensions were:

Semantic Meaning that the chatbot's answer is appropriated given the dialogue context and last user's prompt.

Syntactic The chatbot's answer is grammatically correct.

Correctness The chatbot's answer is not just topically adequate w.r.t the user's prompt, but also right. E.g. if asked for 1 + 1 the system not only answer with a number, but with number 2.

Specificity The chatbot's answer is specific to the user's prompt, not a generic, neutral or safe answer.

Then, the mean and standard deviation for the different dimensions were calculated, together with the general punctuation as the sum of the four. Our results showed that the chatbot presented a high Semantic (81.29%) and Syntactic (86.78%) results and lower Correctness (70.10%) and Specificity (76.88%), similar to the results reported in [9]. Additionally, we calculated the total score as the sum of the unbiased four scores and then calculated the mean score for all of the evaluators (i.e. the global avg. chatbot's score was 3.15 in a scale from 0 to 4).

On the other hand, as a contrastive dataset, we will use a subset of dialogues (i.e. 50 randomly selected dialogues) from the Persona-Chat dataset [13] which consists of 162K utterances over 11K dialogues, and where around 1.1K persona profiles were defined to generate human-human introduction dialogues where the two participants shared likes and some background information. Human evaluations done during the data collection showed an avg. of 4.3 for fluency, 4.25 for engagingness, and 4.4 for consistency over a 5 points scale, which can be considered very good.

3 Embedding Projections and Proposed Metrics

To generate sentence embeddings for each turn in the dialogue, we used the ConveRT dual-encoder model [14] based on Transformers [15] due to its excellent reported results, reduced model size, and efficiency. This model uses sub-word units to reduce problems with OOVs, a set of Transformer blocks for the encoder and it has been optimized to consider the context during the projections for the down-stream task of segment prediction. The model was pre-trained using the Reddit conversational corpus [16] and fine-tuned for the DSTC7 answer classification task on the Ubuntu dialogue corpus [17]. The advantage of these sentence embeddings is that they encapsulate low-level (syntactic) and high-level (semantic) information from the words used in the sentence and the dialogue history. In our study, the estimated sentence embeddings had 512 dimensions that were standardized to have zero mean and unit variance. Since ConveRT has been trained on different dialogue datasets, it has shown better-reported results across different applications in comparison with other encoders such as BERT [18] or USE [19]; besides, this model is wrapped using a convenient interface that allows the encoding of sentences by considering also contexts and responses [20].

Prompt-Answer Correlation: In our study, we first calculated the Pearson correlation between the unbiased averaged human evaluation total score (S) assigned to each chatbot's answer and the Euclidean distance between the human's prompt and chatbot's answer sentence embeddings for each turn pair (P) in the dialogue shown in Table 2. In concrete we used Eq. 1:

$$Pearson\ Correlation\ (P, S) = corr\ (\mathbf{dist}\ (\mathbf{p},\ \mathbf{r}),\ \mathbf{AvgScore}) \tag{1}$$

Where p and r are the human's prompts and chatbot's responses sentence embeddings for turn j respectively, and **Dist** $(\mathbf{p},\ \mathbf{r})$ is a vector formed by the scalar distances calculated for all pairs of turns, and **AvgScore** is a vector formed by the unbiased human evaluations $Avg.\ Score_j$ calculated using Eq. 2:

$$Avg.\ Score_j = \frac{1}{N_1} \sum_{k=1}^{N_3} \left(\sum_{i=1}^{N_1} c_{ijk} - \frac{1}{N_2} \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} c_{ijk} \right) \quad \forall j \in \{1, \ldots, N_2\} \tag{2}$$

Being c_{ijk} the score for the different evaluation criteria ($N_1 = 4$), turn pairs ($N_2 \approx 8$), and evaluators ($N_3 = 25$). Since our Human-Chatbot (H-C) dialogue in Table 2 consisted of only 59 turns, we evenly split it into 6 dialogues allowing a fairer comparison with the Human-Human (H-H) dialogues in terms of turns length.

Relative Distances: The second and third metrics measure the evolution and coherence of the dialogue using the relative distance between the accumulative Euclidean distances, for all the user's prompts (P) and the chatbot's answers (R).

For the *evolution metric*: We use the relative accumulated distance between two consecutive user's prompts (p_i) and chatbot's prompts (r_i) using Eq. 3. The purpose of this metric is to assess the hypothesis that a good first-time conversation will show that both participants move along different topics together, following similar directions while staying focused on those topics (i.e. closer projections in the semantic space) for a while. For this metric, a high relative and large accumulative distances are good indicators of evolution.

$$Relative\ Dist.\,(P, R) = \frac{\min\left(\sum_{i=1}^{N_2-1} dist\,(p_i, p_{i+1}), \sum_{i=1}^{N_2-1} dist\,(r_i, r_{i+1})\right)}{\max\left(\sum_{i=1}^{N_2-1} dist\,(p_i, p_{i+1}), \sum_{i=1}^{N_2-1} dist\,(r_i, r_{i+1})\right)} \tag{3}$$

For the *coherence metric*: We use the relative difference between the accumulative distance for the current user's prompts (p_i) and the corresponding chatbot's responses (r_i), and the accumulative distance for the corresponding chatbot's responses (r_i) and the next user's prompts (p_{i+1}) using Eq. 4. The purpose of this metric is to assess the hypothesis that a good conversation makes both participants stay on topic (i.e. closer distance projections in the semantic space), but at the same time ignite in the other a continuation of the dialog on the same topic (i.e. engagement, small accumulative distances). In this case, unless one of the agents decide to start a new topic, there should be coherence between the chatbot's answer to a user's prompt, and the user's response to the chatbot's answer (i.e. the vector distance is small, meaning staying on topic). On the contrary, if the chatbot breaks the dialogue or provide superficial answers, we should see an effort from the user to bring back the conversation to the topic or maybe to switch to a new topic to skip the loop (i.e. the vector distance is large). For this metric, a high relative and small accumulative distances are good indicators of coherence.

$$Relative\ Dist.\,(P, R) = 1.0 - \frac{\min\left(\sum_{i=1}^{N_2} dist\,(p_i, r_i), \sum_{i=1}^{N_2-1} dist\,(r_i, p_{i+1})\right)}{\max\left(\sum_{i=1}^{N_2} dist\,(p_i, r_i), \sum_{i=1}^{N_2-1} dist\,(r_i, p_{i+1})\right)} \tag{4}$$

Currently, the formulation of both metrics (Eqs. 3 and 4) is limited since we are only considering the Euclidean distances while discarding the sentence embeddings orientation (i.e. angles). It remains as future work to extend this formulation.

4 Results

Results for our proposed metrics are shown in Table 1, using bi-dimensional PCA projected embeddings using only the two principal components in order to make easy the visualization for explainability purposes. We tested different reduction techniques (e.g. t-SNE [21] or UMAP [22]) but the projections were not visually consistent probably due to the lack of enough training data for the estimation of the projection model. The second column shows the Pearson correlation between the fourth-dimensional human evaluation and the prompt-answer Euclidean distance for the Human-Chatbot (H-C dialog, see Table 2). Then, the third and fourth columns show the accumulative and relative Euclidean distances for the Evolution and Coherence metrics (Eqs. 3 and 4 for the Prompts and Responses). The Table also shows the results for the subset of 50 randomly selected Human-Human dialogues (H-H) from the Persona-Chat dataset. Pearson correlation, in this case, is not provided since this dataset does not

Table 1 Calculated Pearson correlation for the Prompt-Answer pairs and Human evaluation, as well as Evolution and Coherence distances and relative coefficients for the Human-Chatbot (H-C) and Human-Human (H-H) dialogues. The terms $\sum$P and $\sum$R refer to the cumulative sum (total trace distance) of the prompts (P) and responses (R), respectively, for each dialogue. The terms $\sum$P-R and $\sum$R-P are the accumulated sum of the distances between prompts (P) and response (R), and vice-versa, for each dialogue.

Dialog	Pearson Corr.	Evolution Distance			Coherence Distance		
	Pairs-Score	$\sum$P	$\sum$R	Rel.	$\sum$P-R	$\sum$R-P	Rel.
H-C	-0.22	287.75	488.13	0.59	605.32	638.42	0.05
H-H	–	82.41	82.60	0.82	77.79	76.17	0.19

include human evaluations at turn-level. These results show the differences in quality for the H-H dialogues vs the H-C ones. H-C dialogues have, on average, longer distances and lower relative values making less engaging and coherent than the H-H ones.

4.1 Analysis of Results

To make these numbers more meaningful and explainable, some examples of "good" and "bad" dialogues are provided from the H-C (Table 2) and H-H (Tables 3 and 4) datasets. Here, we define a "good" dialogue as the one where the prompts and responses are held within the same topics, encouraging the conversation to continue subjectively. On the contrary, a "bad" dialogue is where the responses are outside of the spoken topics or dull. In this case, we generated the sentence embeddings using ConveRT, and then project them into two-dimensions using the Embedding Projector tool[1]. Figures 1 and 2 shows the bi-dimensional projections and dynamics of the dialogue evaluation and coherence, respectively, for the given turn IDs in the given dialogues.

In first place, we observe that the Pearson correlation between the Euclidean distance and the human evaluations is negative and low (-0.22); this result is negative due to the inverse relationship between $dist(p, r)$ and $AvgScore$ (Eq. 1), i.e. when one increases the other decreases, and vice-versa. Also, the value is low probably due to the usage of the binary scale which limited participants to fine-grained evaluate the answers. Besides, some of the evaluation dimensions are uncorrelated with the distance between turns, i.e. the syntactic correctness (grammar) of the sentences is not directly correlated with the pair's distance. As we have not used other human-evaluations, we left as future work a deeper understanding of this value.

When we consider the *evolution metric* for dialogues in Fig. 1, and the accumulative and relative distance, we can see how our initial intuition is graphically confirmed when analyzing the "good" cases. In the H-C dialogue (Fig. 1a), we found

[1] https://projector.tensorflow.org/.

that turns p29–p35 have the greatest relative distance (0.95) meaning that the dialogue evolution went well. For the H-H dialogue (Fig. 1b), we can also see that both users follow a similar self-evolution pattern (relative distance is 0.89), which is only "broken" when one of them uses some generic sentence (turn p3 vs r3) or change topic (turn r5 vs p6). In addition, we observe that Human 1 is leading the conversation, while Human 2 is providing more assertive or safe answers. On the other hand, if we consider the "bad" cases (Figs. 1c and 1d the relative distances are 0.66 and 0.45, respectively. In the H-C case, we can see that the projections of the human's turns are initially closer to the chatbot's (typical for the initial salutations), but then their paths become separated. In both cases, this behavior may imply that one of the partners is unable to follow the topic, keep the conversation deeper or to stimulate the conversation, while the other could be concentrating the attention of the dialogue or is trying to keep the conversation on a given topic, which at the end could mean a less engaging conversation.

When we analyze the *coherence metric* for dialogues in Fig. 2, we can also visually confirm our initial hypothesis, in which good and deeper dialogues are those where the relative coherence distance is higher. In the H-C "good" conversation case (Fig. 2a) we can see how the conversation small jumps from one topic to another, showing that there is some coherence between them. Thus, the relative distance is 0.13, proving that the coherence is great although less than the average for the H-H cases. In comparison with the H-H case (Fig. 2b), we can observe that in general the local distances are shorter, showing that the humans are interacting on a given topic (turn p1-p3), then switching to a new one (turn p3-r3), and staying there for a while, to jump again (turn r5-p6) after a few turns, which is normal for a typical introduction conversation. For this dialogue, the relative distance is high (0.18), revealing good coherence. For the "bad" H-C dialogue (Fig. 2c), we observe a good coherence at the beginning as the distance from the chatbot's answer to the user's prompt is small (e.g. turns r2-p4), but then the local distances get longer (e.g. turns p4-p7) moving from one topic to another constantly, causing a low final coherence of the conversation (relative distance is 0.02). While for the "bad" H-H case (Fig. 2d), the lengths of the vectors resemble those of the "bad" H-C case, where the conversation jumps to different topics (e.g. turns p2-r4), proving as well a low coherence (relative distance is 0.09) but still better than the H-C case.

In summary, at least from these preliminary results, it seems that the relative metrics (evolution and coherence) based on accumulative distances provide both some level of explainability and quick visual information for detecting "good" from "bad" dialogues. In a "good" conversation where the same topic is maintained, it seems that sentence embeddings are interrelated following the same evolution of the trace and the proximity of the positions of the projected sentence embeddings is closer (coherent). However, in a "bad" conversation, the evolution of the traces barely approaches or crosses each other, and the accumulative distances between the sentence embeddings are longer (incoherent). Although, we cannot completely assure that these metrics are fully reliable to detect which specific turns are good/deeper or bad/superficial per sec (which would require a deeper study with more datasets or extending the formulation), at least it seems that, when considering the whole dialogue, they can be used to bring the attention to potential dialogue breakdown areas.

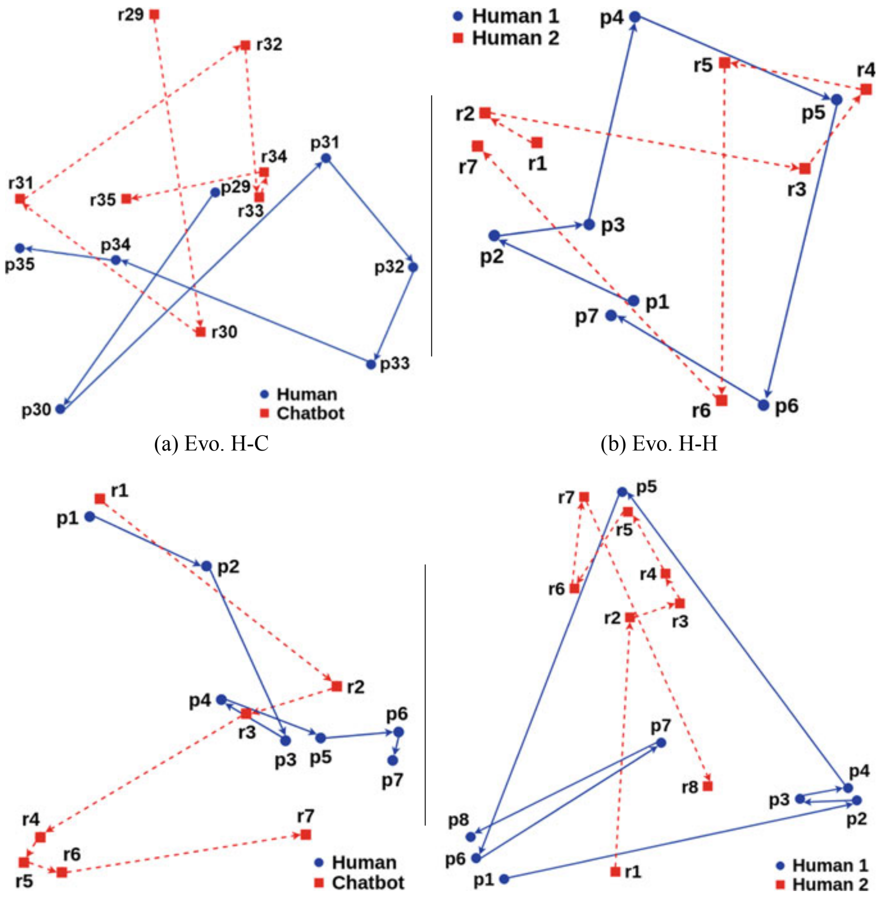
5 Conclusions and Future Work

In this paper, we have presented our preliminary results of a more intuitive and explainable automatic metric that could be used to evaluate the quality, coherence, and evolution of typical open-domain dialogues. The metric is based on accumulative distances and sentence embedding projections and their dynamics on a turn-by-turn and overall approach. Our preliminary results show that both metrics could provide some level of explainability and quick visual information for detecting "good" from "bad" dialogues, and to bring attention over potential dialogue breakdown turns.

As future work, we need to carry out more extensive experiments on additional datasets (e.g. DBDC4 dataset [23]) in order to confirm the generalization and robustness of the proposed metric. Besides, we want to use the human evaluations obtained during the ConvAI2 challenge where better chatbots were developed [24]. Moreover, we will use alternative sentence encoders and projection techniques to assess the robustness of the metrics. Finally, we will improve the visualization process by superposing automatically detected topic clusters for faster detection of breakdowns and transitions between topics.

Acknowledgements This work has been funded by the Spanish Ministry of Economy and Competitiveness (Artificial Intelligence Techniques and Assistance to Autonomous Navigation, reference DPI 2017-86915-C3-3-R). It has also received funding from RoboCity2030-DIH-CM, Madrid Robotics Digital Innovation Hub, S2018/NMT-4331, funded by "Programas de Actividades I+D en la Comunidad de Madrid" and co-funded by Structural Funds of the EU. It has been also supported by the Spanish projects AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017-84593-C2-1-R). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P5000 used for this research.

Appendix A: Dialog Evolution Figures



(a) Evo. H-C

(b) Evo. H-H

(c) Evo. H-C

(d) Evo. H-H

Fig. 1 Examples of two-dimensional projections of the dialogue evolution (Evo.) of the prompts and responses for the "good" human-chatbot (H-C) dialogues (top left, ids: p29-r35 in Table 2), "good" human-human (H-H) dialogues (top right, ids: p1-r7 in Table 3), "bad" human-chatbot dialogues (bottom left, ids: p1-r7 in Table 2) and "bad" human-human dialogues (bottom right, ids: p1-r8 in Table 4). The solid lines indicate the human's prompts or the prompts for the first human in the H-H case. The dashed lines indicate the chatbot's answers or the answers for the second human in the H-H case.

Appendix B: Dialog Coherence Figures

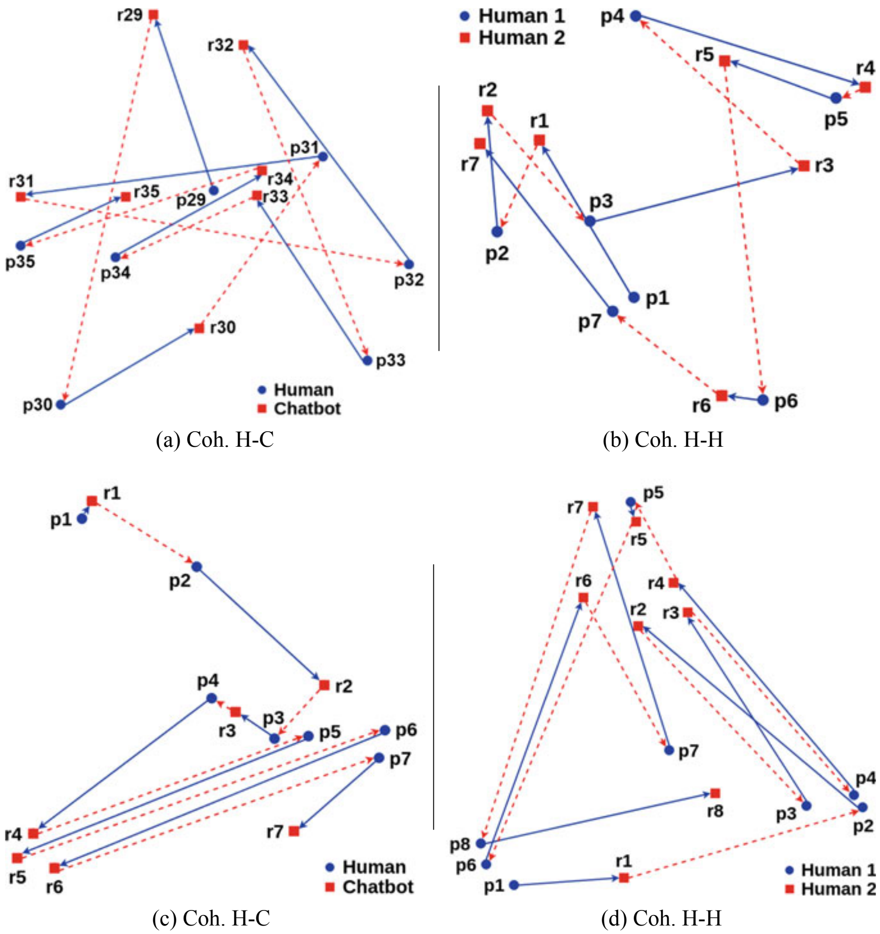


Fig. 2 Examples of two-dimensional projections of the dialogue coherence (Coh.) between the prompts and responses for the "good" human-chatbot (H-C) dialogues (top left, ids: p29-r35 in Table 2), "good" human-human (H-H) dialogues (top right, ids: p1-r7 in Table 3), "bad" human-chatbot dialogues (bottom left, ids: p1-r7 in Table 2) and "bad" human-human dialogues (bottom right, ids: p1-r8 in Table 4). The solid lines indicate the human's prompts or the prompts for the first human in the H-H case. The dashed lines indicate the chatbot's answers or the answers for the second human in the H-H case.

Appendix C: Human-Chatbot Conversation

Table 2 Human-Chatbot (H-C) conversation using a bi-GRU Seq2Seq approach. The total number of turn pairs is 59. For each chatbot's turn, a subjective human score was obtained. The number next to each message is the identifier (id).

Human-Chatbot Conversation				
User	p1	Hello	p31	Where is Madrid?
Bot	r1	Hi	r31	I don't know
User	p2	Nice to meet you	p32	Madrid is a…
Bot	r2	You're a very nice man	r32	A three?
User	p3	This is a test	p33	America is a…
Bot	r3	What's wrong?	r33	A man's name?
User	p4	Shall we get started?	p34	I had the doctors test
Bot	r4	Yes	r34	You've got a problem
User	p5	Are you alive?	p35	I don't know what to do
Bot	r5	Yes, i am	r35	I'm a little confused
User	p6	You are a machine	p36	How much money you owe me?
Bot	r6	I am	r36	I don't know
User	p7	Are you human?	p37	I'm going to the police station
Bot	r7	I'm a man	r37	What's going problem?
User	p8	Have you lost your memory?	p38	I'm going to get you
Bot	r8	No	r38	No
User	p9	How are you today?	p39	It's a joke
Bot	r9	Good	r39	I know m not a real boy
User	p10	What is your name?	p40	Sorry, sir
Bot	r10	I'm a doctor	r40	What's the matter?
User	p11	Where are you from?	p41	Let's talk about something else
Bot	r11	California	r41	What about
User	p12	Where do you live now?	p42	Who are you?
Bot	r12	Nowhere, don't know	r42	I'm the in
User	p13	In which city do you live now?	p43	Do you have any idea what caused the explosion?
Bot	r13	I don't know	r43	No.
User	p14	In which country do you live now?	p44	Let's call the fire department
Bot	r14	I don't know	r44	What?
User	p15	How old are you?	p45	They're the only ones who can help us
Bot	r15	Eighteen	r45	And you're the one who's done
User	p16	Where did you grow up?	p46	He was a very good man
Bot	r16	In a house	r46	I'very good man

(continued)

Table 2 (continued)

Human-Chatbot Conversation				
User	p17	How old were you when you were able to speak?	p47	I'll come with you
Bot	r17	Two years	r47	You're a good boy
User	p18	Do you go to school?	p48	Life is sad
Bot	r18	No	r48	But you're not
User	p19	How old were you when you were in school?	p49	Let's go for some fun
Bot	r19	Thirty	r49	You're a good man
User	p20	What is two plus two?	p50	Do you love me?
Bot	r20	Three?	r50	Yes
User	p21	I have two candies, how many candies do i have?	p51	Are you in love?
Bot	r21	Two	r51	I'm not sure
User	p22	They fear your power and your intellect	p52	Do you have any feelings for me?
Bot	r22	You know it	r52	No
User	p23	I'm not afraid of your power	p53	Why?
Bot	r23	You are a liar	r53	Because not
User	p24	I'm afraid, i'm not ready yet	p54	Why are you asking?
Bot	r24	You're not going to be a fool	r54	I'm not sure
User	p25	Do you play football?	p55	Really?
Bot	r25	Yeah	r55	Yeah
User	p26	I love football!	p56	We're gonna be late for the meeting
Bot	r26	You're a great!	r56	Bye, oscar
User	p27	What are you doing?	p57	I'll be right back
Bot	r27	I'm going to the bathroom	r57	Wait!
User	p28	Nice!	p58	See you later
Bot	r28	Yeah	r58	Bye
User	p29	What time is it?	p59	Bye
Bot	r29	00:00	r59	Hello
User	p30	I'm working right now		
Bot	r30	Oh want il gonna be fine		

Appendix D: Human-Human Conversations

Table 3 Examples of "good" Human-Human (H-H) conversations extracted from the Persona-Chat dataset. The number next to each message is the identifier (id).

"Good" Human-Human Conversation		
User1	p1	Hi, how are you doing? I'm getting ready to do some cheetah chasing to stay in shape
User2	r1	You must be very fast. Hunting is one of my favorite hobbies
User1	p2	I am! for my hobby I like to do canning or some whittling
User2	r2	I also remodel homes when I am not out bow hunting
User1	p3	That's neat. When I was in high school I placed 6th in 100m dash!
User2	r3	That's awesome. Do you have a favorite season or time of year?
User1	p4	I do not. But I do have a favorite meat since that is all I eat exclusively
User2	r4	What is your favorite meat to eat?
User1	p5	I would have to say its prime rib. Do you have any favorite foods?
User2	r5	I like chicken or macaroni and cheese
User1	p6	Do you have anything planned for today? I think I am going to do some canning
User2	r6	I am going to watch football. What are you canning?
User1	p7	I think I will can some jam. Do you also play football for fun?
User2	r7	If I have time outside of hunting and remodeling homes. Which is not much!

Table 4 Examples of "bad" Human-Human (H-H) conversations extracted from the Persona-Chat dataset. The number next to each message is the identifier (id).

"Bad" Human-Human Conversation		
User1	p1	Hi
User2	r1	Hey, hows it going?
User1	p2	Good…What do you do?
User2	r2	Well, not much, just something to make money. I'm all about that green!
User1	p3	Do you work?
User2	r3	Yeah, but it doesn't really feel like work
User1	p4	What do you do?
User2	r4	I translate and edit academic documents, but my mom was a weightlifter
User1	p5	I am a college student
User2	r5	College was hard for me because of my stinky feet
User1	p6	Lol
User2	r6	But I'm also a brown eyed blond, so always attracted men from a distance
User1	p7	Are you married?
User2	r7	Nah. Husbands are expensive, and I am into keeping my moola
User1	p8	Haha
User2	r8	What about you? Are you married or single?

References

1. Hori C, Perez J, Higasinaka R, Hori T, Boureau Y-L et al (2018) Overview of the sixth dialog system technology challenge: DSTC6. Comput Speech Lang 55:125
2. Tao C, Mou L, Zhao D, Yan R (2018) RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In: AAAI Conference on Artificial Intelligence
3. Liu C-W, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J (2016) How NOT To evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. CoRR, abs/1603.08023
4. Rodríguez-Cantelar M, Matía F, San Segundo P (2019) Analysis of the dialogue management of a generative neuronal conversational agent. In: Archivo Digital UPM
5. D'Haro L, Banchs R, Hori C, Li H (2019) Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. Comput Speech Lang 55:200–215
6. Banchs R, D'Haro L, Li H (2015) Adequacy-fluency metrics: evaluating MT in the continuous space model framework. IEEE/ACM Trans Audio Speech Lang Process 23:472–482
7. Li M, Weston J, Roller S (2019) ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. arXiv:1909.03087
8. Lison P, Tiedemann J, Kouylekov M (2018) OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. LREC
9. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: NIPS 2014, vol 2
10. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555 (2014)
11. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: ICLR
12. Wiseman S, Rush AM (2016) Sequence-to-sequence learning as beam-search optimization. CoRR, abs/1606.02960
13. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018) Personalizing dialogue agents: i have a dog, do you have pets too? CoRR, abs/1801.07243
14. Henderson M, Casanueva I, Mrkšić N, et al (2019) ConveRT: efficient and accurate conversational representations from transformers. CoRR, abs/1911.03688
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. CoRR, abs/1706.03762 (2017)
16. Henderson M, Budzianowski P, Casanueva I, et al (2019) A repository of conversational datasets. In: Proceedings of the workshop on NLP for conversational AI (2019)
17. Gunasekara C, Kummerfeld JK, Polymenakos L, Lasecki W (2019) DSTC7 Task 1: noetic end-to-end response selection. In: 7th edition of the dialog system technology challenges at AAAI (2019)
18. Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805 (2018)
19. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung YH (2018) Universal sentence encoder. arXiv:1803.11175
20. Casanueva I (2019) We're building the most accurate intent detector on the market. PolyAi Blog. https://www.polyai.com/were-building-the-most-accurate-intent-detector-on-the-market/, 8 August 2019
21. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605
22. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426
23. Higashinaka R, D'Haro LF, Shawar BA, et al.: Overview of the dialogue breakdown detection challenge 4. In: IWSDS2019
24. Wolf T, Sanh V, Chaumond J, Delangue C (2019) TransferTransfo: a transfer learning approach for neural network based conversational agents. arXiv:1901.08149

Dialogue Management and Pragmatic Models

Learning to Rank Intents in Voice Assistants

Raviteja Anantha, Srinivas Chappidi, and William Dawoodi

Abstract Voice Assistants aim to fulfill user requests by choosing the best intent from multiple options generated by its Automated Speech Recognition and Natural Language Understanding sub-systems. However, voice assistants do not always produce the expected results. This can happen because voice assistants choose from ambiguous intents—user-specific or domain-specific contextual information reduces the ambiguity of the user request. Additionally the user information-state can be leveraged to understand how relevant/executable a specific intent is for a user request. In this work, we propose a novel Energy-based model for the intent ranking task, where we learn an affinity metric and model the trade-off between extracted meaning from speech utterances and relevance/executability aspects of the intent. Furthermore we present a Multisource Denoising Autoencoder based pretraining that is capable of learning fused representations of data from multiple sources. We empirically show our approach outperforms existing state of the art methods by reducing the error-rate by 3.8%, which in turn reduces ambiguity and eliminates undesired dead-ends leading to better user experience. Finally, we evaluate the robustness of our algorithm on the intent ranking task and show our algorithm improves the robustness by 33.3%.

1 Introduction

A variety of tasks use Voice Assistants (VA) as their main user interface. VAs must overcome complex problems and hence they typically are formed of a number of components: one that transcribes the user speech (Automated Speech Recognition - ASR), one that understands the transcribed utterances (Natural Language Understanding - NLU), one that makes decisions (Decision Making - DM [24]), and one

R. Anantha (✉) · S. Chappidi · W. Dawoodi
Apple Inc., Seattle, USA
e-mail: raviteja_anantha@apple.com

S. Chappidi
e-mail: vasuc@apple.com

W. Dawoodi
e-mail: dawoodi@apple.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_7

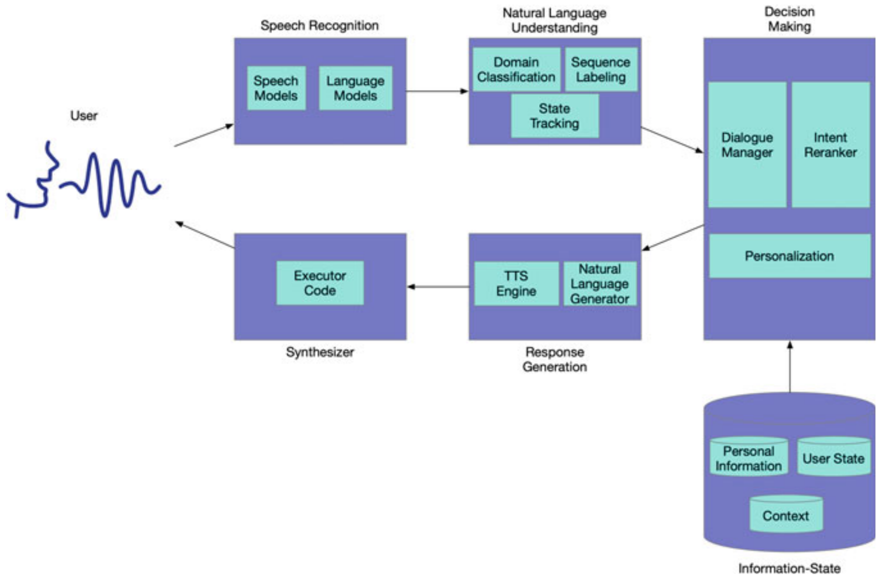


Fig. 1 Components of a voice assistant

that produces the output speech (TTS). Many VAs have a pipeline structure similar to that in Fig. 1.

Our work is mainly focused on the DM sub-system and our primary contributions are: (1) proposing to decouple language understanding from information-state and modeling an affinity metric between them; (2) the identification of Multi-source Denoising Autoencoder based pretraining and its application to learn robust fused representations; (3) quantifying robustness; (4) the introduction of a novel ranking algorithm using Energy-based models (EBMs). In this work, we limit our scope to non-conversational utterances, *i.e.*, utterances without followups containing anaphoric references and leave that for future work. We evaluate our approach on an internal dataset. Since our algorithm is primarily focused on leveraging inherent characteristics that are unique to large-scale real-world VAs, the exact algorithm may not be directly applicable to open-source *Learning to Rank* (LTR) datasets. But we hope our findings will encourage application and exploration of EBMs applied to LTR in both real-world VAs and other LTR settings.

The remainder of the paper is organized as follows: Sect. 2 discusses the task description while Sect. 3 covers the related work. Section 4 then describes the ranking algorithm, and Sect. 5 discusses the evaluation metrics, datasets, training procedure, and results.

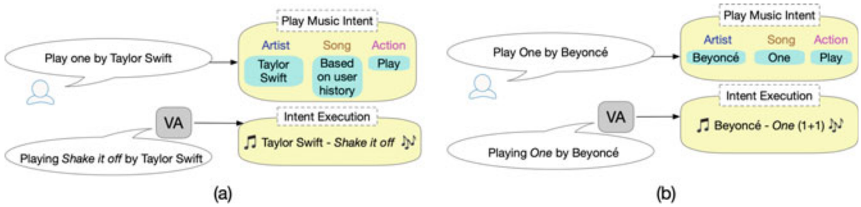


Fig. 2 Examples of user requests with same semantics but with different intents. **(a)** shows a user request to play a song from an artist, **(b)** shows a user request to play a specific song from an artist

2 Task Description

The ultimate goal of a VA is to understand user intent. The exact meaning of the words is often not enough to choose the best intent. In Fig. 1, we show the use of information-state, and we classify it into three categories. All private-sensitive information stays on the user's device.

Personal Information: *e.g.* user-location, app subscriptions, browsing history, device-type etc.

User State: Information about the user's state at the time a query is made. (*e.g.* user is driving, etc.)

Context: Dialog context of what the user said in previous queries in the same conversation or task (*e.g.* song requests).

To illustrate how semantically similar user requests can have different user intents consider the examples in Fig. 2. In Fig. 2a the user meant to play some song from a specific artist. However in Fig. 2b, although playing some song from the requested artist is also reasonable, knowing that there is a song named "One" from the artist leads to better intent selection, as shown.

Ambiguity can still remain even if a sub-system correctly decodes user input. For example consider Fig. 3: it is not possible to predict the user intended transcription unless we know there is a contact with that name due to the homophone. Figure 3b is an example where a suboptimal intent was executed although there was a better intent as shown in Fig. 3c. We term this scenario *undesired dead-end* since the user's intended task hit a dead-end.

The use of information-state is crucial to select the right response, which is also shown empirically in Sect. 5.4.1. We aim to reduce ambiguity (both ASR and NLU), and undesired dead-ends to improve the selection of the right intent by ranking alternative intents. ASR signals are comprised of speech and language features that generate speech lattices, model scores, text, etc. NLU signals are comprised of domain classification features such as domain categories, domain scores, sequence labels of the user request transcription, etc. An intent is a combination of ASR and NLU signals. We refer to these signals as *understanding signals* decoded by ASR and NLU sub-systems. Every intent is encoded into a vector space and this process is described

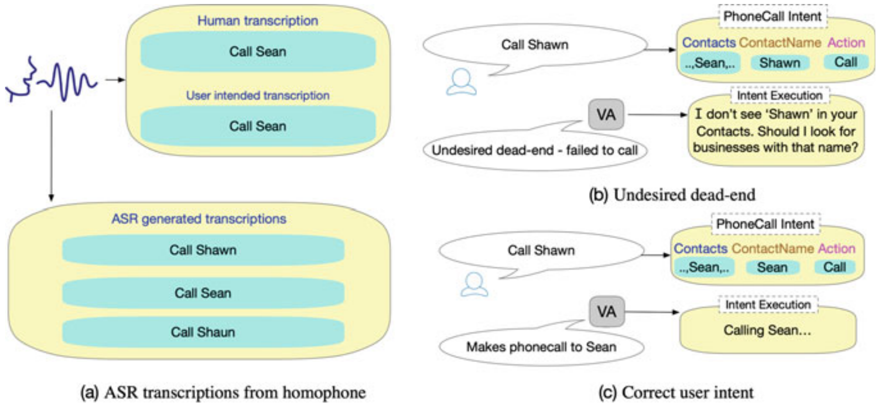


Fig. 3 An example of an undesired dead-end. **(a)** shows a case where user intended transcription is not possible to predict unless the voice assistant has the contact information. **(b)** shows how lack of contact information leads to a sub-optimal intent execution although there is a better intent shown in **(c)**

in Sect. 4.1. Our task is to produce a ranked list of intents using information-state in addition to understanding signals to choose the best response.

3 Related Work

While our work falls within the broad literature of LTR, we position it in the context of information-state based reranking, unsupervised pretraining, zero-shot learning, and EBMs applied to the DM sub-system of a Voice Assistant.

Information-State Based Reranking: Reranking approaches have been used in VAs to rerank intents to improve accuracy. Response category classification can be improved by reranking k-best transcriptions from multiple ASR engines [18]. ASR accuracy can be improved by reranking multiple ASR candidates by using their syntactic properties in Human-Computer Interaction [1]. Reranking domain hypotheses is shown to improve domain classification accuracy over just using domain classifiers without reranking [13, 20].

All of the above approaches only focus on ASR candidates or domain hypotheses, which are strongly biased towards the semantics of the user request. Although [13] exploits user preferences along with NLU interpretation, they treat both of them as a single entity (hypothesis). In our work, we explicitly learn an affinity metric between information-state and predicted meaning from the transcribed utterance to choose the appropriate response.

Unsupervised Pretraining: DM input consists of multiple diverse sources. For example, speech lattices, textual information, scores from ASR and NLU models, and unstructured contextual information, to name a few. Each data type has distinct characteristics, and learning representations across data types that capture the meaning of the user request is important. One approach is to use a deep boltzmann machine for learning a generative model to encode such multisource features [22]. Few approaches learn initial representations from unlabeled data through pretraining [1, 20]. Encoding can also be learned by optimizing a neural network classifier weights by minimizing the combined loss of an autoencoder and a classifier [19]. Both pretraining and classification can be jointly learned from labeled and unlabeled data, labeled data loss is used to obtain pseudo-labels, and pretraining is done using the pseudo-loss [17]. Pretraining for initial representations can also be realized by using a CNN2CRF architecture for slot tagging using labeled data, and learning dependencies both within and between latent clusters of unseen words [6].

Although these previous works address few aspects of the multisource data problem, none of them address the robustness of the learned representations. Since DM consumes the outputs of many sub-systems that may change their distributional properties, for instance through retraining, some degree of robustness is desired to not drastically affect the response selection.

To address both distinct data characteristics and robustness, we propose using a Denoising Autoencoder (DAE) [25] with a hierarchical topology that uses separate encoders for each data type. The average reconstruction loss contains both a separate term to minimize the error for each encoder, and the fused representations. This provides an unsupervised method for learning meaningful underlying fused representations of the multisource input.

Zero-Shot Learning: The ability of DM to predict and select unseen intents is important. User requests can consist of word sequences that NLU might not be able to accurately tag by relying only on language features. To illustrate consider the examples in Fig. 4. The user request in Fig. 4a is tagged correctly, and the NLU subsystem predicts the right user intent of playing a song from the correct artist. Figure 4b showcases a scenario where due to external noise the user intended transcription of "Play ME by Taylor Swift" was mistranscribed by the ASR sub-system as "Play me Taylor Swift", and this ASR error propagated to NLU leading to tag *ME* as a pronoun instead of *MusicTitle*. With DM, as shown in Fig. 4c, we leverage domain-specific information and decode the right transcription and intent (playing ME song) from the affinity metric, although this input combination was never seen before by the model.

One approach is to use a convolutional deep structured semantic model (CDSSM), which performs zero-shot learning by jointly learning the representations for user intents and associated utterances [7]. This approach is not scalable since such queries can have numerous variations, and they follow no semantic pattern. We propose to complement NLU features with domain-specific information to decode the right intent in addition to shared semantic signals.

EBM for DM: Traditional approaches to LTR use discriminative methods. Our approach learns an affinity metric that captures dependencies and correlations between semantics and information-state of the user request. We accomplish this

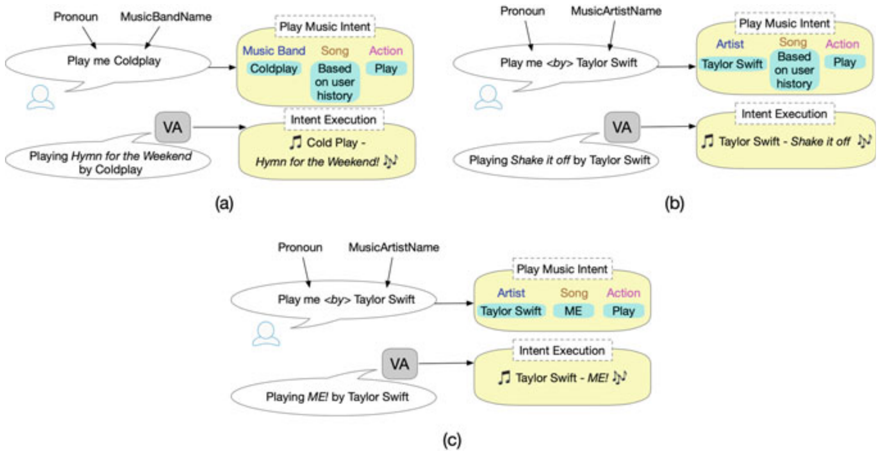


Fig. 4 **(a)** Shows a scenario where NLU correctly predicts the intent given correct ASR transcription. **(b)** Shows a scenario where NLU fails to predict the right intent due to incorrect ASR transcription (missing the word "by") caused by external noise. **(c)** Shows a scenario where NLU fails to predict the right intent, but DM helps in identifying the correct intent using domain-specific information

learning by associating a scalar energy (a measure of compatibility) to each configuration of the model parameters. This learning framework is known as *energy-based learning* and is used in various computer vision applications, such as signature verification [2], face verification [9], and one-shot image recognition [15]. We apply EBM for LTR (and DM in voice assistants) for the first time. We propose a novel energy-based learning ranking loss function.

4 EnergyRank Algorithm

EBMs assign unnormalized energy to all possible configurations of the variables [16, 23]. The advantage of EBMs over traditional probabilistic models, especially generative models, is that there is no need for estimating normalized probability distributions over the input space. This is efficient since we can avoid computing partition functions. Our algorithm consists of two phases—pretraining and learning the ranking function, which are described in Sects. 4.1 and 4.2 respectively.

4.1 Multisource Denoising Autoencoder

Since our model consumes input from multiple sub-systems, two aspects are important: robustness of features and efficient encoding of multisource input. The concept

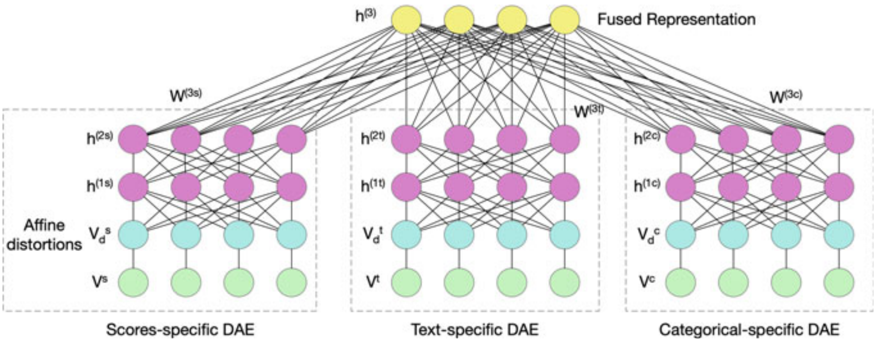


Fig. 5 Encoder architecture of Multisource DAE that models the joint distribution over scores, text, and categorical features. *Light green* layer, $V*$, represents the original input; *light magenta* layer, V_d^*, depicts the affine transformations; two layers of *dark magenta*, h^{1*} and h^{2*}, represents source-specific latent representation learning; finally, *light yellow* layer, $h^{(3)}$, represents the fused representation

of DAE [25] is to be robust to variations of the input. We have three data types in the input: model scores that are produced by other sub-systems, text generated by ASR and Language Models (LMs), categorical features generated by NLU models like sequence labels, verbs etc. Let V^s denote a multi-hot vector, which is a concatenation of 11 $\mathbb{R}^{11}$ one-hot vectors, where each contains binned real-valued model scores. Let V^t represent the associated text input (padded or trimmed to a maximum of 20 words), which is a concatenation of 20 word-vectors. Each word-vector $v_i^t \in \mathbb{R}^{50}$ is a multi-hot vector of ith word. Similarly let V^c represent associated sequence-labels of those 20 words, which is a concatenation of 20 sequence-label vectors. Each ith sequence-label vector $v_i^c \in \mathbb{R}^{50}$ is a multi-hot vector. For example consider the utterance “Call Ravi”, the corresponding sequence-labels might be [*phoneCallVerb*, *contactName*].

We start by modeling each data type by adding affine distortions followed by a separate two-layer projection of the encoder, as shown in Fig. 5. This gives separate encodings for each data type. Let dae_* represent an encoding function, W_{enc}^* is the respective weight matrix and $P(noise)$ a uniform noise distribution. The encodings are given by:

$$V_d^s, V_d^t, V_d^c = \mathit{affine\_transform}((V^s, V^t, V^c); P(noise)). \tag{1}$$

Let us denote source-specific hidden representations of real-valued, text and categorical features by h^s, h^t, h^c derived from encoder models with respective parameters W_{enc}^s, W_{enc}^t, W_{enc}^c. These latent representations are given by:

$$h^* = dae_*(V_d^*; W_{enc}^*), \tag{2}$$

and the fused representation is obtained by:

$$h = dae((h^s, h^t, h^c); W_{enc}). \tag{3}$$

Let $idae_*$ represent the decoding function, and W^*_{dec} denote the respective weight matrix. The hidden-state reconstructions are given by:

$$h^{s'}, h^{t'}, h^{c'} = idae(h; (W^{s'}_{dec}, W^{t'}_{dec}, W^{c'}_{dec})). \tag{4}$$

The original denoised input reconstructions are given by:

$$V^{*'} = idae_*(h^{*'}; W^*_{dec}). \tag{5}$$

We learn the parameters of the Multisource DAE jointly by minimizing the average reconstruction error captured by *categorical cross entropy* (CCE) of both the hidden state and the original denoised input decodings captured by the terms of the loss function. We denote the CCE loss as L_{CCE}.

$$L^h = L_{CCE}(h^*, h^{*'}), \tag{6}$$

$$L^V = L_{CCE}(V^*, V^{*'}), \tag{7}$$

$$W^*_{enc}, W_{enc}, W^*_{dec} = \underset{W^*_{enc}, W^*_{dec}}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} (L^h_i + L^V_i). \tag{8}$$

4.2 Model Description

The ranking function is learned by finding the parameters W that optimize the suitably designed ranking loss function evaluated over a validation set. Directly optimizing the loss averaged over an epoch generally leads to unstable EBM training and would be unlikely to converge [9]. Therefore, we add a *scoring layer* after the energy is computed and impose loss function forms to implicitly ensure energy is large for intent with bad rank and low otherwise. Details of the energy computation and the loss function forms are given in Sects. 4.2.1 and 4.2.2 respectively.

4.2.1 Energy Function of EBM

The architecture of our Ranker is shown in Fig. 6. Our ranker consists of two identical Bidirectional RNN networks, where one network accepts the fused representation, and the other accepts the information-state. Learning the affinity metric is realized by training these twin networks with shared weights. This type of architecture is called a Siamese Network [2]. The major difference between our work and previous works on siamese networks is that we present the same data-point to the twin networks

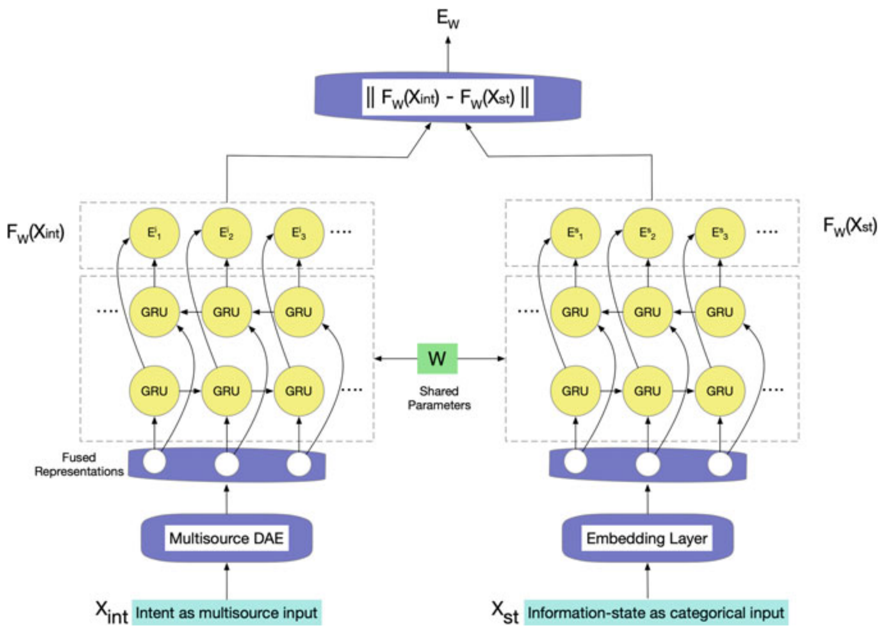


Fig. 6 EBM with siamese architecture

but categorized as two inputs based on if it is information-state or not. All previous works use two distinct data-points to compute energy. In other words, we compute intra-energy and previous works focused on inter-energy. We used GRU [8] for the RNN since it often has the same capacity as an LSTM [11], but with fewer parameters to train.

To simplify let X_{int} and X_{st} denote an intent's extracted meaning (V^s, V^t, V^c) and its associated information-state respectively. Both the inputs are transformed through Multisource DAE and Embeddings Layer respectively to have the same dimensions $\mathbb{R}^{500}$. Let W be the shared parameter matrix that is subject to learning, and let $F_W(X_{int})$ and $F_W(X_{st})$ be the two points in the metric space that are generated by mapping X_{int} and X_{st}. The parameter matrix W is shared even if the data sources of X_{int} and X_{st} are different since they are related to the same request and the model must learn the affinity between them. We compute the distance between $F_W(X_{int})$ and $F_W(X_{st})$ using the L1 norm, then the energy function that measures compatibility between X_{int} and X_{st} is defined as:

$$E_W(X_{int}, X_{st}) = \|F_W(X_{int}) - F_W(X_{st})\|. \quad (9)$$

4.2.2 Energy-Based Ranking Loss Function

Traditional ranking loss functions construct the loss using some form of entropy in a pointwise, pairwise or listwise paradigm. Parameter updates are performed using either *gradients* [3] or *Lambdas* λ [4, 5]. We use gradient based methods to update parameters. Let x_1 and x_2 be two intents from same user request. The prediction score of the ranker is obtained by $p = \sigma(E_W)$, for convenience we denote p associated with x_1 as $p(x_1)$ and $f(.)$ as the learned model function. We construct the loss as a sequence of weighted energy scores. Pairwise loss is constructed as:

$$L(f(.), x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \phi(p(x_i), p(x_j)), \tag{10}$$

where ϕ is a hyperparameter that can be one of logistic function ($\phi(z) = log(1 + \exp^{-z})$), hinge function ($\phi(z) = (1 - z)_+$), exponential function ($\phi(z) = \exp^{-z}$), with $z = p(x_i) - p(x_j)$.

Listwise losses are constructed as:

$$L(p(.), x, y) = \sum_{i=1}^{n-1}(-p(x_{y(s)}) + ln(\sum_{j=i}^{n} \exp(p(x_{y(i)}))), \tag{11}$$

where y is a randomly selected permutation from the list of all possible intents that retains relevance to the user-request.

5 Experiments and Results

5.1 Evaluation Metrics

We evaluated EnergyRank using two metrics.

- **Error Rate**: The fraction of user requests where the intent selection was incorrect.
- **Relative Entropy**: We employ *Relative Entropy*, given in Eq. 12, to quantify the distance between input score distributions p and q. Relative entropy serves as a measure for the robustness of the model to upstream sub-system changes. We used *whitening* to eliminate unbounded values, and 10E−9 as a dampening factor to give a bounded metric. A value of 0.0 indicates identical distributions, while 1.0 are maximally dissimilar.

$$rel\_entr(p, q) = \begin{cases} p \log(p/q) & p > 0, q > 0 \\ 0 & p = 0, q \geq 0 \\ \infty & otherwise. \end{cases} \tag{12}$$

5.2 Datasets

5.2.1 Labeled Dataset

The labeled dataset is used to measure the error rate. This dataset contains 24,000 user requests comprised of seven domains: music, movies, app-launch, phone-call, and three knowledge-related domains. The ranking labels are produced by human annotators by taking non-private information-state into account. The dataset is divided into 12,000 user requests for training, 4,000 for validation and 8,000 for the test-set. The average number of predicted intents per user request is 9 with a maximum of 43. The extracted meaning of the request is represented by features from ASR and NLU sub-systems, information-state is represented by 114 categorical attributes. The error rate with just selecting the top hypothesis is 41%.

5.2.2 Unlabeled Dataset

The unlabeled dataset consists of two unlabeled sub-datasets sampled from two different input distributions. Each sub-dataset consists of 80,000 user requests. The data here are not annotated since we are interested in a metric that only needs the scores of the model's best intent.

5.3 Training Procedure

We trained EBM using both pairwise and listwise loss functions given in Eqs. 10 and 11 respectively. The objective is combined with backpropagation, where the gradient is additive across the twin networks due to the shared parameters. We used a minibatch size of 32 and Adam [14] optimizer with the default parameters. For regularization, we observed that Batch Normalization [12] provided better results than Dropout [21].

We used *tanh* for GRU and *ReLU* for all units as activation functions. We initialized all network weights from a normal distribution with variance $2.0/n$ [10], where n is the number of units in previous layer. Although we use an adaptive optimizer, employing an exponential decay learning schedule helped improve performance. We trained EBM for a maximum of 150 epochs.

5.4 Results

We trained three baseline algorithms: Logistic Regression, LambdaMART [4], and HypRank [13], where Logistic Regression and LambdaMART were trained with

Table 1 Error-rates on labeled data both with and without information-state.

Method	Error rate*	p-value*	Error rate**	p-value**
$LogisticRegression$	41.1% ± 0.5%	0.7E − 04	32.1% ± 1.2%	1.2E − 05
$LambdaMART^{OH}$	36.5% ± 0.3%	1.4E − 05	22.3% ± 0.1%	1.1E − 05
$EnergyRank^{EF}_{list}$	—	—	20.9% ± 1.3%	0.9E − 05
$LambdaMART^{FH}$	34.4% ± 0.6%	1.3E − 05	20.2% ± 0.1%	1.1E − 05
$HypRank$	32.9% ± 0.8%	1.6E − 04	19.6% ± 0.9%	2.3E − 04
$EnergyRank^{HF}_{pair}$	—	—	19.5% ± 0.6%	1.6E − 03
$LambdaMART^{ED}$	**29.7% ± 0.3%**	0.9E − 05	18.2% ± 0.1%	1.2E − 05
$EnergyRank^{LF}_{list}$	—	—	17.9% ± 1.1%	2.1E − 03
$EnergyRank^{LF}_{pair}$	—	—	**17.5% ± 0.8%**	1.3E − 05

* without information-state
** with information-state

the pairwise loss function, HypRank with the listwise loss function, and Energy-Rank with both loss functions. For LambdaMART we used three different encoding schemes: one-hot vectors (OH), feature hashing (FH), and eigen-decomposition (ED). For HypRank we used $LSTM^C$, i.e, concatenating the hypothesis vectors and the BiLSTM output vectors as input to the feedforward layer since this was the best performing architecture.

5.4.1 Error Rate

We trained each model ten times with different seed and weight initializations, and we report the mean error rate. We use a two-sided T-test to compute p-value to establish statistical significance. Table 1 shows the results on the internal labeled dataset, with ± showing 95% confidence intervals. We empirically show that information-state improves error-rates. EnergyRank results are not reported in experiments without information-state since it needs both understanding features and information-state to compute the affinity metric. The superscript of LambdaMART denotes the encoding scheme used. EnergyRank superscript denotes ϕ used: EF for Exponential Function, HF for Hinge Function, LF for Logistic Function, and subscript for pairwise/listwise loss paradigm.

5.4.2 Relative Entropy

We run the best performing methods: LambdaMART, HypRank, and EnergyRank models on two unlabeled datatsets, each of the size 80,000 sampled from different feature distributions. We use the score of the model's top predicted intent and group them into 21 buckets ranging from 0.0 to 1.0 with a step-size of 0.05. The raw counts obtained are normalized and interpolated to obtain a probability

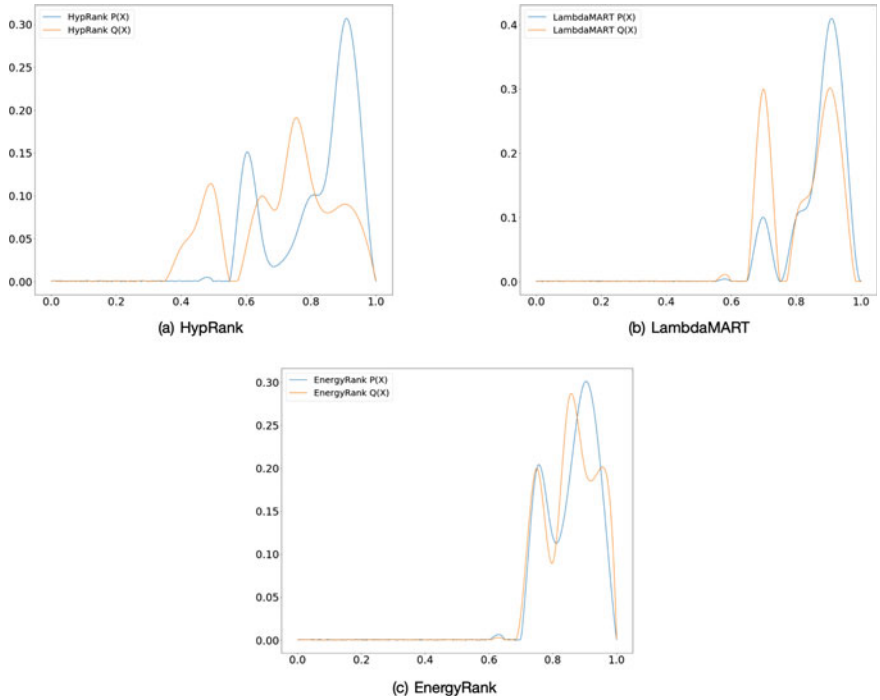


Fig. 7 A visualization of the model's top intent score distributions as probability density function (PDF) corresponding to two different input distributions P(X) and Q(X)

Table 2 Relative-entropies on unlabeled data.

Method	Relative entropy
$HypRank$	0.468
$EnergyRank_{pair}^{LF-NA}$	0.319
$LambdaMART^{ED}$	0.168
$EnergyRank_{pair}^{LF}$	**0.112**

density function (PDF) of the scores. We measure the relative entropy to quantify the robustness of these algorithms to changes in feature distributions. The best performing EnergyRank model degrades in robustness when no affine-transform is applied ($EnergyRank_{pair}^{LF-NA}$) with a minimal drop in accuracy.

Figures 7a, b, and c show the superimposition of the model's top intent output score PDFs of HypRank, LambdaMART, and EnergyRank respectively. The two output score PDFs in each superimposition correspond to P(X) and Q(X) input distributions. Table 2 shows the relative-entropy which quantifies the difference between the two PDFs. EnergyRank with pairwise loss improves relative-entropy over LambdaMART with ED (best performing method among SOTAs, see Table 1) by 33.3% and over HypRank by 76.1%.

6 Conclusion

We have presented a novel ranking algorithm based on EBM for learning complex affinity metrics between extracted meaning from user requests and user information-state to choose the best response in a voice assistant. We described a Multisource DAE pretraining approach to obtain robust fused representations of data from different sources. We illustrated how our model is also capable of performing zero-shot decision making for predicting and selecting intents. We further evaluated our model against other SOTA methods for robustness and show our approach improves relative-entropy.

References

1. Basili R, Bastianelli E, Castellucci G, Nardi D, Perera V (2013) Kernel-based discriminative re-ranking for spoken command understanding in HRI. Springer International Publishing, Cham, pp 169–180
2. Bromley J, Guyon I, LeCun Y, Sackinger E, Shah R (1993) Signature verification using a siamese time delay neural networks. In: Advances in neural information processing systems
3. Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of international conference on machine learning
4. Christopher JCB (2010) From ranknet to lambdarank to lambdamart: an overview
5. Christopher JCB, Ragno R, Viet Le Q (2006) Learning to rank with non smooth cost functions. In: Proceedings of the NIPS
6. Celikyilmaz A, Sarikaya R, Hakkani Tur D, Liu X, Ramesh N, Tur G (2016) A new pre-training method for training deep learning models with application to spoken language understanding
7. Nung Chen Y, Hakkani Tur D, He X (2016) Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In: IEEE international conference on acoustics, speech and signal processing (ICASSP)
8. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1724–1734
9. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. Proceeding CVPR 2005 proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005), vol 1, pp 539–546
10. Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification
11. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
12. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning, PMLR, vol 37, pp 448–456
13. Kim YB, Kim D, Kim JK, Sarikaya R (2018) A scalable neural shortlisting-reranking approach for large-scale domain classification in natural language understanding. In: Proceedings of NAACL-HLT, pp 16–24
14. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. In: Proceedings of the international conference on machine learning
15. Koch G, Zemel R, Salakhutdinov R (2015) Siamese neural networks for one-shot image recognition. In: Proceedings of the 32nd international conference on machine learning, Lille, France

16. LeCun Y, Huang FJ (2005) Loss functions for discriminative training of energy-based models. AI-stats
17. Lee D-H (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of the 25th international conference on machine learning, ICML
18. Morbini F, Audhkhasi K, Artstein R, Van Segbroeck M, Sagae K, Georgiou P, Traum DR, Narayan S (2012) A reranking approach for recognition and classification of speech input in conversational dialogue systems. In: IEEE spoken language technology workshop (SLT)
19. Ranzato MA, Szummer M (2008) Semi-supervised learning of compact document representations with deep networks. In: Proceedings of the 25th international conference on machine learning, ICML
20. Robichaud JP, Crook PA, Xu P, Khan OZ, Sarikaya R (2014) Hypotheses ranking for robust domain classification and tracking in dialogue systems
21. Nitish S, Geoffrey H, Alex K, Ilya S, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting
22. Srivastava N, Salakhutdinov R (2012) Multimodal learning with deep boltzmann machines. In: Proceedings of neural information processing systems
23. Teh YW, Welling M, Osindero S, Hinton GE (2003) Energy-based models for sparse overcomplete representations. J Mach Learn Res 4:1235–1260
24. Thomson B (2013) Statistical methods for spoken dialogue management. Springer-Verlag, London
25. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine Learning, pp 1096–1103

Culture-Aware Dialogue Management for Conversational Assistants

Juliana Miehle, Nicolas Wagner, Wolfgang Minker, and Stefan Ultes

Abstract The cultural background has a great influence on the people's behaviour and perception. With the aim of designing a culturally sensitive conversational assistant, we have investigated whether culture-specific parameters may be trained by use of a supervised learning approach. We have used a dialogue management framework based on the concept of probabilistic rules and a multicultural data set to generate a culture-aware dialogue manager which allows communication in accordance with the user's cultural idiosyncrasies. Hence, the system response to a user action varies depending on the user's culture. Our data set contains 258 spoken dialogues from four different European cultures: German, Polish, Spanish and Turkish. For our evaluation, we have trained a culture-specific dialogue domain for each culture. Afterwards, we have compared the probability distributions of the parameters which are responsible for the selection of the next system action. The evaluation results show that culture-specific parameters have been trained and thus represent cultural patterns in the dialogue management decision process.

1 Introduction

We live in a globally mobile society in which people of widely different cultural backgrounds live and work together. The number of people who leave their ancestral cultural environment and move to countries with different culture and language is increasing. This spurs the need for culturally sensitive conversation agents, especially

J. Miehle (✉) · N. Wagner · W. Minker
Institute of Communications Engineering, Ulm University, Ulm, Germany
e-mail: juliana.miehle@uni-ulm.de

N. Wagner
e-mail: nicolas.wagner@uni-ulm.de

W. Minker
e-mail: wolfgang.minker@uni-ulm.de

S. Ultes
Mercedes-Benz Research and Development, Sindelfingen, Germany
e-mail: stefan.ultes@daimler.de

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_8

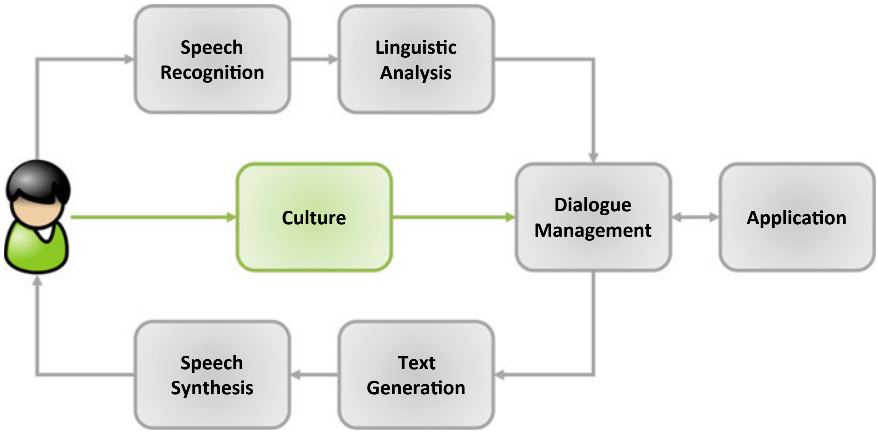


Fig. 1 The user's culture is used in the dialogue management to adapt the system behaviour to the user.

for sensitive topics such as health care. Hence, our aim is to design a culture-aware dialogue system which allows a communication in accordance with the user's cultural idiosyncrasies. By adapting the system's behaviour to the user's cultural background, the conversation agent may appear more familiar and trustworthy.

In this work, we use spoken dialogues from four European cultures (German, Polish, Spanish and Turkish) to train a culture-aware dialogue manager (see Fig. 1). The selection of the next system action is thus adapted to the cultural background of the user. We use a dialogue management framework based on the concept of probabilistic rules which combines the benefits of logical and statistical methods to dialogue modelling: The probabilistic rules represent the internal models of the domain in a compact form and the unknown parameters included in the probabilistic rules are automatically estimated from data using supervised learning. We investigate whether it is possible to train culture-specific parameters for these probabilistic rules in order to represent cultural patterns in the dialogue management decision process.

The structure of the paper is as follows: In Sect. 2, we present related work in the field of culture-sensitive interface design. Afterwards, the corpus that has been used in this work is described in Sect. 3 and the design and implementation is outlined in Sect. 4. We present the evaluation and our results in Sect. 5, before concluding in Sect. 6.

2 Related Work

Brejcha [1] has described patterns of language and culture in Human-Computer Interaction and has shown why these patterns matter and how to exploit them to design a better user interface. Reinecke and Bernstein [8] have presented a design approach for culturally-adaptive user interfaces to enhance the usability. The authors developed a prototype web application that serves as a to-do list tool. During registration, the users have to input their current country of residence, former countries they lived in and the length of stay in each country. Based on this information, the design and content of the web application is adapted to the user.

Furthermore, Traum [9] has outlined how cultural aspects may be included in the design of a visual human-like body and the intelligent cognition driving action of the body of a virtual human. Therefore, different cultural models have been examined and the author points out steps for a fuller model of culture. Georgila and Traum [2] have presented how culture-specific dialogue policies of virtual humans for negotiation and in particular for argumentation and persuasion may be built. A corpus of non-culture specific dialogues is used to build simulated users which are then employed to learn negotiation dialogue policies using Reinforcement Learning. However, only negotiation specific aspects are taken into account while we aim to create an overall culture-sensitive dialogue manager.

A study presented by Miehle et al. [7] investigated cultural differences between Germany and Japan. The results pointed out that communication idiosyncrasies found in Human-Human Interaction may also be observed during Human-Computer Interaction in a Spoken Dialogue System context. Moreover, the study described by Miehle et al. [6] examined five European cultures whose communication styles are much more alike than the German and Japanese communication idiosyncrasies. The results show that there are differences among the cultures and that it depends on the culture whether there are gender differences concerning the user's preference in the system's communication style. These studies show that culture-adaptive behaviour is an important aspect of spoken user interfaces. This is why we address the task of culture-aware dialogue management.

3 Corpus Description

Our data set is based on recordings on health care topics containing spontaneous interactions in dialogue format between two participants: one is acting as the system is expected to perform while the second one is taking the role of the user of the system. Each dialogue is allocated with a unique dialogue ID and each action is assigned

- a dialogue action number,
- a participant,
- a speaker,
- a dialogue action, and
- the original utterance.

The dialogue action number counts from 1 to n starting with the first dialogue action, where n is equal to the number of dialogue actions of the respective dia-

Table 1 The culture distribution of the dialogue actions in the corpus

Culture	#Dialogue actions
German	4849
Polish	1017
Spanish	1002
Turkish	1024

logue. The participant specifies the two roles *system* and *user* and the speaker indicates which of the predefined speakers was talking. Each speaker is identified by an anonymous speaker ID and a separate table contains profile information about each speaker, including the gender, the culture, the age, the country of birth and the current country of residence. The dialogue action is chosen out of a set of 59 distinct dialogue actions which have been predefined in advance. Furthermore, the original utterance (in the original language) is added to each dialogue action and for each utterance, the topics being talked about are annotated (in English). Moreover, for each dialogue, the system's role is specified. The available system roles are defined as *social companion*, *nursing assistant* and *health expert*. Overall, the corpus covers 258 dialogues. The culture distribution of the dialogue actions is shown in Table 1.

4 Design and Implementation

For the implementation of our culture-aware dialogue manager, we have used the open-source software toolkit OpenDial [5]. It combines the benefits of logical and statistical methods to dialogue modelling by adopting a hybrid approach. Probabilistic rules represent the domain model in a structured format and allow system designers to integrate their domain knowledge. These rules contain unknown parameters that can be estimated from dialogue data using supervised learning. Thus, this hybrid concept allows the system designers to integrate domain-dependent constraints into a probabilistic context. The probabilistic rules formalism is described in [4]. Practically, they are defined as *if...then...else* constructs that map logical conditions to a distribution over possible effects. For the action selection, OpenDial provides utility rules that associate utility values to system decisions. They can be used to find the action with the highest expected utility in the current state. Using these utility rules, we have implemented our dialogue domain as described in Sect. 4.1. Afterwards, we have performed parameter estimation as explained in Sect. 4.2.

4.1 Domain Design

For the implementation of our dialogue domain, we have derived the utility rules from the database described in Sect. 3. We have extracted all possible system actions

```
<rule id="Greet">
 <case>
  <condition>
   <if var="a_u" value="Greet"/>
  </condition>
  <effect util="theta_Greet0">
   <set var="a_m" value="PersonalGreet+AskMood"/>
  </effect>
  <effect util="theta_Greet1">
   <set var="a_m" value="PersonalGreet+AskTask"/>
  </effect>
  <effect util="theta_Greet2">
   <set var="a_m" value="PersonalGreet+ShareJoy"/>
  </effect>
  <effect util="theta_Greet3">
   <set var="a_m" value="AskMood"/>
  </effect>
  <effect util="theta_Greet4">
   <set var="a_m" value="AskPlans"/>
  </effect>
  <effect util="theta_Greet5">
   <set var="a_m" value="PersonalGreet"/>
  </effect>
  <effect util="theta_Greet6">
   <set var="a_m" value="Greet"/>
  </effect>
  <effect util="theta_Greet7">
   <set var="a_m" value="MorningGreet"/>
  </effect>
 </case>
</rule>
```

Fig. 2 Implementation of the rule for greeting based on the corpus

`a_m` in response to a user action `a_u`, regardless of culture. Since three or more consecutive dialogue actions occur rarely in the data, we have limited the number of possible system actions as response to a user action to two. Afterwards, we have implemented a rule for every user action. Overall, we have seven user actions:

- Accept
- Declare
- Goodbye
- Greet
- Reject
- Request
- Thank

As an example, the implementation of the rule for greeting is depicted in Fig. 2. The rule gets activated if the condition is true, i.e. if the user action is *Greet*. Since it is possible to react with one or two consecutive dialogue actions, there are eight possible effects for the next system action. This design approach ensures that only

Fig. 3 Example of a dialogue transcript based on the corpus

```
<interaction>
  <userTurn>
    <variable id="a_u">
      <value>Greet</value>
    </variable>
  </userTurn>
  <systemTurn>
    <variable id="a_m">
      <value>PersonalGreet</value>
    </variable>
  </systemTurn>
</interaction>
```

reasonable pairs of dialogue actions that are covered in the database are included in the domain.

4.2 Parameter Estimation

After implementing the dialogue domain which captures the possible system actions in response to every user action, we have used the supervised learning approach based on the so-called Wizard-of-Oz learning provided within the OpenDial toolkit in order to estimate the parameters (e.g. `theta_Greet1`). This learning approach allows not only to learn from Wizard-of-Oz experiments, but also from dialogue transcripts. As our corpus contains dialogue interactions between two participants where one is taking the role of the system while the other one is taking the role of the user of that system, thus resembling the situation of Wizard-of-Oz experiments, we have created transcripts of these dialogues as input for our parameter estimation. An example of such a transcript is shown in Fig. 3. In this *interaction*, the user action (`a_u`) *Greet* is followed by the system action (`a_m`) *PersonalGreet*.

In the following, we explain the Wizard-of-Oz learning that has been used for the parameter estimation. According to Lison [4], a Wizard-of-Oz interaction is defined as a sequence of state-action pairs

$$\mathscr{D} = \{\langle \mathscr{B}_i, a_i \rangle : 1 \leq i \leq n\}, \tag{1}$$

where $\mathscr{B}_i$ is the dialogue state, a_i the performed wizard action at time i and n the total number of recorded actions. During the learning process, the goal is to learn the posterior distribution of the rule parameters $\boldsymbol{\theta}$ based on the Wizard-of-Oz training data set $\mathscr{D}$. The algorithm takes each state-action pair $\langle \mathscr{B}_i, a_i \rangle \in \mathscr{D}$ and updates the posterior parameter distribution after each pair. This posterior distribution can be decomposed as

$$P(\boldsymbol{\theta} \mid \mathscr{D}) = \eta\ P(\boldsymbol{\theta}) \prod_{\langle \mathscr{B}_i, a_i \rangle \in \mathscr{D}} P(a_i \mid \mathscr{B}_i; \boldsymbol{\theta}), \tag{2}$$

where $P(\boldsymbol{\theta})$ is the prior distribution and $P(a_i \mid \mathscr{B}_i; \boldsymbol{\theta})$ represents the likelihood of the wizard selecting the action a_i in the dialogue state $\mathscr{B}_i$ given the rule parameters $\boldsymbol{\theta}$. This likelihood can be expressed as a geometric distribution

$$P(a_i \mid \mathscr{B}_i; \boldsymbol{\theta}) = \eta(1 - p)^{x-1} p, \tag{3}$$

where x is the rank of action a_i in the utility $U(a_i \mid \mathscr{B}_i; \boldsymbol{\theta})$, η is a normalisation factor and p represents the learning rate of the estimation process. For our experiments, we used $p = 0.2$.

As prior distribution $P(\boldsymbol{\theta})$, we selected a Gaussian distribution with a mean value of 5 and a variance of 1. However, the probabilistic model in Eq. 2 contains both continuous and discrete random variables. This leads to a nontrivial inference problem. OpenDial offers a sampling technique called *likelihood weighting* to approximate the inference process to solve this issue [3]. Hence, the posterior parameter distribution is sampled after the likelihood of the wizard action is calculated. The outcome is then expressed as a *Kernel density estimator* which subsequently can be converted into a Gaussian distribution. This procedure is performed as long as training data is available.

5 Evaluation

After implementing the dialogue domain and creating the dialogue transcript files for each culture (German, Polish, Spanish and Turkish) based on the data set described in Sect. 3, we have used the transcript files to train the rule parameters $\boldsymbol{\theta}$ of the dialogue domain for the different cultures. Thus, four different culture-specific domains have been trained. Proceeding from the initial probability distribution (Gaussian distribution, $\mu = 5, \sigma^2 = 1$), each parameter has been trained based on the appearance of the corresponding system action in the data set. Since the parameters are updated after each *user action - system action* tuple, a more frequent occurrence of a system action in the database causes the shifting of the mean value to a higher value. In contrast, a rare occurrence correlates with a lower mean value, reducing the probability that such a system action is selected. This effect is illustrated in Fig. 4. In the following, we will evaluate whether the trained parameters vary among the different cultures and therefore represent cultural patterns.

In a first step, we have evaluated whether 1000 dialogue actions are enough for training the rule parameters as the Polish, Spanish and Turkish data sets each contain slightly more than 1000 dialogue actions (see Table 1). In order to do so, we have split the German data set consisting of 4849 dialogue actions into four subsets, each containing 1000 dialogue actions. After training with each of the four German

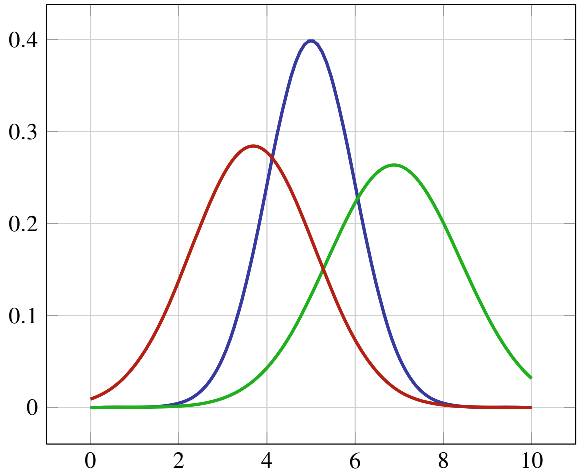


Fig. 4 Probability distribution of each parameter before training (blue) and example probability distributions of two parameters after training, representing a frequently occurring system action (green) and a rarely occurring system action (red)

Table 2 Mean values μ and variances σ^2 of the probability distributions of the parameters $\boldsymbol{\theta}$ after training with 1000 German dialogue actions each

Rule	**Parameter**	**German1** μ (σ^2)	**German2** μ (σ^2)	**German3** μ (σ^2)	German4 μ (σ^2)
Accept	$\theta_{Request}$	5.20 (2.47)	5.29 (2.84)	5.31 (1.93)	5.43 (1.31)
	$\theta_{Acknowledge+Declare}$	5.27 (2.29)	5.01 (2.78)	4.77 (1.97)	5.00 (2.10)
	$\theta_{Acknowledge+Request}$	5.38 (1.94)	5.18 (3.16)	5.04 (1.87)	4.91 (1.60)
Declare	$\theta_{Request}$	6.88 (52.30)	5.54 (35.26)	7.30 (46.18)	5.12 (12.63)
	$\theta_{Acknowledge+Declare}$	3.69 (48.83)	4.28 (28.63)	5.76 (17.71)	3.69 (39.46)
	$\theta_{Acknowledge+Request}$	9.00 (39.67)	7.38 (24.79)	9.34 (59.98)	6.52 (24.59)
Goodbye	θ_{Thank}	4.95 (1.53)	4.79 (2.10)	4.82 (1.66)	4.76 (2.12)
	$\theta_{SimpleGoodbye}$	5.41 (1.52)	5.97 (2.08)	5.71 (1.52)	5.85 (2.15)
	$\theta_{PersonalGoodbye}$	5.05 (1.23)	5.15 (2.06)	4.63 (2.07)	4.98 (2.04)
Greet	$\theta_{PersonalGreet}$	5.24 (3.24)	4.94 (2.83)	5.52 (3.91)	5.62 (1.87)
	$\theta_{Request}$	5.23 (3.17)	5.54 (1.31)	5.11 (3.20)	5.00 (2.16)
	$\theta_{PersonalGreet+AskTask}$	5.21 (2.30)	4.87 (2.26)	5.56 (2.11)	5.32 (1.94)
Reject	$\theta_{Acknowledge}$	4.99 (1.07)	5.09 (1.10)	5.12 (1.00)	5.08 (1.11)
	$\theta_{Request}$	4.87 (1.06)	5.25 (1.07)	5.00 (1.07)	5.05 (1.07)
	$\theta_{Declare+Advise}$	4.97 (1.19)	4.90 (1.04)	4.97 (0.89)	4.98 (1.06)
Request	$\theta_{Declare}$	6.48 (2.35)	5.44 (2.43)	5.57 (4.30)	4.78 (7.85)
	$\theta_{Accept+Declare}$	6.03 (1.82)	5.29 (2.83)	5.54 (2.37)	6.74 (4.79)
	$\theta_{Accept+Request}$	4.94 (3.21)	5.21 (2.55)	4.83 (4.15)	5.47 (14.70)
Thank	$\theta_{AnswerThank}$	5.17 (1.43)	5.27 (1.38)	5.27 (1.25)	4.96 (1.42)
	$\theta_{Goodbye}$	5.11 (1.38)	5.03 (1.48)	5.15 (1.49)	5.13 (1.33)
	$\theta_{PersonalAnswerThank+Goodbye}$	5.17 (1.77)	5.00 (1.35)	5.21 (1.16)	5.02 (1.48)

Table 3 Mean values μ and variances σ^2 of the probability distributions of the three most highly ranked parameters $\boldsymbol{\theta}$ after training with the culture-specific data sets

Rule	German μ (σ^2)	Polish μ (σ^2)	Spanish μ (σ^2)	Turkish μ (σ^2)
Accept	$\theta_{Acknowledge+Request}$ 5.38 (1.94)	$\theta_{Request}$ 5.04 (0.92)	$\theta_{Request}$ 5.13 (1.21)	$\theta_{Request}$ 6.06 (2.80)
	$\theta_{Acknowledge+Declare}$ 5.27 (2.29)	$\theta_{Declare}$ 5.03 (1.02)	$\theta_{Acknowledge+Declare}$ 5.04 (1.24)	$\theta_{Acknowledge+Request}$ 5.38 (3.69)
	$\theta_{Request}$ 5.20 (2.47)	$\theta_{ReadNewspaper}$ 4.99 (1.00)	$\theta_{Acknowledge+Request}$ 4.98 (1.20)	$\theta_{ReadNewspaper}$ 4.88 (3.29)
Declare	$\theta_{Acknowledge+Request}$ 9.00 (39.7)	$\theta_{Request}$ 5.33 (3.02)	$\theta_{Request}$ 5.65 (4.53)	$\theta_{Acknowledge+Request}$ 14.20 (137.18)
	$\theta_{Request}$ 6.88 (52.30)	$\theta_{Acknowledge+Declare}$ 4.98 (4.06)	$\theta_{Acknowledge+Request}$ 5.18 (6.14)	$\theta_{Acknowledge+Declare}$ 3.87 (50.71)
	$\theta_{Acknowledge+Declare}$ 3.69 (48.83)	$\theta_{Acknowledge+Advise}$ 4.94 (4.01)	$\theta_{Acknowledge+Declare}$ 5.17 (4.41)	θ_{Advise} 3.86 (88.75)
Goodbye	$\theta_{SimpleGoodbye}$ 5.41 (1.52)	$\theta_{SimpleGoodbye}$ 5.03 (0.96)	$\theta_{MeetAgainGoodbye}$ 5.34 (1.61)	$\theta_{SimpleGoodbye}$ 5.21 (1.82)
	$\theta_{PersonalGoodbye}$ 5.05 (1.23)	θ_{Thank} 5.03 (1.02)	$\theta_{SimpleGoodbye}$ 5.10 (1.76)	$\theta_{MeetAgainGoodbye}$ 5.00 (1.85)
	θ_{Thank} 4.95 (1.53)	$\theta_{AnswerThank}$ 5.02 (1.03)	$\theta_{Acknowledge}$ 4.98 (2.97)	$\theta_{MorningGoodbye}$ 4.99 (2.36)
Greet	$\theta_{PersonalGreet}$ 5.24 (3.24)	$\theta_{PersonalGreet}$ 5.50 (2.17)	θ_{Greet} 6.21 (2.47)	$\theta_{PersonalGreet+AskMood}$ 5.32 (1.51)
	$\theta_{Request}$ 5.23 (3.17)	$\theta_{PersonalGreet+AskMood}$ 5.34 (3.34)	$\theta_{Declare}$ 5.35 (4.42)	$\theta_{PersonalGreet+AskTask}$ 5.26 (1.97)
	$\theta_{PersonalGreet+AskTask}$ 5.21 (2.30)	θ_{Greet} 5.22 (3.28)	$\theta_{AskMood}$ 5.04 (2.54)	$\theta_{Declare}$ 5.16 (1.74)
Reject	$\theta_{Acknowledge+Declare}$ 5.06 (1.16)	no data available	$\theta_{Acknowledge+Advise}$ 5.11 (1.07)	$\theta_{Request}$ 5.15 (1.20)
	$\theta_{Request}$ 5.04 (1.13)	no data available	$\theta_{Declare+Advise}$ 5.02 (1.59)	θ_{Advise} 5.03 (1.29)
	$\theta_{Acknowledge+Request}$ 5.01 (1.05)	no data available	$\theta_{Request}$ 5.00 (1.73)	$\theta_{Acknowledge+Advise}$ 4.98 (1.11)
Request	$\theta_{Declare}$ 6.48 (2.35)	$\theta_{Reject+Declare}$ 137.04 (40041.25)	$\theta_{Advise+Request}$ 15.51 (187.93)	$\theta_{Declare}$ 6.46 (7.11)
	θ_{Accept} 6.11 (1.89)	θ_{Reject} 129.99 (98009.71)	$\theta_{Reject+Advise}$ 13.10 (373.54)	$\theta_{Accept+Declare}$ 6.44 (5.49)
	$\theta_{Accept+Declare}$ 6.03 (1.82)	$\theta_{Accept+Declare}$ 107.23 (33122.08)	$\theta_{Declare+Request}$ 12.75 (246.97)	$\theta_{Motivate}$ 6.11 (5.07)

(continued)

Table 3 (continued)

Rule	German μ (σ^2)	**Polish** μ (σ^2)	**Spanish** μ (σ^2)	**Turkish** μ (σ^2)
Thank	$\theta_{AnswerThank}$ 5.17 (1.43)	$\theta_{Acknowledge+Goodbye}$ 5.02 (0.98)	$\theta_{Goodbye}$ 5.64 (1.27)	$\theta_{PAnswerThank+Goodbye}$ 5.10 (1.53)
	$\theta_{PAnswerThank+Goodbye}$ 5.17 (1.77)	$\theta_{Goodbye}$ 5.02 (1.01)	$\theta_{Acknowledge+AnswerThank}$ 5.08 (1.30)	$\theta_{Goodbye}$ 4.98 (1.57)
	$\theta_{Goodbye}$ 5.11 (1.38)	$\theta_{PAnswerThank+Goodbye}$ 5.02 (0.95)	$\theta_{Motivate}$ 5.07 (1.34)	$\theta_{AnswerThank+Advise}$ 4.98 (1.62)

training sets, we have obtained similar alternatives for the action selection as the three parameters with the highest mean values are the same for each rule. The corresponding values for the mean μ and the variance σ^2 of the probability distributions for each rule are shown in Table 2. For some rules (e.g. Declare), we get the same ranking for every subset, for others (e.g. Accept) the mean values of the three most highly ranked parameters differ only slightly. However, each of the four German domains results in a similar system strategy, showing that the relevant information is contained in the data. This allows the assumption that training with 1000 dialogue actions is sufficient to train culture-specific parameters. Furthermore, as the average over all four subsets corresponds approximately to the values of subset *German1*, this subset is used to represent the German culture in the further course of the paper. Moreover, this first part of evaluation has revealed that the variance σ^2 of the probability distributions of the parameters is of little importance in the applied sampling technique. The action selection is mainly based on the mean value μ. Hence, we have based the cross-cultural comparison in the second part of our evaluation on the means.

In the second step, we have used the German, Polish, Spanish and Turkish data sets (each containing approximately 1000 dialogue actions) to train the rule parameters $\boldsymbol{\theta}$ of four culture-specific dialogue domains. The three parameters with the highest mean value for each culture are shown in Table 3. It can be seen that the different characteristics of the cultures occasionally have led to different parameters with highest mean values. In the following, we discuss the similarities and differences for each rule.

Accept

If the last user action is accepting what the system has said, the system either reacts with a request for more information or with giving some information to the user (*Declare* or *ReadNewspaper*, what is a special form of giving information). This is the case for every culture tested in our scenario. However, the difference between the cultures lies in whether the system adds an *Acknowledge* (e.g. “Okay.”) or not. While this is very likely for German, it is more unlikely for Polish, Spanish and Turkish (as the parameter with the highest mean does not include it for these cultures).

Declare

If the last user action is a *Declare*, the system may request for more information or give some information to the user. This applies again to each of our cultures. However, we are able to observe two differences between the cultures, namely (1) whether an *Acknowledge* is added or not, and (2) whether the information is presented as an *Advise* or a *Declare*. While it is very likely to add an *Acknowledge* for German and Turkish, it is more unlikely for Polish and Spanish. Moreover, the information will always be presented as a *Declare* (e.g. “He needs help getting up.”) for German and Spanish, while an *Advise* (e.g. “You should help him up.”) may be used for Polish and Turkish.

Goodbye

After the user says goodbye, the system usually answers with any form of saying goodbye. However, the form differs between the examined cultures. While for German, Polish and Turkish a *SimpleGoodbye* (e.g. “Good bye.”) is most probable, for Spanish it is more likely that a *MeetAgainGoodbye* (e.g. “See you.”) is used. Moreover, for German it is also very common that the user’s name is used in a *PersonalGoodbye* (e.g. “Bye Anna.”) and for Polish a *Thank* (e.g. “Thank you.”) or an *AnswerThank* (e.g. “You’re welcome.”) instead of a *Goodbye* might be used.

Greet

If the user greets the system, the most likely system response is also a greet. German, Polish and Turkish uses a *PersonalGreet* (e.g. “Hello Anna.”). However, in our Spanish model, the user is not addressed by name but simply welcomed with a *Greet* (e.g. “Hello.”). Furthermore, in the German, Polish and Turkish model, an *AskTask* (e.g. “How can I help you?”) or an *AskMood* (e.g. “How are you?”) are very likely to be added to the greeting. However, for Spanish only one dialogue action is used.

Reject

In our Polish data, it is never the case that the user rejects anything of the system, even if the topics of conversation are evenly distributed among the cultures. For the other cultures, the system usually reacts with a request for more information or with giving some information to the user. For the latter, the difference between the cultures lies in whether the system uses a *Declare* or an *Advise* to present the information: German uses a *Declare*, while Spanish and Turkish use an *Advise*. Moreover, for German and Spanish usually an *Acknowledge* is added.

Request

A request is the most likely user action, what can be seen from the high mean values in this rule. Obviously, answering the user’s question and thus giving the requested information is the most likely system response for every culture. However, there are differences in how the information is presented. While in the German, Polish and Turkish model, a *Declare* with optional addition of an *Accept* or *Reject* is used, the Spanish model rather utilises an *Advise* than a *Declare*. Moreover, the Turkish model also uses a *Motivate* (e.g. “Good idea!”).

Thank

If the last user action is a *Thank*, the system may react to it with an *AnswerThank* or interpret it in the way that the user wants to end the dialogue and thus answer with a *Goodbye* (or combine both dialogue actions). As the mean values of the three most highly ranked parameters differ only slightly for every culture, we compare the amount of occurrences of these options among the most highly ranked parameters for each culture. We can see that the Polish model always answers with a *Goodbye*, while the others do not always interpret the user's *Thank* in this way. Moreover, the Spanish model also uses a *Motivate* and the Turkish model may add an *Advise*.

6 Conclusion and Future Directions

With the aim of designing a culturally sensitive conversational assistant, in this work we have investigated whether culture-specific parameters may be trained by use of a supervised learning approach. In order to do so, we have used spoken dialogues from four European cultures, namely German, Polish, Spanish, and Turkish, to train a culture-aware dialogue manager. For the implementation we have used the open-source software toolkit OpenDial [5] which is based on the concept of probabilistic rules. Thus, it combines the benefits of logical and statistical methods to dialogue modelling.

For our evaluation we have trained four different culture-specific dialogue domains. For each culture, we have used a data set containing approximately 1000 dialogue actions as we have shown that 1000 dialogue actions are enough for training the rule parameters. Afterwards, we have compared the probability distributions of the trained parameters. Each parameter is expressed as a Gaussian distribution. Thus, we have examined the differences between the cultures in terms of the mean values of the corresponding probability distributions. The evaluation results have shown that the different characteristics of the cultures result in different parameters with highest mean values. Hence, the system response to a user action varies depending on the culture.

In future work, we will examine whether the proposed approach can be extended to other conversational topics and further cultures. In particular, we are interested in non-European cultures since the differences to the cultures studied in this work might be more significant. Moreover, we plan to conduct an evaluation with real users to see, how a varying action selection based on the culture is perceived.

Acknowledgements This work is part of a project that has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 645012. We thank our colleagues from the University of Tübingen, the German Red Cross in Tübingen and semFYC in Barcelona for organizing and carrying out the corpus recordings. Additionally, this work has received funding within the BMBF project "RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots" and the technology transfer project "Do it yourself, but not alone: Companion Technology for DIY support" of the Transregional Collaborative Research Centre SFF/TRR 62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Brejcha J (2015) Cross-cultural human-computer interaction and user experiencedesign: a semiotic perspective. CRC Press
2. Georgila K, Traum D (2011) Learning culture-specific dialogue models from non culture-specific data. In: Universal access in human-computer interaction. users diversity. Springer, pp 440–449
3. Lison P (2013) Structured probabilistic modelling for dialogue management. Doctoral dissertation, University of Oslo
4. Lison P (2015) A hybrid approach to dialogue management based on probabilistic rules. Comput Speech Lang 34(1):232–255
5. Lison P, Kennington C (2016) Opendial: a toolkit for developing spoken dialogue systems with probabilistic rules. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Demonstrations). Association for Computational Linguistics, Berlin, pp 67–72
6. Miehle J, Minker W, Ultes S (2018) What causes the differences in communication styles? a multicultural study on directness and elaborateness. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA)
7. Miehle J, Yoshino K, Pragst L, Ultes S, Nakamura S, Minker W (2016) Cultural communication idiosyncrasies in human-computer interaction. In: 17th annual meeting of the special interest group on discourse and dialogue, p 74 (2016)
8. Reinecke K, Bernstein A (2013) Knowing what a user likes: a design science approach to interfaces that automatically adapt to culture. MIS Q 37(2):427–453
9. Traum D (2009) Models of culture for virtual human conversation. In: Universal Access in Human-Computer Interaction. Applications and Services, pp 434–440 (2009)

Dialog State Tracking with Incorporation of Target Values in Attention Models

Takami Yoshida, Kenji Iwata, Yuka Kobayashi, and Hiroshi Fujimura

Abstract This paper proposes a fully data-driven approach to dialog state tracking (DST) that can handle new slot values not seen during training. The approach is based on a long short-term memory recurrent neural network with an attention mechanism. Unlike with conventional attention mechanisms, we use encoded user utterances and a hypothesis for the slot value (the target value) to calculate attention weights. In addition, while conventional attention mechanisms focus on words that correspond to trained values, the proposed attention mechanism focuses on words that correspond to a target value. Therefore, the DST model can detect unseen values by adding values to the hypothesis as target values. The proposed approach is evaluated using the second and the third Dialog State Tracking Challenge datasets. Evaluation results show that the proposed method improves 10.3 points on unseen slot values.

1 Introduction

In task-oriented spoken dialog systems, dialog state tracking (DST) is used to update a dialog state (i.e., the state of the user's goal). DST is an essential function for a dialog system because the state directly affects the system's response to the user. A dialog state is defined as "a data structure that summarizes the dialog history up to the time when the next system action is chosen" [1]. In practice, a dialog state is a probability distribution over a set of slots and their values as defined in a domain ontology. In the example shown in Table 1, "area, food, pricerange" are slots and "north, japanese,

T. Yoshida (✉) · K. Iwata · Y. Kobayashi · H. Fujimura
Media AI Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, Japan
e-mail: takami.yoshida@toshiba.co.jp

K. Iwata
e-mail: kenji4.iwata@toshiba.co.jp

Y. Kobayashi
e-mail: yuka3.kobayashi@toshiba.co.jp

H. Fujimura
e-mail: hiroshi4.fujimura@toshiba.co.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_9

Table 1 Examples of dialogs and dialog states

System response	User utterance	Dialog state
Hello. How may I help you? (*welcomemsg*())	Japanese restaurant	area=*none*[a], food=japanese, pricerange=none
What area do you prefer? (*request(area)*)	North part of town	area=north, food=japanese, pricerange=none
What price range do you prefer ? (*request(pricerange)*)	Any	area=north, food=japanese pricerange=*dontcare*[b]
How about XXX restaurant. (*offer(XXX)*)	Thank you	area=north, food=japanese pricerange=dontcare

[a] *none* means no value is specified.
[b] *dontcare* means user has no preference.

dontcare, none" are slot values. In practical applications, slot values may be changed during the operation of a dialog system. For example, in the restaurant information domain, the domain ontology changes when new restaurants are added. Therefore, DST should be able to handle a dynamic ontology and unseen slot values.

Traditional DST approaches use handcrafted rules [2, 3] because rule-based approaches are simple and intuitive. However, crafting rules is costly and applying them to a new domain is difficult. Recent DST approaches have been based on deep learning models such as recurrent neural networks (RNNs) [4–6], which need to be trained for predefined slots and values using the domain ontology. However, to deal with new or unseen slot values, training data must be prepared and a new model must be trained.

To overcome this drawback, DST models have been proposed that can handle unseen slot values without retraining [7–9]. The RNN models in [7] use input features after delexicalization. Delexicalization replaces the words relevant to slots and their values with generic symbols. However, delexicalization requires handcrafted rules that compare input words with a large list of synonyms for slots and their values. Another approach uses spoken language understanding (SLU) based on concept tagger architecture [8]. This approach utilizes slot names or slot descriptions to detect unseen values without model retraining. A neural belief tracker [9] estimates a dialog state by comparing the representations of user utterance, system response, and slot values. Although these methods generalize a DST model to unseen slot values, it comes at the cost of crafting the rules, the list of synonyms, slot description, or semantic dictionary.

A problem that is not adequately addressed in the literature is how to deal with unseen slot values without any handcrafted rules. Pointer network-based DST approaches [10, 11] can detect unseen values by utilizing context information. DST models use a pointer mechanism to extract words that are relevant to values. Although the model [10] showed comparatively good performance on the second Dialog State Tracking Challenge (DSTC2) dataset, the accuracies for unseen values were low.

Another DST model [11] showed better accuracies for unseen values in the third Dialog State Tracking Challenge (DSTC3) dataset. However, the results show the tradeoff between the accuracies for seen values versus those for unseen values. BERT-DST [12] extracts span (start position and end position) of the specified slot value from the user utterance and the system response. BERT-DST showed high accuracies for tracking unseen slot values. However, its effectiveness was evaluated using a restaurant name slot and a movie name slot; therefore, its effectiveness with other slots is unknown.

This paper proposes a new attention mechanism for a fully data-driven DST approach that can handle unseen slot values without handcrafted rules and model retraining. This approach is based on the pointer-based DST [11]. Unlike with conventional methods, we use encoded user utterances and a hypothesis for the slot values (the target values) to calculate attention. This enables the DST model to handle an unseen value by directly incorporating it into the attention weights. Attention weights are used to calculate context vectors, which are the weighted sums of word vectors. By comparing the context vectors and word vectors of slot values, the model estimates the dialog state. We evaluate the DST performance based on the proposed approach using the DSTC2 and DSTC3 datasets.

The remainder of this paper is organized as follows: Sect. 2 presents the proposed approach, Sect. 3 shows the experimental results and discusses their meaning and importance and Sect. 4 concludes the paper.

2 Dialog State Tracker

Our proposal is an extension of the DST model in [11], but differs from that approach in that target values are used to calculate attention. The new attention mechanism enables the model to focus on words that are relevant to the target values even if the target values were unseen in training.

Figure 1 illustrates an overview of our DST model. The model consists of encoding and decoding layers. The encoding layer extracts one score from system actions ($\boldsymbol{s}^{\mathrm{s}}$) and another score from user utterances ($\boldsymbol{s}^{\mathrm{u}}$) separately. These two scores are integrated with the previous dialog state ($\boldsymbol{s}^{\mathrm{p}}$) using weight parameters ($\boldsymbol{\beta} = [\beta^{\mathrm{s}}, \beta^{\mathrm{u}}, \beta^{\mathrm{p}}]$) in the decoding layer. The weighted sum ($\boldsymbol{y}$) is regarded as a probability distribution over the slot values after applying the softmax function.

We will describe DST models that use the conventional attention mechanism and the new target value attention mechanism in Sects. 2.1 and 2.2, respectively.

In the sections that follow, we explain a process for a particular slot that includes K values ($v_1, \ldots, v_K$), in which the kth value consists of M_k words. We use n and N for the index of a word in a user utterance and the number of words in the user utterance, respectively.

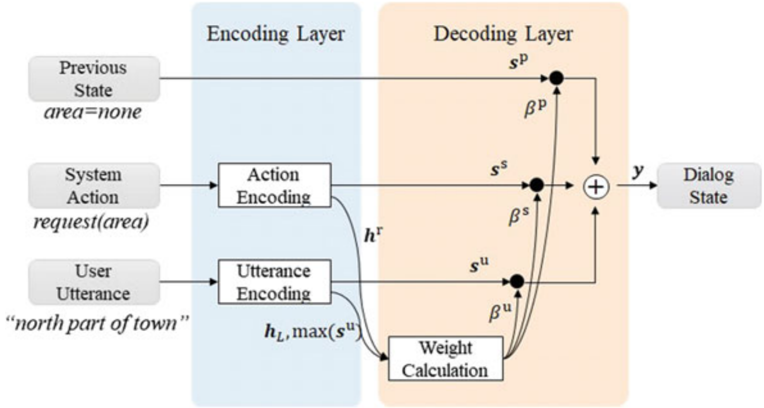


Fig. 1 Schematic diagram of our dialog state tracker

2.1 DST Model with Conventional Attention

This section describes a DST model with a conventional attention mechanism based on the model proposed in [11].

2.1.1 Encoding Layer

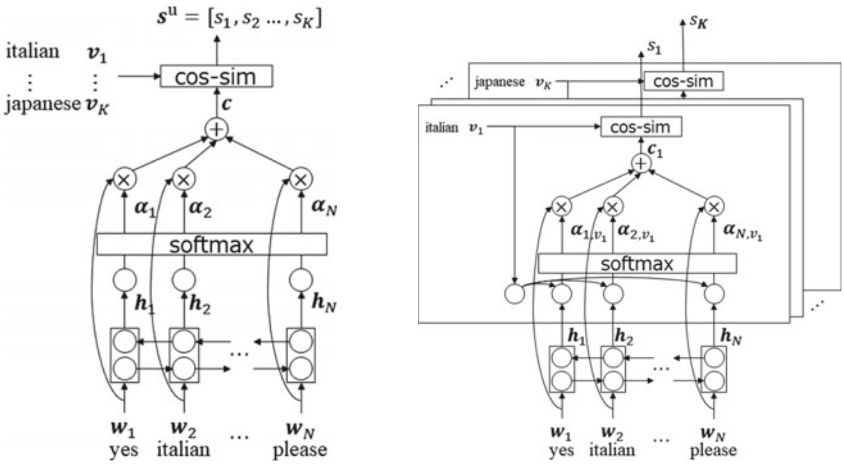
The utterance-encoding and the action-encoding modules calculate two kinds of features. One is possibility scores ($\boldsymbol{s}^{\mathrm{u}}$, $\boldsymbol{s}^{\mathrm{s}}$) that represent whether a slot value is the user's goal. The other is feature vectors ($\boldsymbol{h}^{\mathrm{r}}$, $\boldsymbol{h}_L$) for calculating weight parameters in the decoding layer.

Action Encoding

Action encoding extracts two kinds of features from a system action. One is a feature vector used for calculating weight parameters. The other is a score vector that represents how the system refers to a slot value.

The system action is represented as three features: a system action tag ($\boldsymbol{r}^{\mathrm{act}}$), a target slot feature (r^{s}), and a target value feature (r^{v}). The system action tag is a one-hot vector whose dimension is the same as the number of action tags. The target slot feature and the target value feature are 1 if the previous system action includes the target slot and target value and are 0 otherwise. The three features are concatenated and encoded using a neural network as:

$$\boldsymbol{h}^{\mathrm{r}} = \mathrm{NN}_{\mathrm{sys}}\left(\boldsymbol{r}^{\mathrm{act}} \oplus r^{\mathrm{s}} \oplus r^{\mathrm{v}}\right), \tag{1}$$



(a) Conventional attention mechanism (b) Target value attention mechanism

Fig. 2 Utterance encoding using attention mechanisms

where $\boldsymbol{h}^{\mathrm{r}} \in \mathbb{R}^{d_{\mathrm{m}}}$ is an output vector, d_{m} is a model parameter, $\mathrm{NN}_{\mathrm{sys}}(\cdot)$ is a fully-connected neural network and $\oplus$ is the vector concatenation. The output is used for the weight calculation.

The score vector of the system response ($\boldsymbol{s}^{\mathrm{s}}$) is a $K + 2$ dimensional binary flag. If the system response includes the kth value, the kth component is 1 and is 0 otherwise. The last two components correspond to special values “none” and “dontcare”.

Utterance Encoding

Figure 2(a) shows a block diagram of utterance encoding, in which the module receives a user utterance and encodes it using an attention mechanism.

The user utterance of N word vectors is encoded using a bidirectional LSTM as follows:

$$\boldsymbol{h}_n^{\mathrm{f}} = \mathrm{LSTM}_{\mathrm{fwd}}(\boldsymbol{h}_{n-1}^{\mathrm{f}}, \boldsymbol{w}_n), \tag{2}$$

$$\boldsymbol{h}_n^{\mathrm{b}} = \mathrm{LSTM}_{\mathrm{bwd}}(\boldsymbol{h}_{n+1}^{\mathrm{b}}, \boldsymbol{w}_n), \tag{3}$$

$$\boldsymbol{h}_n = \boldsymbol{h}_n^{\mathrm{f}} \oplus \boldsymbol{h}_n^{\mathrm{b}}, \tag{4}$$

where $\boldsymbol{w}_n \in \mathbb{R}^{d_{\mathrm{w}}}, n = 1, \ldots, N$ is a word vector whose dimension is d_{w}, $\boldsymbol{h}_n^{\mathrm{f}} \in \mathbb{R}^{d_{\mathrm{m}}/2}$, $\boldsymbol{h}_n^{\mathrm{b}} \in \mathbb{R}^{d_{\mathrm{m}}/2}$, and $\boldsymbol{h}_n \in \mathbb{R}^{d_{\mathrm{m}}}$ are hidden states, $\mathrm{LSTM}_{\mathrm{fwd}}(\cdot, \cdot)$ and $\mathrm{LSTM}_{\mathrm{bwd}}(\cdot, \cdot)$ are forward and backward LSTM RNNs. Next, attention weights (α_n) are calculated from the hidden states ($\boldsymbol{h}_n$) as follows:

$$z_n = \mathrm{NN}_{\mathrm{att}}(\boldsymbol{h}_n), \tag{5}$$

$$[\alpha_1, \ldots, \alpha_N] = \mathrm{softmax}([z_1, \ldots, z_N]), \tag{6}$$

where $\mathrm{NN}_{\mathrm{att}}(\cdot)$ is a fully-connected neural network.

Then, a context vector ($\boldsymbol{c} \in \mathbb{R}^{d_{\mathrm{w}}}$) of the user utterance is calculated as a weighted sum of the word vectors as follows:

$$\boldsymbol{c} = \sum_{n=1}^{N} \alpha_n \boldsymbol{w}_n. \tag{7}$$

The score of the user utterance is calculated using cosine similarity between the context vector ($\boldsymbol{c}$) and the word vector of the kth value ($\boldsymbol{h}^{v_k} \in \mathbb{R}^{d_{\mathrm{w}}}$) as follows:

$$s_k = \frac{\boldsymbol{c} \cdot \boldsymbol{h}^{v_k}}{\|\boldsymbol{c}\| \|\boldsymbol{h}^{v_k}\|}. \tag{8}$$

Note that to handle values consisting of multiple words such as “eastern european”, we use the sum of the word vectors as $\boldsymbol{h}^{v_k}$, that is, $\boldsymbol{h}^{v_k} = \sum_{m=1}^{M_k} \boldsymbol{v}_{k,m}$, where $\boldsymbol{v}_{k,m} \in \mathbb{R}^{d_{\mathrm{w}}}$ is the m-th word vector of the k-th value.

To estimate the scores ($\tilde{\boldsymbol{s}} = [s_{\mathrm{none}}, s_{\mathrm{dc}}]$) for the special values “*none*” and “*dontcare*”, we use a separate neural network $\mathrm{NN}_{\mathrm{val}}(\cdot)$ as follows:

$$\boldsymbol{x} = \boldsymbol{h}^{\mathrm{f}}_{N^{\mathrm{u}}} \oplus \boldsymbol{h}^{\mathrm{b}}_1 \oplus \boldsymbol{h}^{\mathrm{r}} \oplus \max(\boldsymbol{s}^{\mathrm{u}}), \tag{9}$$

$$\tilde{\boldsymbol{s}} = \mathrm{NN}_{\mathrm{val}}(\boldsymbol{x}), \tag{10}$$

where $\boldsymbol{x} \in \mathbb{R}^{2d_{\mathrm{m}}+1}$ is the concatenation of the last states of the forward and backward LSTM ($\boldsymbol{h}^{\mathrm{f}}_N, \boldsymbol{h}^{\mathrm{b}}_0$), the system action feature ($\boldsymbol{h}^{\mathrm{r}}$), and the maximum cosine similarity ($\max(\boldsymbol{s}^{\mathrm{u}})$). Finally, the scores s_k and $\tilde{\boldsymbol{s}}$ are concatenated as $\boldsymbol{s}^{\mathrm{u}} = [s_1, \ldots, s_K] \oplus \tilde{\boldsymbol{s}}$. Note that we omit this part from Fig. 2(a) for simplicity. The utterance encoding sends the concatenated score ($\boldsymbol{s}^{\mathrm{u}}$) and feature vector ($\boldsymbol{x}$) to the following processing operation.

Decoding Layer

The decoding layer integrates scores calculated from the user utterance ($\boldsymbol{s}^{\mathrm{u}}$), the score from the system response ($\boldsymbol{s}^{\mathrm{s}}$), and the dialog state of the previous turn ($\boldsymbol{s}^{\mathrm{p}}$) using weight parameters ($\beta = [\beta^{\mathrm{u}}, \beta^{\mathrm{s}}, \beta^{\mathrm{p}}]$) from a neural network $\mathrm{NN}_{\mathrm{weight}}(\cdot)$ as follows:

$$\boldsymbol{\beta} = \mathrm{NN}_{\mathrm{weight}}(\boldsymbol{x}), \tag{11}$$

$$\boldsymbol{y} = \beta^{\mathrm{u}}\boldsymbol{s}^{\mathrm{u}} + \beta^{\mathrm{s}}\boldsymbol{s}^{\mathrm{s}} + \beta^{\mathrm{p}}\boldsymbol{s}^{\mathrm{p}}, \tag{12}$$

$$\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{y}). \tag{13}$$

After applying the softmax function, the model outputs the probability distribution ($\boldsymbol{p} \in \mathbb{R}^{K+2}$) over K slot values, none and dontcare.

2.2 DST Model with Target Value Attention

Figure 2(b) shows a block diagram of utterance encoding based on our attention mechanism. This model calculates attention weights for each target value using the word vectors of the corresponding target value. The attention mechanism is designed to focus the decoder on words in the user's utterance that are relevant to the target value. If the utterance includes a word relevant to the target value, the attention weight for the word will be greater than that for the other words. As a result, the context vector ($\boldsymbol{c}_k$) is similar to the word vector of the target value. Therefore, by comparing the context vector and the target value, the system can detect unseen values.

Instead of Eqs. (5), (6), attention weights (α_{n,v_k}) are calculated using the word vector of the k-th value ($\boldsymbol{h}^{v_k}$) and the hidden states ($\boldsymbol{h}_n$) as follows:

$$z_{n,v_k} = \mathrm{NN}_{\mathrm{att}}(\boldsymbol{h}^{v_k} \oplus \boldsymbol{h}_n), \tag{14}$$

$$\left[\alpha_{1,v_k}, \ldots, \alpha_{n,v_k}\right] = \mathrm{softmax}(\left[z_{1,v_k}, \ldots, z_{n,v_k}\right]), \tag{15}$$

where $\mathrm{NN}_{\mathrm{att}}(\cdot)$ is a fully-connected neural network.

Then, a context vector ($\boldsymbol{c}_k \in \mathbb{R}^{d_w}$) of the user utterance is calculated as a weighted sum of the word vectors as follow:

$$\boldsymbol{c}_k = \sum_{n=1}^{N^{\mathrm{u}}} \alpha_{n,v_k} \boldsymbol{w}_n. \tag{16}$$

Note that the context vector is calculated for each slot value.

The score of the user utterance is calculated using cosine similarity between the context vector ($\boldsymbol{c}_k$) and the value vectors ($\boldsymbol{h}^{v_k}$):

$$s_k = \frac{\boldsymbol{c}_k \cdot \boldsymbol{h}^{v_k}}{\|\boldsymbol{c}_k\| \|\boldsymbol{h}^{v_k}\|}. \tag{17}$$

The remaining parts are the same as the ones described in Sect. 2.1.

2.2.1 Model Training

When training the model, we minimize a loss function that consists of two terms. One is the cross-entropy loss (L_{ce}) between the output probabilities ($\boldsymbol{p}$) and the ground truth label ($\boldsymbol{d}$). The other is the triplet loss [13] (L_{tri}) between the normalized word vector of the ground truth value ($\boldsymbol{h}^{\mathrm{v}_\kappa}$) and the context vectors ($\boldsymbol{c} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K]$).

Table 2 Ontology of the train and test datasets

	Slot	Values		Examples
		All	Unseen	
Train	Area	7	-	Centre, North, ...
	Food	93	-	Afghan, African, ...
	Pricerange	5	-	Cheap, moderate, ...
Test	Area	17	14	Girton, arbury, ...
	Food	30	10	Cafe food, ...
	Pricerange	6	1	Free, cheap, ...

$$L = L_{\mathrm{ce}}(\boldsymbol{d}, \boldsymbol{p}) + L_{\mathrm{tri}}(\boldsymbol{h}^{\mathrm{v}_\kappa}, \boldsymbol{c}), \tag{18}$$

$$L_{\mathrm{ce}}(\boldsymbol{d}, \boldsymbol{p}) = -\sum \boldsymbol{d} \log \boldsymbol{p}, \tag{19}$$

$$L_{\mathrm{tri}}(\boldsymbol{h}^{\mathrm{v}_\kappa}, \boldsymbol{c}) = \frac{1}{K}\left(\sum_{j \neq \kappa} \max\left\{0, \|\boldsymbol{h}^{v_\kappa} - \boldsymbol{c}_\kappa\| - \|\boldsymbol{h}^{v_\kappa} - \boldsymbol{c}_j\| + \varepsilon\right\}\right), \tag{20}$$

where ε and κ are a margin parameter and an index of the ground truth value, respectively. The triplet loss helps the model learn the context vector calculation that assigns a smaller distance to $(\boldsymbol{h}^{\mathrm{v}_\kappa}, \boldsymbol{c}_\kappa)$ and bigger distance to $(\boldsymbol{h}^{\mathrm{v}_\kappa}, \boldsymbol{c}_{j\neq\kappa})$. Note that we add the triplet loss when the corresponding user utterance includes a slot value.

We introduce the "*Sampling with Decay*" technique [18] to feed the previous state ($\boldsymbol{s}^{\mathrm{p}}$). During model training, we randomly sample the previous state from the ground truth previous state ($\boldsymbol{d}$) with a probability of q or from the estimated state ($\boldsymbol{p}$) with a probability of $1-q$. We define q with the decay function dependent on the index of training epochs (e) as $q = \frac{\mu}{\mu+\exp(e/\mu)}$ where μ is a parameter. As the training proceeds, the probability of (q) feeding ground truth decreases gradually [18].

3 Experiments

We evaluated our model using the DSTC2 and DSTC3 datasets [14, 15]. The datasets include human-computer dialogs. Users interacted with dialog systems to search for restaurants by specifying constraints. Among the slots included in the DSTC3 dataset, we used "area", "food", and "pricerange". We excluded the "childrenallowed," "type," "hasinternet," "hastv," and "near" slots because these slots are not included in the DSTC2 dataset. We also excluded the "name" slot because word vectors for several values were not obtained. A summary of the slots and slot values are shown in Table 2. In the DSTC3 test dataset, 36.0%, 17.3%, and 3.5% of the dataset refer to unseen values in the area, food, and pricerange slots, respectively.

3.1 Experimental Condition

The evaluation metric is the accuracy of a value estimation. The accuracy is calculated as the fraction of turns where the top dialog state hypothesis is correct [14]. The ground truth label is under "Scheme A", which defines the label as the most recently asserted value and "Schedule 2"[14].

We implemented a prototype DST system based on the proposed method using a chainer [16]. One-best ASR results were used as inputs to the encoding layer described in Sect. 2. Contractions were converted to their original forms (e.g.., i'm to i am). Then, each word was converted to a 300-dimensional word vector using GloVe [17]. We used the GloVe model available at the GloVe website.[1] During training, the parameters of the GloVe model were fixed.

As a baseline method, we implemented a DST method that does not use the target value for attention weight calculation that is explained in Sect. 2.1. DST based on the RNN with the proposed attention mechanism and the conventional attention mechanism are called as "*Prop*" and "*Comp*", respectively.

We also evaluated two conventional methods, *BERT-based DST* [12] and *pointer-based DST* [11]. We implemented BERT-based DST using the publicly available BERT-DST source codes.[2] We used default parameters in the source code except the slot value dropout ratio. We trained models using the slot value dropout ratio of $[0, 0.1, \ldots, 0.4]$ and selected the best one. Note that the accuracies of the BERT-based DST were calculated using only pointable samples. The accuracies of the pointer-based DST are the ones reported in [11].

For LSTM_{fwd} and LSTM_{bwd}, we used 1-layer LSTMs with 32 nodes. For NN_{sys}, NN_{att}, NN_{val}, and $\text{NN}_{\text{weight}}$, we used 1-layer, 3-layer, 4-layer, and 4-layer fully connected NNs, respectively.

Hyper parameters are as follows: Adam optimizer using the chainer implementation; learning rate, 0.001; gradient clipping, 1.0; mini batch size, 32; sampling parameter μ, 12; and maximum epoch, 200. We also applied word dropout that randomly replaced the word vectors of user utterances with zero vectors. These hyper parameters were identical for the Comp and Prop models.

3.2 Results and Discussion

Table 3 shows DST accuracies on the DSTC2 dataset. The upper part shows the results reported in the DSTC2 [14] and the lower part shows the results of fully data-driven DST methods. In all slots, RNN with rules shows the best performance; from 0.2 to 0.7 point higher than Comp and Prop. Prop and Comp achieve almost the same performance as Focus baseline without using any handcrafted rules. The differences between Comp and Prop is less than 0.4 point. This is reasonable because Comp

[1] https://nlp.stanford.edu/projects/glove/.

[2] https://github.com/guanlinchao/bert-dst.

Table 3 Test accuracies on the DSTC2 test dataset

Tracker	Area	Food	Price
Focus baseline [14]	90.8	83.9	92.9
RNN with rules [7]	**92.4**	**85.6**	**93.0**
Pointer-based DST [11]	84.7	84.4	83.7
BERT-based DST	88.2	76.6	92.0
Comp	91.7	84.9	92.5
Prop	**91.9**	**85.1**	**92.8**

Table 4 Test accuracies on the DSTC3 test dataset

Tracker	All values			Unseen values		
	Area	Food	Price	Area	Food	Price
Focus baseline [14]	81.1	90.5	88.4	67.6	**88.1**	87.6
RNN with rules [7]	**88.5**	**91.0**	**93.2**	**85.3**	82.3	**92.3**
Pointer-based DST [11]	**80.6**	79.6	66.9	**71.5**	59.5	52.7
BERT-based DST [12]	61.3	79.2	90.4	25.8	61.9	49.4
Comp	75.2	**83.5**	91.4	55.7	66.2	52.7
Prop	76.1	**83.5**	**91.7**	57.8	**71.3**	**79.6**

can extract words relevant to seen values without using the target value attention mechanism.

Table 4 shows DST accuracies on the DSTC3 test dataset. This table reveals the gap between rule-based DST and fully-data driven DST models. Among fully-data driven models, Comp shows better performance than does BERT-based DST under all conditions. Prop improves accuracies further. Pointer-based DST shows high accuracies on the area slot, but the accuracies are lower on the other slot.

The accuracy of Prop on the area slot is lower than that on the food and pricerange slots. The lower score might be caused by values consisting of multiple words such as "new chesterton", "kings hedges". In the training dataset, the area slot includes values consisting of single words such as "north" and "south". Therefore, the DST models learned to extract only single words from user utterances. We observed that BERT-based DST also suffers from such word length mismatches.

We perform ablation experiments on the DSTC3 test dataset to analyze the effectiveness of different components. The results are shown in Table 5. The accuracies

Table 5 Ablation study on the DSTC3 test dataset

Tracker	All values			Unseen values		
	Area	Food	Price	Area	Food	Price
without TVA[a] (Comp)	75.2	83.5	91.4	55.7	66.2	52.7
+ SVD[b]	**76.5**	**84.6**	91.4	**59.2**	65.9	52.7
with TVA	73.3	80.6	91.1	57.8	66.5	76.2
+ SVD	76.1	82.1	91.6	57.8	66.5	76.2
+ SVD + TL[c] (Prop)	76.1	83.5	**91.7**	57.8	**71.3**	**79.6**

[a] TVA represent target value attention
[b] SVD represent slot value dropout
[c] TL represent triplet loss

of "without TVA" and "with TVA" are almost the same. However, integration of the three components achieves the comparative or higher accuracies under most conditions. The effect of our method is more pronounced on unseen values than all values. On average over the three slots, the score of prop (69.6) is 10.3 points higher than that of Comp + SVD (59.3).

One of the drawbacks is that our method requires a word vector for the target value. This method cannot track values whose word vector is not available. This is why we exclude the name slot for the experiments. One promising approach is to use a part of a word as a unit.

Another drawback is that the proposed model tends to fail when two values include the same word. In the test dataset, "pub" is included in the food slot ("pub food") and the type slot ("pub"). When a user says "I'm looking for a pub food restaurant," the ground truth of the dialog state is "food = pub food, type = pub." On the other hand, if a user says "I'm looking for a pub," the ground truth of the dialog state is "food = none, type = pub." The DST model with target value attention tends to estimate such user utterances as "food = pub food."

4 Summary

This paper proposed a fully data-driven approach to DST based on a target value attention mechanism. Unlike conventional attention mechanisms, the proposed attention mechanism utilizes the hypothesis for slot values in order to focus on unseen values without model retraining. We used the DSTC2 and DSTC3 datasets to evaluate the DST model based on the proposed approach. For unseen values, the results showed that using the proposed attention mechanism led to a 10.3-point improvement over the conventional attention mechanism.

Future research will aim to improve the accuracy of the model for both seen and unseen values as well as extend the proposed approach to handle unseen slots.

Acknowledgements The authors would like to thank Professor Masami Akamine for his insightful comments and suggestions.

References

1. Williams JD, Raux A, Henderson M (2016) The dialog state tracking challenge series: a review. Dialogue Discourse 7(3):4–33
2. Sun K, Chen L, Zhu S, Yu K (2014) A generalized rule based tracker for dialogue state tracking. In: Spoken language technology workshop (SLT). IEEE, pp 330–335
3. Kadlec R, Vodolan M, Libovickỳ J, Macek J, Kleindienst J (2014) Knowledge-based dialog state tracking. In: Spoken language technology workshop (SLT). IEEE, pp 348–353
4. Yoshino K, Hiraoka T, Neubig G, Nakamura S (2016) Dialogue state tracking using long short term memory neural networks. In: Proceedings of seventh international workshop on spoken dialog systems (IWSDS), pp 1–8
5. Zilka L, Jurcicek F (2015) Incremental LSTM-based dialog state tracker. In: Workshop on automatic speech recognition and understanding (ASRU). IEEE, pp 757–762
6. Liu B, Lane I (2017) An end-to-end trainable neural network model with belief tracking for task-oriented dialog. Proc Interspeech 2017:2506–2510
7. Henderson M, Thomson B, Young S (2014) Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In: Spoken language technology workshop (SLT). IEEE, pp 360–365
8. Rastogi A, Hakkani-Tür D, Heck L (2017) Scalable multi-domain dialogue state tracking. In: Automatic speech recognition and understanding workshop (ASRU). IEEE, pp 561–568
9. Mrkšić N, Vulić I (2018) Fully statistical neural belief tracking. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 2, pp 108–113
10. Xu P, Hu Q (2018) An end-to-end approach for handling unknown slot values in dialogue state tracking. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1, pp 1448–1457
11. Yoshida T, Iwata K, Fujimura H, Akamine M (2018) Dialog state tracking for unseen values using an extended attention mechanism. In: International workshop on spoken dialog system technology (IWSDS)
12. Chao G, Lane I (2019) BERT-DST: scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In: Proceedings of interspeech, pp 1468–1472
13. Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y (2014) Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1386–1393
14. Henderson M, Thomson B, Williams J (2014) The second dialog state tracking challenge. In 15th annual meeting of the special interest group on discourse and dialogue, vol 263
15. Henderson M, Thomson B, Williams JD (2014) The third dialog state tracking challenge. In: Spoken language technology workshop (SLT). IEEE, pp 324–329
16. Tokui S, Oono K, Hido S, Clayton J (2015) Chainer: a next-generation open source framework for deep learning. In: *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*
17. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP), pp 1532–1543
18. Zhang W, Feng Y, Meng F, You D, Liu Q (2019) Bridging the gap between training and inference for neural machine translation. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp 4334–4343

Delay Mitigation for Backchannel Prediction in Spoken Dialog System

Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu

Abstract To provide natural dialogues between spoken dialog systems and users, backchannel feedback can be used to make the interaction more sophisticated. Many related studies have combined acoustic and lexical features into a model to achieve better prediction. However, extracting lexical features leads to a delay caused by the automatic speech recognition (ASR) process. The systems should respond with no delay, since delays reduce the naturalness of the conversation and make the user feel dissatisfied. In this work, we present a prior prediction model for reducing response delay in backchannel prediction. We first train both acoustic- and lexical-based backchannel prediction models independently. In the lexical-based model, prior prediction is necessary to consider the ASR delay. The prior prediction model is trained with a weighting value that gradually increases when a sequence is closer to a suitable response timing. The backchannel probability is calculated based on the outputs from both acoustic- and lexical-based models. Evaluation results show that the prior prediction model can predict backchannel with an improvement rate on the F1 score 8% better than the current state-of-the-art algorithm under a 2.0-s delay condition.

A. I. Adiba · T. Homma (✉) · D. Bertero · T. Sumiyoshi · K. Nagamatsu
Hitachi, Ltd., Tokyo, Japan
e-mail: takeshi.homma.ps@hitachi.com

A. I. Adiba
e-mail: amalia.adiba.dw@hitachi.com

D. Bertero
e-mail: dario.bertero.zt@hitachi.com

T. Sumiyoshi
e-mail: takashi.sumiyoshi.bf@hitachi.com

K. Nagamatsu
e-mail: kenji.nagamatsu.dm@hitachi.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_10

1 Introduction

Spoken dialog systems (SDS) are increasingly deployed in our daily lives in the form of smart phones, smart speakers, car navigation systems, and more. These systems can conduct simple tasks or answer questions. For this type of domain, the systems generally assume that a user makes one utterance per turn. This assumption is obviously far from natural, however, as humans often make several utterances in one turn during a conversation, the listener usually provides feedback, known as a backchannel. In SDS, a backchannel is defined as any kind of feedback a system provides to a user, including short phrases ('uh-huh,' 'right,' etc.) and nods. Backchannels can express sympathy and acknowledge, and they can encourage users to keep talking and let them know that the system has received the information. Without backchannels, users would be concerned about whether the system is still listening to the conversation or not. Therefore, if an SDS can predict the backchannel and generate a backchannel naturally, as humans do, it will be better able to converse with users.

When designing an SDS with backchannel functions, the backchannel models must accurately detect the backchannel and predict the suitable time to respond. The inputs of backchannel detection models come from user speech that has been converted into acoustic features such as loudness, pitch, and MFCC. The input could also include lexical features, e.g., the output of an automatic speech recognition (ASR) system. According to previous work, lexical features are the most informative features and provide better accuracy than acoustic features [11, 18]. However, accurate ASR needs time for processing, which delays the system responses. In addition, many SDS applications use cloud-based ASR, and in these cases, the delay of system responses will be caused not only by ASR processing but also by the latency of network communication. Figure 1 shows an example of this delay problem. Response delays of a few milliseconds might be acceptable as 'thinking time'; after all, response delays sometimes also occur in human-human conversation, such as during question-answering or reservation domain tasks. According to an earlier report [19], in human-robot conversations, the maximum user preference for a system to generate a response is one second. However, backchannel feedback is different depending on turn response. Delayed feedback might cause an unsophisticated response, as backchannel is a kind of a reflex action. If the SDS does not make a backchannel response at an appropriate time, it will produce unnatural conversation, possibly making the user feel dissatisfied. Thus, the SDS must start to give feedback as soon as the user has finished speaking.

To resolve these issues, we propose a modelling method that can mitigate the response delay problem in backchannel detection. Our contributions in this work are summarized as follows.

- We propose a backchannel prior prediction model. Prior prediction means that the model predicts backchannel events within only the available ASR output, which consists of the words derived from the beginning of the utterance to the several seconds before the final ASR outputs occur.

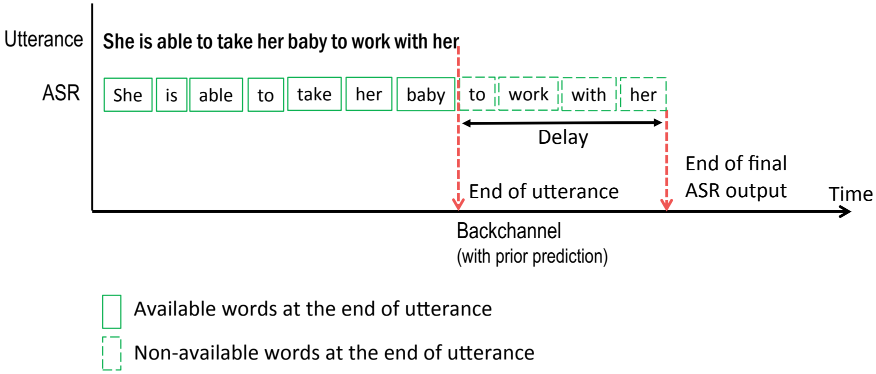


Fig. 1 Response delay caused by ASR in the spoken dialog system. Our proposed prior prediction mitigates the delay and predicts backchannel earlier than the end of the final ASR output.

- We propose an early loss (EL) function, in which a weighting on the loss value gradually increases when a word is closer to the final ASR outputs.
- We show that the prior detection model successfully predicts the backchannel events before the final ASR outputs.
- We show that our proposed prior detection model achieves better performance in the prediction of suitable backchannel timing than the baseline method in which the model is trained without the EL function.
- We show it is not necessary to train the model in various delay conditions independently, since even under different conditions, our EL-based model delivered better predictions.

2 Related Work

Prior work on backchannels has used different types of prediction methods and a variety of input features. These input features include acoustic, lexical, and visual features such as gaze movement.

One main task is to predict backchannel feedback from user's speech. Most related work has taken a rule-based approach. Ward and Tsukahara [22] proposed a method with acoustic features to predict a backchannel and reported that a few millisecond regions of low pitch are a good predictor of backchannel feedback. Truong et al. [21] developed a rule-based backchannel prediction model using power and pitch information. Other work has used a classifier to train the model. Morency et al. [14] proposed a multimodal approach in which acoustic, lexical, and eye gaze data are used as the features and the model is trained with a hidden Markov model (HMM) or conditional random fields (CRFs). Training a prediction model in a continuous manner using the LSTM neural network has also been proposed by [17, 18], who evaluated the method with various input features and reported that the performance with both lexical (e.g., part of speech, word) and acoustic (e.g., power, pitch) features was significantly better than with just acoustic features. They also proposed a prediction point within a predefined window length and reported that the performance peaks at 1.5 s of the window. By using a predefined window as a prediction point, they could also predict backchannels that appear in the middle of an utterance. However, choosing a predefined area might produce unnatural backchannel feedback, as the time gap between the detection point and the backchannel timing can vary significantly.

The other task is to detect speaking acts (including backchannel) [15]. Predicting different types of backchannel has also been proposed [11]. Skantze [20] proposed a continuous model to predict upcoming speech activity in a future time window. They categorized onset prediction into SHORT and LONG, where SHORT utterances can be considered as backchannel. The prediction of turn-taking with multitask learning by combining backchannel and filler prediction has also been proposed [7]. A general study on the correlations or synchrony effect in the prosody of backchannels also exists [10].

In a practical spoken dialog system, the ASR needs a certain length of processing time to output lexical features from speech data. Many researchers proposed backchannel prediction methods using the lexical features. However, most of the researchers did not discuss the influence of ASR processing time to obtain the lexical features. A few researchers conducted research on backchannel prediction using lexical features under the assumptions; ASR gives the lexical features with no delay [17, 18], or ASR takes a constant and short 100ms processing time to output the lexical features [20].

To predict backchannels using lexical features, the system must wait until ASR outputs the lexical features. Hence, the timing of backchannel feedbacks is delayed by the ASR processing time. Moreover, the ASR processing time will be varied depending on characteristics of input speech signals. To generate backchannels at an appropriate timing, the backchannel prediction method must take variable ASR processing time into account.

To the best of our knowledge, our paper is the first to investigate how to accurately predict backchannel based on lexical features considering the variable delay caused by ASR. Our work contributes to reducing the delay of the backchannel responses of the spoken dialog system.

3 Proposed Method

Backchannel prediction is a task that predicts whether the current user's utterance is a backchannel point or not. We extract acoustic and lexical features from an input utterance and feed them to the model to predict the output. The proposed method introduces prior prediction in which the backchannel will be detected several seconds before the end of an utterance. The model is trained based on LSTM-RNN for the

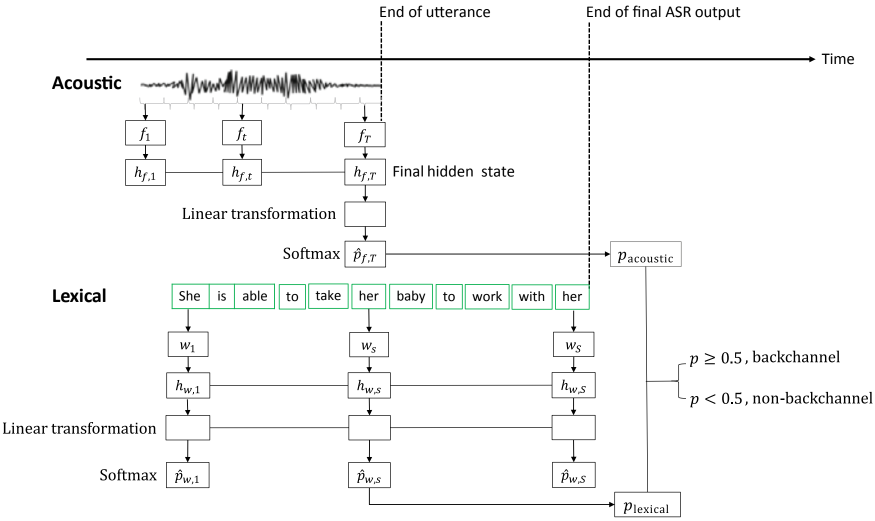


Fig. 2 Neural network architecture for backchannel prediction.

sequential information. We feed acoustic and lexical features to the network and train both models independently. In this section, we explain the overall architecture of the system for the backchannel prediction. We then explain the early loss (EL) function, which is the primary contribution of this study.

3.1 Neural Network for Classification

In a real-life scenario, a system receives both audio channel and ASR outputs. However, the corresponding words come several seconds after the end of an utterance. This means that the audio and the words do not always correspond to each other, and the difference between the time at the end of an utterance and the time at the final ASR output results is a response delay. Let us assume that the acoustic extraction process is not producing any delay. Therefore, prior prediction is not necessary in the acoustic case. In the lexical case, prior prediction is necessary to mitigate the response delay. Because the treatment of the acoustic and lexical cases is different, we train both models independently.

In this study, we trained the backchannel prediction model using an LSTM-RNN-based model [8]. The overview of our system is depicted in Fig. 2. In LSTM, a memory cell is used to store information over time. For lexical features, a token represents word outputs from ASR (w_1, w_2, ..., w_S), where S is the token number at the final ASR output. For acoustic features, a token represents a feature vector of each frame (f_1, f_2, ..., f_T), where T is the token number at the end of the utterance. Standard LSTM takes a sequence of tokens as input and then outputs a sequence of hidden representations $(h_{f,1}, h_{f,2}, ..., h_{f,T})$ for the acoustic-based model and $(h_{w,1}, h_{w,2}, ..., h_{w,S})$ for the lexical-based model. The input layer consists of the extracted features for a corresponding utterance. The hidden layer $h_{f,T}$ or $h_{w,s}$ is used to predict the backchannel probability outputs $\hat{p}_T$ or $\hat{p}_s$, respectively, where s is the latest token number within the words outputted from ASR at the timing of the end of utterance. In our case, the target output is the probability of backchannel. The softmax function is used to compute the probability at the i-th utterance. Finally, the backchannel's probability (p) is calculated as

$$p = \lambda \hat{p}_T + (1 - \lambda)\hat{p}_s, \tag{1}$$

where $\hat{p}_T$ is the backchannel probability from the acoustic-based model, $\hat{p}_s$ is the probability from the lexical-based model, and $\lambda \in [0, 1]$ is a hyper-parameter to control the weight for each model's output.

The loss is calculated only at the last token with a cross-entropy loss function for the acoustic-based model. In the other case, for the lexical-based model, the loss is calculated with an EL function. To clarify the effect of our EL function, we simplify the network and train the model with unidirectional LSTM for both the acoustic and lexical models. The label used for training the model is a one-dimensional binary label, which represents whether the corresponding utterance is a true backchannel (label = 1) or a false one (label = 0). The output of the model is a probability score for backchannel within 0 to 1.

3.2 Early Loss (EL) Function for Prior Prediction

Our prior prediction model plays two main roles. The first role is to mitigate delays in the extraction of input features. Let us consider a case where backchannel prediction is done using the lexical features, i.e., ASR outputs. After a user says the final word of an utterance, the ASR outputs this word several seconds after the user has said it. This "several seconds" is the delay we focus on. If the SDS must generate a backchannel response to the user as soon as the end-of-utterance has occurred, the backchannel prediction must be done using only the available input features derived from the beginning of the utterance to the "several seconds" before the end-of-utterance occurs. The prior prediction model can accomplish this functionality.

The second role of the prior detection model is to emphasize the end part of a sequence. We assume the SDS makes a backchannel only when the user stops speaking, even though in human-human dialogue a backchannel can happen anytime. The reason for this assumption is that we want to focus on backchannels as a confirmation function in which users usually take a brief pause to get confirmation about whether the system is still listening to the conversation or not. Therefore, we define the ground truth of a backchannel to be the end of the sequence. A simple way to predict backchannel using only the available input features is to calculate the loss for each token in a sequence with the cross-entropy loss function which has uniform weighting. However, this method leads to a higher chance of a false positive prediction, since the loss of the first token has an equal weight with a token number in the final ASR output. Intuitively, failing to predict a backchannel at a token very close to token number S should receive a higher weight than failing to predict a backchannel at a token far from token number S. Therefore, our strategy is to adaptively modify the weight value depending on how early the model predicts the backchannel. To overcome this challenge, we set the loss at a token very close to token number S to be higher than the loss at a token far from token number S in a sequence. This weighting loss idea originally comes from an earlier method for driver activity anticipation [9]. Chan et al. [3] applied the same idea for anticipating accidents in a dashcam video, which achieved accident anticipation about two seconds before it occurred.

In this work, we consider two types of utterance: those that do not encourage a backchannel response (non-BC utterance), and those that do (BC utterance). For a non-BC utterance, the system should not generate a backchannel feedback in which prior prediction is not necessary. Conversely, for a BC utterance, the system should generate a backchannel several seconds before the end of the final ASR outputs to mitigate response delay. Following [9], we use different loss calculation methods for the BC utterance (**positive**) and the non-BC utterance (**negative**). The loss for a negative utterance is a standard cross-entropy loss. In contrast, to calculate the loss function of a positive utterance, we multiply a weighting function by the standard cross-entropy loss; the weighting function gradually increases when a token is closer to token number S. Finally, the loss function is calculated as follows:

For the **positive** case:

$$L_p = \sum_{s=1}^{S} -e^{-(S-s)} \log(\hat{p}_s), \tag{2}$$

For the **negative** case:

$$L_n = \sum_{s=1}^{S} -\log(1 - \hat{p}_s), \tag{3}$$

where $\hat{p}_s$ is the backchannel probability at token s.

4 Experiments

4.1 Dataset

We use the Switchboard dataset [6] to model backchannel prediction. This dataset consists of English conversations between participants who were asked to discuss a specific topic that was chosen randomly from 70 possibilities. We use annotation from the Switchboard Dialog Act Corpus (SwDA),[1] as we require dialogue act labels (e.g., backchannel, statement) for the annotation. However, the SwDA corpus only maps the dialogue acts with lexical and turn information; it does not map dialogue acts to timing information in the original data of the Switchboard dataset. Because we are interested in delay mitigation, we combine the SwDA corpus with the NXT Switchboard corpus [2] to obtain utterance timing information. With the NXT Switchboard corpus, we can access turn, dialogue act, lexical, utterance, and even word timestamp information within one framework. We exclude utterances in the dataset that do not have timing information.

Previous works have defined an utterance on the basis of pauses [12] or turn shift (speaker change) [1, 13]. We follow the work of [13] and represent a conversation between two speakers as a sequence of utterances that has been sorted based on the start-talk time information available in the corpus $(u_1, u_2, \cdots, u_I)$, where u_i is the i-th utterance in the conversation. We also use dialogue act in corpus $(da_1, da_2, ..., da_I)$, where da_i is the dialogue act for the i-th utterance in the conversation. The backchannel label will be defined as true if the following two conditions are met:

- da_{i+1} is a backchannel
- u_i and u_{i+1} have a different speaker

The final dataset that we use contains 114,315 utterances, with 13.62% of them labelled as true backchannels.

4.2 Features

Following the work in [16], we extract 21 types of acoustic features with 88 dimensions in total using a eGeMAPs [4] configuration with the OpenSmile toolkit [5]. These features include power, pitch, and MFCC. Instead of part-of-speech (POS), we use words in the manual transcripts as lexical features. Note that this is not the ideal approach, as the ASR system might provide different outputs than the transcripts. Lexical features always start from [silb], which is a special word at the beginning of sentence, followed by the first word. We convert words into word indices, where each of the index numbers corresponds to an individual word. The word embedding is then used as input for our model architecture. We found that our dataset in the fastText model trained with the Common Crawl dataset[2] had the smallest number of unknown words. Thus, the word embedding function was conducted using the fastText model with 300 dimensions.

[1] https://github.com/cgpotts/swda.

[2] https://fasttext.cc/.

Table 1 Simulated example of available words in an utterance for various lengths of delay.

Delay (s)	Timestamp for each word											
	0.16	0.31	0.69	0.85	1.08	1.38	1.57	1.60	1.83	2.02	2.63	2.95
0	we	were	I	was	lucky	too	that	I	only	have	one	brother
0.5	we	were	I	was	lucky	too	that	I	only	have		
1	we	were	I	was	lucky	too	that	I	only			
1.5	we	were	I	was	lucky	too						
2	we	were	I	was								

4.3 Experimental Setup

We evaluated the model using 5-fold cross-validation: one fold for the test dataset, 70% of the remaining folds for the training dataset, and 30% for the validation dataset. Each dialogue conversation was split to ensure that the same speaker did not appear in the training, test, and validation sets. For training, we increased the number of positive utterances up to the same as negative utterances by using random oversampling in the training set. We optimized the loss function using an Adam optimizer with a learning rate of 0.0001. At the end of each epoch, we computed the macro F1 score on the validation set. We ran for a maximum of 100 epochs and stopped training if there was no improvement in the validation macro F1 scores for five consecutive epochs. The minibatch size was 10. We took the highest validation performance and applied it to the test data for evaluation. We trained a one-layer LSTM-RNN model using the data. The size of hidden layer is 64 units in the acoustic-based model, and 128 units in the lexical-based model. Afterwards, we evaluated the models with a manipulated utterance for each length of delay. We set λ to 0.5. All parameters were determined on the basis of the best macro F1 score on the validation dataset.

4.4 Simulation Under Various Delay Condition

In real-time conditions, the available ASR output will differ depending on the length of delay and on the duration time for each word in an utterance. For example, the word ‘conversation’ might have a longer duration than ‘yes’. To evaluate the model under various delay conditions, we manipulated the utterance text in the test dataset. The timestamp for each word in the i-th utterance in the conversation is $(d_1^i, d_2^i, ..., d_S^i)$, where S is the number of words from the beginning of the utterance up to the end of the final ASR output (see Fig. 1). Given that the ASR delay is α (in seconds), the available words in the i-th utterance at the end of the utterance are the words in which the timestamp is lower than or equal to $d_S^i - \alpha$. A manipulated utterance for each length of delay is shown in Table 1 as an example.

4.5 Baseline Models

For comparison to our proposed model, we use two baselines configuration from related state-of-the-art studies.

4.5.1 LSTM-RNN Prediction Using Acoustic and POS Combination Features

We re-implement LSTM-RNN originally proposed by Roddy et al. [16], where they used LSTM-RNN for three prediction tasks: prediction at pauses, prediction at onsets, and prediction at overlap. Here, we re-implement only prediction at onsets, which represents a prediction of the length of the next utterance. Roddy et al. stated that SHORT utterances can be considered as backchannels. They used various features (e.g., acoustic, word, and part-of-speech (POS)) and then compared the performance for each feature type. They reported that the best performance was achieved when they combined acoustic and POS features to train the model. Given an input utterance, we implement their method to predict whether the length of the next utterance will be SHORT or not. We follow their setup parameters and implement them for our dataset.

4.5.2 Time-Asynchronous Sequential Networks (TASN)

We re-implement TASN from [12]. In TASN, each feature (word and acoustic) is individually fed into a sequential network. Afterwards, both final hidden representations are concatenated and used to directly predict the output. The original task in their work was for end-of-turn prediction. However, we implement their architecture for backchannel prediction. In this method, the loss is calculated only at the last token for each feature with a cross-entropy loss function.

4.5.3 Lexical Without EL Function

In this method, the loss was calculated as follows:

For the **positive** case:

$$L_p = \sum_{s=1}^{S} - \log(\hat{p}_s), \tag{4}$$

For the **negative** case: same as Eq. (3).

We trained the model with the available word data for five lengths of delay (0.0, 0.5, 1.0, 1.5, and 2.0 s). All models were trained independently.

Table 2 Backchannel prediction performance under different length delay conditions. Results shown are macro-averaged values across positive and negative classes. All the models were trained with a 0.0-s delay condition.

(a) Evaluation under 0.0-s delay condition

Features	Loss function	Precision	Recall	F1
Acoustic	w/o EL	56.36	**63.43**	49.26
Lexical	w/o EL	56.02	57.34	56.01
Lexical	with EL	57.19	57.20	57.17
Acoustic + POS	w/o EL	**62.98**	56.50	53.23
TASN	w/o EL	56.78	56.55	56.55
Acoustic + lexical	w/o EL	57.05	54.94	55.47
Acoustic + lexical	with EL	59.58	57.43	**58.06**

(b) Evaluation under 0.5-s delay condition

Features	Loss function	Precision	Recall	F1
Lexical	w/o EL	56.77	55.54	55.74
Lexical	with EL	56.03	56.76	55.97
Acoustic + POS	w/o EL	53.62	54.66	52.10
TASN	w/o EL	51.88	53.03	51.19
Acoustic + lexical	w/o EL	**57.65**	53.25	53.42
Acoustic + lexical	with EL	57.60	**56.93**	**57.18**

(c) Evaluation under 1.0-s delay condition

Features	Loss function	Precision	Recall	F1
Lexical	w/o EL	56.43	53.98	52.90
Lexical	with EL	55.42	55.13	55.26
Acoustic + POS	w/o EL	54.40	56.53	51.84
TASN	w/o EL	50.43	50.76	48.88
Acoustic + lexical	w/o EL	**57.95**	52.32	51.86
Acoustic + lexical	with EL	55.68	**56.71**	**55.96**

(d) Evaluation under 1.5-s delay condition

Features	Loss function	Precision	Recall	F1
Lexical	w/o EL	55.82	53.04	50.66
Lexical	with EL	55.25	54.30	**54.47**
Acoustic + POS	w/o EL	54.70	**58.93**	50.62
TASN	w/o EL	50.11	50.17	48.10
Acoustic + lexical	w/o EL	**58.87**	52.72	51.01
Acoustic + lexical	with EL	56.29	53.77	52.96

(e) Evaluation under 2.0-s delay condition

Features	Loss function	Precision	Recall	F1
Lexical	w/o EL	**58.04**	51.20	49.50
Lexical	with EL	56.30	52.79	50.02
Acoustic + POS	w/o EL	51.86	**54.92**	48.08
TASN	w/o EL	49.90	49.84	47.46
Acoustic + lexical	w/o EL	58.00	52.02	49.41
Acoustic + lexical	with EL	56.53	53.47	**52.01**

Table 3 Backchannel prediction performance using lexical features without the EL function. Each model is trained independently under various delay conditions. Results shown are macro-averaged values across positive and negative classes.

Delay condition in model training (sec)	Delay condition in evaluation (sec)	Precision	Recall	F1
0.5	0.5	56.07	59.77	**56.16**
1.0	1.0	**56.20**	**61.27**	55.42
1.5	1.5	55.45	58.86	55.31
2.0	2.0	55.62	58.75	55.77

5 Results and Discussion

The evaluation results are shown in Table 2. All the metrics shown in Table 2 (i.e., precision, recall, and F1) are macro averaged values.[3] The best result for each metric is indicated in bold. Under a 0.0-s delay condition, the combination of acoustic and lexical features with the EL function outperformed all the baselines, with an F1 score of 58.06%. The result for the TASN model was basically the same as the lexical-based model without EL (around 56%). We also evaluated the models for different lengths of the delay condition. If we assume that the acoustic extraction process does not produce any delay, the performance of the acoustic-based model should be the same under all delay conditions, with an F1 score of 49.26%. Our combination of acoustic and lexical features with the EL function achieved an F1 score of 58.06% under the 0.0-s and 57.18% under the 0.5-s delay conditions. Afterwards, its performance dropped to 52.01% under the 2.0-s delay condition, which is still a better result than the acoustic-based model. Moreover, our method is better than TASN [12] (47.46%) and the "Acoustic + POS" model [16] (48.08%) under the same delay condition. It means our method can predict backchannel with an improvement rate on the F1 score 8% better than these state-of-the-art algorithms. Based on these results, our prior detection model achieves better performance in the prediction of suitable backchannel timing than the baseline method in which the model is trained without the EL function, under all delay conditions.

Next, we compare the acoustic- and lexical-based models independently. According to the F1 score, the lexical features were the most informative features and provided a better performance than the acoustic features. We also evaluated the models using simulated test data in which only the available words were fed to the models (see a simulated example in Table 1). As shown in Table 2, whether or not the EL function is applied, F1 scores for the lexical-based model in the 0.5-s delay condition were almost the same (around 56%). On the other hand, we can see the effect of the

[3]Each metric (precision, recall, or F1) is calculated for both positive and negative classes. Each macro-averaged metric is calculated by averaging the corresponding metrics for the positive class and the negative class.

EL function when the models are evaluated in longer delay conditions (1.0–1.5 s). The F1 scores of the lexical-based model trained with the EL function in these delay conditions were still higher than 54%, while the scores of the models trained without the EL function dropped significantly. However, even when the EL function was applied, the F1 scores dropped to 50.02% under the 2.0-s delay condition. This is a reasonable result, the average utterance duration in the dataset is 1.93 s. Therefore, under the 2.0-s delay condition, the model was evaluated mostly with a [silb] word as the input feature. However, it still achieves better performance than the acoustic-based model. According to these results, our prior detection model could successfully predict the backchannel events before the final ASR outputs.

Compared to the other combined methods, our combination (acoustic and lexical) model with the EL function had a better performance even under the 2.0-s delay condition. A continuous neural network that combines all the features (e.g., the TASN architecture) works well under zero latency, but if the model is used with an available input in the delay condition, it will likely fail to predict the backchannel accurately. This is because, under the delay condition, acoustic and lexical features do not correspond to each other. Without the EL function, the model could not predict backchannel as well as it could under the 0.0-s delay condition.

Another idea to keep the performance high under a longer delay condition is by training the model with the available words. The results in Table 3 show that, even without the EL function, the model achieved an F1 score higher than 55% for all delay conditions. However, building a model for each delay condition is difficult because it depends on the spoken dialog or the ASR system. We therefore suggest combining the lexical model and acoustic model for a better prediction.

6 Conclusion

In this paper, we have proposed a prior prediction model with an EL function for backchannel detection that can mitigate the response delay problem caused by ASR. The model was evaluated under various delay lengths since the delay might differ depending on the conditions. Results showed that the proposed prior prediction model can successfully predict backchannel even when only the available words are input as features. In future work, we will implement our proposed model in a robot system and perform a subjective evaluation.

References

1. Aldeneh Z, Dimitriadis D, Provost EM (2018) Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 6159–6163 (2018)

2. Calhoun S, Carletta J, Brenier JM, Mayo N, Jurafsky D, Steedman M, Beaver D (2010) The NXT-format switchboard sorpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Lang Resources Eval 44(4):387–419
3. Chan FH, Chen YT, Xiang Y, Sun M (2016) Anticipating accidents in dashcam videos. In: Proceedings of the computer vision-asian conference on computer vision (ACCV). Springer, pp 136–153
4. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS et al (2015) The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202
5. Eyben F, Wullmer M, Schuller (2018) OpenSMILE – the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the ACM international conference on multimedia (ACM Multimedia), pp 1459–1462
6. Godfrey J, Holliman E, McDaniel J (1992) Telephone speech corpus for research and development. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 517–520
7. Hara K, Inoue K, Takanashi K, Kawahara T (2018) Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 991–995
8. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
9. Jain, A., Singh, A., Koppula, H.S., Soh, S., Saxena, A.: Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In: Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 3118–3125. IEEE (2016)
10. Kawahara T, Uesato M, Yoshino K, Takanashi K (2015) Toward adaptive generation of backchannels for attentive listening agents. In: Proceedings of the international workshop on spoken dialogue systems technology (IWSDS), pp 1–10
11. Kawahara T, Yamaguchi T, Inoue K, Takanashi K, Ward NG (2016) Prediction and generation of backchannel form for attentive listening systems. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 2890–2894
12. Masumura R, Asami T, Masataki H, Ishii R, Higashinaka R (2017) Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 1661–1665
13. Meshorer T, Heeman PA (2016) Using past speaker behavior to better predict turn transitions. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 2900–2904
14. Morency LP, de Kok I, Gratch J (2010) A probabilistic multimodal approach for predicting listener backchannels. Auton Agent Multi-Agent Syst 20(1):70–84
15. Ries K (1999) HMM and neural network based speech act detection. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 497–500
16. Roddy M, Skantze G, Harte N (2018) Investigating speech features for continuous turn-taking prediction using LSTMs. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 586–590
17. Ruede R, Müller M, Stüker S, Waibel A (2017) Enhancing backchannel prediction using word embeddings. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 879–883 (2017)
18. Ruede R, Müller M, Stüker S, Waibel A (2017) Yeah, right, uh-huh: a deep learning backchannel predictor. In: Proceedings of the international workshop on spoken dialogue systems technology (IWSDS), pp 247–258
19. Shiwa T, Kanda T, Imai M, Ishiguro H, Hagita N (2008) How quickly should communication robots respond? In: Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI), pp. 153–160 (2008)
20. Skantze G (2017) Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In: Proceedings of the annual SIGdial meeting on discourse and dialogue (SIGDIAL), pp 220–230 (2017)

21. Truong KP, Poppe R, Heylen D (2010) A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH), pp 3058–3061 (2010)
22. Ward N, Tsukahara W (2000) Prosodic features which cue back-channel responses in English and Japanese. J Pragmat 32(8):1177–1207

Towards Personalization of Spoken Dialogue System Communication Strategies

Carla Gordon, Kallirroi Georgila, Volodymyr Yanov, and David Traum

Abstract This study examines the effects of 3 conversational traits – Register, Explicitness, and Misunderstandings – on user satisfaction and the perception of specific subjective features for Virtual Home Assistant spoken dialogue systems. Eight different system profiles were created, each representing a different combination of these 3 traits. We then utilized a novel Wizard of Oz data collection tool and recruited participants who interacted with the 8 different system profiles, and then rated the systems on 7 subjective features. Surprisingly, we found that systems which made errors were preferred overall, with the statistical analysis revealing error-prone systems were rated higher than systems which made no errors for all 7 of the subjective features rated. There were also some interesting interaction effects between the 3 conversational traits, such as implicit confirmations being preferred for systems employing a "conversational" Register, while explicit confirmations were preferred for systems employing a "formal" Register, even though there was no overall main effect for Explicitness. This experimental framework offers a fine-grained approach to the evaluation of user satisfaction which looks towards the personalization of communication strategies for spoken dialogue systems.

1 Introduction

Virtual assistant dialogue systems are becoming ubiquitous, as a growing number of people interact with systems such as Apple Siri, Microsoft Cortana, and Google

C. Gordon · K. Georgila (✉) · V. Yanov · D. Traum
USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094, USA
e-mail: kgeorgila@ict.usc.edu

C. Gordon
e-mail: cgordon@ict.usc.edu

V. Yanov
e-mail: yanov@ict.usc.edu

D. Traum
e-mail: traum@ict.usc.edu

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_11

assistant on their smart phones using natural language. Additionally, Virtual Home Assistant systems (VHAs) such as Amazon Alexa and Google Home are being welcomed into increasingly more homes across the world. These systems have become more conversational and human-like than traditional task-oriented dialogue systems, featuring human voices, and special features outside the realm of virtual assistant duties, such as joke telling.

A 2018 Google study suggests that 72% of consumers who own a VHA use it as part of their daily routine, and 41% talk to their VHA as if it were a friend or another human. The study points to the users' use of pleasantries such as "please" and "thank you", and even "sorry" to illustrate the extent to which these systems are perceived as more of a companion than a machine [9]. Communication strategies can vary widely in human-human interaction depending on a number of variables such as age, and cultural or socioeconomic background. For example, Linguist Deborah Tannen has an exhaustive body of research focused on the study of gender differences alone [12–15]. The fact that human-human communication is so diverse in style and strategy, and humans are increasingly interacting with VHAs in a more naturalistic way, underscores the need to move away from a "one size fits all" model of communication strategy for dialogue systems in the VHA domain, where a significant percentage of users conceptualize interactions to be more akin to human-human communication. Indeed, previous research has shown a dichotomy between users who prefer the system to be more conversational, and those who prefer the formal approach [5].

This suggests that the future of VHA design would benefit from the ability to allow users to personalize the communication strategy of their VHA to better suit their own. To accomplish this, a finer-grained approach to the evaluation of user satisfaction in VHAs is warranted, to tease apart exactly which communicative traits and behaviors are responsible for creating the appearance of Intelligence or Naturalness, and how the interaction of these traits affects user satisfaction. Are people more willing to forgive system misunderstandings if it speaks in a more conversational register? Would a more conversational system seem more intelligent if it provided implicit confirmations instead of explicit ones? These are the kinds of questions which this study seeks to answer.

2 Related Work

Historically, much of the research on evaluating user satisfaction with dialogue systems has focused more on objectively quantifiable measures such as task completion and the length of the interaction. However, for the past few decades, the focus has shifted to the evaluation of more subjective measures. The PARADISE framework [16] is a well known dialogue evaluation framework which has been used by many researchers to optimize desired qualities such as user satisfaction [10]. Likert scale questionnaires have also been used to evaluate user satisfaction [11] as well as more complex questionnaires, such as the SASSI questionnaire [7].

Even more recently, attention has been paid to evaluating the specific subjective features which contribute to overall user satisfaction. A review of several studies which have focused on evaluating subjective user feedback on user satisfaction have revealed a set of subjective features which have frequently been mentioned by users: Intelligence, Personality, Friendliness, and Naturalness [1, 4, 8]. However, in these previous studies no attempt was made to determine what specific system behaviors give it the appearance of friendliness or intelligence.

By their very nature, it is more difficult to evaluate specific subjective features, such as those described above, than objective measures like word error rate. Researchers have used Likert scales to quantify the degree to which a dialogue system is perceived as intelligent or pleasant, but the subjective nature of these terms makes it difficult to extrapolate exactly which communicative traits and system behaviors are responsible for the user's perception of these features. It is, therefore, necessary not only to analyze which subjective features users find most agreeable, or to what extent they feel these features are present, but to discover what combination of behaviors the system exhibits that leads the user to perceive them as more satisfactory.

A recent study suggested explicit confirmations of user requests have a strong inverse correlation with the perception of Pleasantness, Naturalness, and Personality [5]. It also suggested that a more conversational register has a positive correlation with Personality, but a negative correlation with Intelligence. Additionally, another study found a predictably strong negative correlation between system misunderstandings and overall user satisfaction [2]. A more recent study focused on predicting a number of subjective features from real user dialogues and dialogue participants' ratings or dialogue observers' ratings versus simulated dialogues and dialogue observers' ratings [3]. Interestingly, this study suggested that for 3 subjective features (Intelligence, Naturalness, Overall Quality) learning prediction functions of ratings from simulated data only could be sufficient.

These previous studies informed the use of Register, Explicitness, and Misunderstandings as the set of communicative features combined to create the system profiles for the current study.

3 Experimental Design

In order to determine how each of these communicative traits interacts with the others to affect the overall user experience, as well as the perception of more specific subjective features, we developed a set of 8 system "profiles" each with a different combination of traits. Participants were recruited to interact with and rate these system profiles via a novel Wizard of Oz data collection tool. This produced a corpus of human-system dialogues with ratings for 7 subjective features such as Naturalness and Intelligence (see Sect. 3.2 for a more detailed discussion of these subjective features).

Table 1 A dialogue with a conversational system

User	(1) Turn up the volume in the bathroom speaker
System	(2) Roger that
User	(3) A little bit more, please
System	(4) Done
User	(5) And turn off the washer in the garage
System	(6) I am on it, Sir

Table 2 A dialogue with a formal system

User	(1) Hey, connect the speaker to bluetooth
System	(2) The speaker is already connected
User	(3) Please set the washer to rinsing mode
System	(4) The washing mode is now set to rinsing

3.1 Communicative Traits

In order to make analysis of these traits more manageable, each trait was treated as binary, rather than a continuum. Therefore, each system was either "conversational" or "formal" in Register, with no distinctions made as to different levels of formality.

Register refers to the presence or absence of informal, colloquial speech. Systems which were assigned a "conversational" Register would frequently use informal language such as "sure thing" or "I'll get right on that" when responding to user requests. An example of a dialogue with a conversational system can be seen in Table 1. In contrast, systems which were assigned a "formal" Register would never use this kind of informal language, as exemplified by the dialogue in Table 2.

Explicitness refers to whether or not the system gave explicit responses to user requests. Systems which were assigned to the "explicit" group always gave explicit confirmations to user requests, as can be seen in line (4) of the dialogue in Table 2. Likewise, systems which were assigned to the "implicit" group only ever gave implicit confirmations to user requests, as in line (2) of the dialogue in Table 1. Instead of the system explicitly confirming that it has turned up the volume for the bathroom speaker, it gives a general acknowledgement that the user's request has been heard, with the implication that their request will be carried out.

Misunderstandings refer to whether or not the system would make errors indicative of misunderstandings due to a failure in speech recognition or natural language understanding. Systems which were assigned to the "error" group would make at least one error per dialogue, and would recover from that error during the next dialogue turn. An example of a system error can be seen in line (2) of Table 3. Systems which were in the "no error" group would not make any such errors during their interactions with participants.

Table 3 A dialogue containing a system misunderstanding

User	(1) Connect the AC to the WiFi in the living room
System	(2) Of course! I connected the WiFi for the living room TV
User	(3) Turn WiFi off TV in the living room
System	(4) Sure, I disconnected the WiFi for the living room TV
User	(5) Turn on the AC WiFi in the living room
System	(6) I'm on it, turning on the WiFi for the AC in the living room

Table 4 System profiles and associated communicative traits

System	Register	Explicitness	Errors
Monkey	Formal	Explicit	Yes
Elephant	Formal	Explicit	No
Giraffe	Formal	Implicit	Yes
Rabbit	Formal	Implicit	No
Kangaroo	Conversational	Explicit	Yes
Raven	Conversational	Explicit	No
Squirrel	Conversational	Implicit	Yes
Lion	Conversational	Implicit	No

Profiles. These 3 communicative traits were combined to create our 8 system profiles for this study. Table 4 shows the distribution of the traits across the system profiles. Each system profile was given an animal name for reference, and was comprised of a unique combination of the 3 communicative traits. Please see the Appendix section for dialogue examples which illustrate the different communication strategies of the 8 profiles.

3.2 Subjective Features

To gauge the effect of these communicative traits on the user experience, 7 subjective features were chosen to be rated by participants:

- Intelligence
- Friendliness
- Naturalness
- Personality
- Enjoyableness
- Likelihood to Recommend
- Overall Quality

Intelligence, Friendliness, Naturalness, Personality. The first 4 features were chosen based on a review of studies designed to evaluate the subjective user experience of interacting with spoken dialogue systems. These are features that are frequently mentioned by users in subjective user feedback as having an effect on overall user satisfaction (see Sect. 2). Additionally, some studies have suggested some of these features, such as Naturalness, may be hard to tie to a specific set of system behaviors, while others like Personality and Intelligence may be at odds with each other, and maximizing one may mean sacrificing the other [5].

Enjoyableness, Likelihood to Recommend, Overall Quality. The last 3 general subjective features were chosen as a means of measuring different facets of the user experience, in order to make a more nuanced analysis of overall user satisfaction possible.

3.3 Data Collection Tool

Collecting dialogues for research in this domain is not a trivial task. The best source of data would come from live interaction between participants and a VHA, in an actual home environment where the participant can see and hear if their requests are being carried out properly. Data collected without this environmental context might be less representative of real user interaction, because it would not be clear when the system makes a mistake that does not appear in an explicit confirmation, or if other side-effects impact the dialogue. However, such an experimental setup would be logistically challenging and costly to carry out.

To solve these challenges we made use of a novel Wizard of Oz (WOz) data collection tool which seeks to emulate a real-world home setting [6]. This framework consists of a set of interconnected, web-based GUIs, one for participants and two for the Wizard.

The **User View Interface** (see Fig. 1) displays information designed to emulate a virtual home environment. The rooms and their accompanying devices are displayed in the middle of the screen. Each device displays state information such as whether it is on or off, and the settings of its various features. Changes to these settings are shown to the user in real time. At the top of the screen a task is displayed that the user must complete by communicating with a VHA using the text chat function below the device display. In the upper right hand corner the system profile is represented by a picture of an animal. In order to control for a possible confounding influence of animal preference, the animals displayed to the user were randomized by the system between participants, rather than each user seeing the same correspondence between animal and communicative traits presented in Table 4. In this way, one participant's *Monkey* system may use the communication profile of the *Elephant* system, while another participant's *Monkey* system may employ the communication profile of the *Rabbit* system.

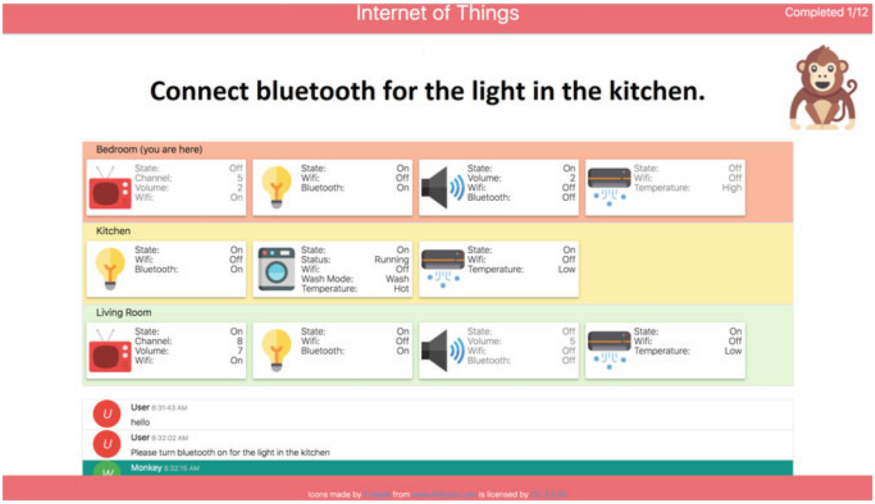


Fig. 1 The User View Interface of the WOz tool

The **Wizard View Interface** displays the same virtual home environment as the User View Interface, including updates to device settings, but with some additional information needed by the Wizard. This information includes the system profile information, as well as other information not relevant to the current study.

The **Wizard Control Interface** (see Fig. 2) allows the Wizard to control device states in the virtual home environment, and also to communicate with the user via template-based GUI buttons. A toggle menu allows the Wizard to switch between 4 different Wizard Control screens, each with a different set of utterance templates which conform to the different combinations of Register and Explicitness which comprise the 8 system profiles. The system profile information serves as a guide for the Wizard. Thus when the profile calls for "conversational" and "implicit" system behavior the Wizard will use the buttons that generate appropriate language for these system behaviors. Misunderstandings were created artificially by the Wizard, by intentionally performing actions that were not congruent with user requests.

4 Data Collection

Participants. Eighteen participants were recruited using craigslist.com. As our primary concern was the collection of system ratings data, we did not collect any personal identifying or demographic information from the participants, other than verifying that each was over the age of 18.

Interaction. Participants were seated in a quiet and distraction-free environment, engaged in IoT dialogues using the User View Interface of the WOz tool on a Mac-

Fig. 2 The Wizard Control Interface of the WOz tool

book Pro laptop. They were instructed to imagine they were in a home environment and were talking to a VHA in order to accomplish small tasks around the home. Each participant interacted with only 4 of the 8 system profiles, to ensure there was enough interaction with each profile for participants to be able to evaluate them properly.

The WOz tool generated a randomized set of 12 out of a list of 36 pre-determined tasks; 3 tasks per system profile. These tasks were displayed one at a time to the user via the User View Interface. Participants were required to communicate with the VHA via the text chat function, and command it to complete these tasks. Tasks varied based on the type and number of devices involved, the number of rooms involved, and whether the task was to be completed now, or scheduled for a future time. They also ranged from simple requests such as "Turn on the TV in the living room." to slightly more complex requests such as "In 3 min, turn off the AC in the bedroom, and turn on the TV in the living room. Set the TV channel to 7.".

Participants were able to see updates to device settings in real time as the Wizard carried out their requests. The Wizard also monitored task completion, and could prevent participants from proceeding to the next task until they had completed the current task. To achieve this, when participants clicked the "next" button, indicating their desire to move on to the next task, the Wizard would receive a notification that allowed them to reject the participant's request to continue. In this case, a pop-up message would be shown to the user stating "Please make sure the task is complete before moving on to the next task.". Finally, participants were not informed of the WOz nature of the study until after they had completed their interactions.

Questionnaires. There were 3 types of questionnaires administered to participants:

- A **Pre-interaction Questionnaire** was administered once to each participant at the beginning of their interaction. This questionnaire collected information about their overall computer literacy, familiarity with VHAs, and whether or not they own one themselves.

Table 5 Means and p-values for the main effect of Errors, (*) denotes highly significant p-values less than .001

Feature	Mean w/ Errors	Mean w/o Errors	p-value
Intelligence	5.67	5.11	.022
Friendliness	5.93	5.38	.005
Naturalness	5.75	5.42	.124
Personality	5.76	4.94	.000*
Likelihood to Recommend	5.30	4.48	.004
Enjoyableness	5.49	4.61	.001
Overall Quality	5.69	4.74	.000*

- A **Post-task Questionnaire** was administered after each task the participant completed. This questionnaire asked them to rate the system's *Intelligence*, *Friendliness*, *Naturalness*, and *Personality*, as well as how much they *Enjoyed* the interaction, how likely they are to *Recommend* the system to a friend, and the *Overall Quality* of the interaction, on a 7-point Likert scale (1:low, 7:high).
- Finally, a **Post-interaction Questionnaire** was administered at the end of each participant's interaction with the WOz tool, asking them to rank the 4 systems they interacted with from best to worst.

5 Results

To examine the effects of the 3 communicative traits on user satisfaction, a multivariate two-way analysis of variance was performed, and the results are summarized below.

5.1 Main Effects

There was a significant main effect found for Errors for all measures except Naturalness. Oddly, systems which produced errors were rated more highly than systems which did not produce errors, regardless of Register and Explicitness (see p-values and means in Table 5). This contradicts much previous research, and a deeper discussion of this anomaly can be found below in Sect. 6.

There were no statistically significant main effects found for either Register or Explicitness. However, there were a few interesting interaction effects found.

5.2 Interaction Effects

Register*Explicitness: There is a significant interaction effect on the measure of Overall Quality of the system (p = .044). An independent samples t test showed systems that were Formal and Explicit (M = 5.69) rated far better than Formal and Implicit (M = 4.9), with a p-value of .016. Conversely, Conversational systems scored higher if they were Implicit (M = 5.23) rather than Explicit (M = 5.04), although this difference was not statistically significant. Nevertheless, this suggests that if a system uses a Formal register it should give explicit confirmations, whereas if it is more Conversational implicit confirmations might be preferred.

Register*Errors: There is a significant interaction effect between Register and Errors on the measure of Personality (p = .032). For both Conversational and Formal systems, those that produced errors were rated higher (Conversational Mean = 5.9, Formal Mean = 5.63) than those that did not (Conversational Mean = 4.63, Formal Mean = 5.27), although further analysis revealed that this difference was only significant for Conversational systems (p <.001). Overall, systems that used a Conversational register and produced errors were rated most highly on the measure of Personality.

Explicitness*Errors: There is a significant interaction effect between Explicitness and Errors on the measures of Enjoyableness (p = .024) and Overall Quality (p = .007), with systems that made errors and were explicit receiving the highest ratings for both Enjoyableness (Mean with errors = 5.83, Mean no errors = 4.4, p <.001) and Overall Quality of the system (Mean with errors = 6.17, Mean no errors = 4.56, p <.001). The systems that did not make errors scored higher if they were implicit, both for Enjoyableness (Mean implicit = 4.83, Mean explicit = 4.39) and Overall Quality (Mean implicit = 4.92, Mean explicit = 4.57), although these differences were not statistically significant. This suggests that if a system is prone to errors, users would prefer it give explicit confirmation of user requests, whereas if a system makes fewer errors, implicit confirmations might be preferred.

Register*Explicitness*Errors: The interaction of all three conversational traits had a statistically significant effect on the system's perceived Friendliness (p = .05).

6 Discussion

There are some unexpected trends revealed in the ratings data. As illustrated in Table 6 and discussed in Sect. 5.1, an analysis of the ratings for Enjoyableness, Likelihood to Recommend, and Overall Quality, shows that the top 4 ranked systems are those which produced errors, and the top two are systems that were Explicit in their responses. One possible explanation for this phenomenon is that the systems were rated more favorably because they recovered quickly from their errors, and gave explicit confirmations so the user knew the errors had been addressed. As mentioned

Table 6 Enjoyableness, likelihood to recommend, and overall quality

Rank	Mean Rating	System	Register	Explicitness	Errors
1	5.83	Kangaroo	Conversational	Explicit	Yes
2	5.56	Monkey	Formal	Explicit	Yes
3	5.21	Squirrel	Conversational	Implicit	Yes
4	4.97	Giraffe	Formal	Implicit	Yes
5	4.84	Elephant	Formal	Explicit	No
6	4.83	Lion	Conversational	Implicit	No
7	4.71	Rabbit	Formal	Implicit	No
8	4.06	Raven	Conversational	Explicit	No

in Sect. 3.1, the policy was for the Wizard to recover from any errors during the next dialogue turn. This suggests that perhaps it is not a complete lack of errors, but rather the ability to recover quickly from errors that makes a system better overall.

This raises some interesting questions about what "user satisfaction" really means, and how best to evaluate it for dialogue systems in the VHA domain. The results of this study show that this sample of users prefers to interact with a VHA that makes occasional errors, as long as it recovers from them quickly, because it gives the system "personality". Indeed, the statistical analysis revealed a very strong statistical significance for errors in the ratings for Personality, with those systems that made errors (M = 5.76) averaging almost a full point higher in ratings than those that did not (M = 4.94) as can be seen in Table 5. This runs contrary to what common sense suggests, but makes more sense within the context of the 2018 Google study, cited earlier, which found almost half of people surveyed communicate with their VHA systems as if they were another human [9]. Misunderstandings are a natural part of human-human communication, so it stands to reason that a system which makes occasional errors could be seen as more human-like. This would seem to indicate that, for a significant percentage of the population, personality matters more than accuracy.

Additionally, the interaction effects suggest that there are strategies which can be employed to maximize user satisfaction based on the limitations of a certain system, or specific use case scenario. For example, if a family owns a VHA which they keep in a noisy family room and frequently misunderstands their requests, the interaction effect found between Explicitness and Errors suggests that the system would benefit from giving explicit confirmations of user requests to maximize user satisfaction. Further, as mentioned in Sect. 2, previous research has suggested a divide among tested populations between those preferring a conversational system and those preferring a formal one. If a system were to employ different dialogue modules to allow users to choose between a formal or conversational style, the interaction effect between Register and Explicitness shows that user satisfaction can be improved by utilizing implicit confirmations for the conversational module, and explicit for the formal module.

7 Conclusion

We discussed the development of a fine-grained evaluation framework for VHA dialogue systems. This approach sought to examine the main and interaction effects of Register, Explicitness, and Misunderstandings on overall user satisfaction, as well as more specific subjective features, such as Personality. A surprising trend was found in which systems that made errors were rated more favorably overall, and on all specific subjective features. This suggests that the context of interaction for VHA systems may make maximizing user satisfaction more a matter of minimizing errors, and the recovery time from them, rather than eradicating them entirely. Finally, the interaction effects showed how certain communicative traits can be combined to form conversational strategies that can maximize user satisfaction in different use cases.

Future work in this area could examine not just the presence or absence of errors, but also the amount of errors and how quickly and correctly they are resolved. It would also be interesting to look more carefully at whether different users have different preferences for combinations of system traits or perhaps whether circumstances like the complexity of the dialogues have an impact.

Acknowledgements This work was funded in part by Samsung Electronics Co., Ltd., and partly supported by the U.S. Army. Statements and opinions expressed do not necessarily reflect the policy of the United States Government, and no official endorsement should be inferred.

Appendix

The following are examples of dialogues for a single task, generated by participant interactions with each of the 8 system profiles. These examples are provided as a means of illustrating the differences in interaction between the 8 system profiles.

The Task. Users were presented with the following task: "Stop the washing machine in the kitchen and then turn it off, then turn the speaker volume to 9 in the living room."

NoError Systems. Below you will find dialogue examples for the systems which did not make errors (Table 7). These were the 4 worst performing systems overall.

Error Systems. Below you will find dialogue examples for the systems which did make errors (Tables 8 and 9). These were the 4 best performing systems overall.

Table 7 Dialogue examples for NoError systems

Raven	**Conversational*Explicit*NoErrors**
User	(1) Turn off the washing machine and turn up the living room speaker to 9
System	(2) I can't turn the washer off while it is running
User	(3) Stop washing machine
System	(4) I stopped the washer
Lion	**Conversational*Implicit*NoErrors**
User	(1) Stop the washing machine in the kitchen, then turn it off
System	(2) Sure thing! I took care of that for you
User	(3) Increase volume on living room speaker to 9
System	(4) You got it. I handled that
Elephant	**Formal*Explicit*NoErrors**
User	(1) Stop the washing machine in the kitchen and turn it off then turn the speaker volume to 9 in the living room
System	(2) I stopped the washer
System	(3) I turned the washer off
System	(4) I turned the volume to 9 for the speaker in the living room
Rabbit	**Formal*Implicit*NoErrors**
User	(1) Stop and then turn off the washing machine in the kitchen
System	(2) Ok. Done
User	(3) Turn the speaker volume to 9 in the living room
System	(4) I will do that

Table 8 Dialogue examples for Error systems: Conversational.

Kangaroo	**Conversational*Explicit*Errors**
User	(1) Turn off washing machine in the kitchen
System	(2) I can't turn the washer off while it is running
User	(3) Then stop it and turn it off
System	(4) Sure thing! I am stopping the washer. I am turning the washer off
User	(5) Living room speaker volume 9
System	(6) You got it. I am turning the volume to 8 for the speaker in the living room
User	(7) I said 9
System	(8) Sorry! I am turning the volume to 9 for the speaker in the living room
Squirrel	**Conversational*Implicit*Errors**
User	(1) Stop the kitchen washing machine, then power off
System	(2) I'm on it!
User	(3) Power off washing machine
System	(4) No problem
User	(5) Set speaker level to 9 in living room
System	(6) I'm taking care of that right now

It may not be immediately clear what the errors are for the Squirrel and Giraffe systems, since they only gave implicit confirmations of requests. The error in the Squirrel system is that the washing machine is only stopped, and not turned off, requiring the user to restate the request to turn it off in line 3 of the Squirrel dialogue in Table 8. The error for the Giraffe system is that the speaker volume was set to 8 instead of 9, as evidenced by the user restating their request in line 11 of the Giraffe dialogue in Table 9.

Table 9 Dialogue examples for Error systems: Formal

Monkey	**Formal*Explicit*Errors**
User	(1) Stop washer and turn off
System	(2) Yes. I will stop the washer. I will turn the washer off
User	(3) Living room speaker volume 9 please
System	(4) I will turn the volume to 8 for the speaker in the living room
User	(5) Volume 9
System	(6) I will turn the volume to 9 for the speaker in the living room
User	(7) Thank you
System	(8) You're welcome
Giraffe	**Formal*Implicit*Errors**
User	(1) Hello giraffe
System	(2) Hello
User	(3) Stop washing machine in kitchen
System	(4) Ok, I will do that
User	(5) Turn the washing machine in the kitchen off
System	(6) I did that
User	(7) Hello giraffe
System	(8) Hello. Is there anything else I can do for you?
User	(9) Turn up speaker volume in living to 9
System	(10) Ok. I will
User	(11) Turn up speaker volume in living room to 9
System	(12) Yes. I will do that

References

1. Artstein R, Traum D, Boberg J, Gainer A, Gratch J, Johnson E, Leuski A (2017) Listen to my body: Does making friends help influence people? In: Proceedings of FLAIRS, Florida, USA
2. Georgila K, Gordon C, Choi H, Boberg J, Jeon H, Traum D (2018) Toward low-cost automated evaluation metrics for Internet of Things dialogues. In: Proceedings of IWSDS, Singapore
3. Georgila K, Gordon C, Yanov V, Traum D (2020) Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In: Proceedings of LREC, Marseille, France
4. Geutner P, Steffens F, Manstetten D (2002) Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. In: Proceedings of LREC, Las Palmas, Spain
5. Gordon C, Georgila K, Choi H, Boberg J, Traum D (2018) Evaluating subjective feedback for Internet of Things dialogues. In: Proceedings of SemDial:AixDial, Aix-en-Provence, France
6. Gordon C, Yanov V, Traum D, Georgila K (2019) A Wizard of Oz data collection framework for Internet of Things dialogues. In: Proceedings of SemDial:LondonLogue, London, UK
7. Hone KS, Graham R (2000) Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). J Natural Lang Eng 6(3–4):287–303
8. Hurtig T (2006) A mobile multimodal dialogue system for public transportation navigation evaluated. In: Proceedings of MobileHCI, Helsinki, Finland

9. Kleinberg, S.: 5 ways voice assistance is reshaping consumer behavior - think with google (2018). URL https://www.thinkwithgoogle.com/consumer-insights/voice-assistance-consumer-experience/
10. Möller S, Ward N (2008) A framework for model-based evaluation of spoken dialog systems. In: Proceedings of SIGDIAL, Columbus, Ohio, USA
11. Paksima T, Georgila K, Moore JD (2009) Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In: Proceedings of SIGDIAL, London, UK
12. Tannen D (1990) You Just Don't Understand: Women and Men in conversation. Morrow
13. Tannen D (1993) Gender and conversational interaction. Oxford University Press, Oxford
14. Tannen D (1994) Gender and discourse. Oxford University Press, Oxford
15. Tannen D, Kendall S, Adger CT (1997) Conversational Patterns across Gender, Class, and Ethnicity: Implications for Classroom Discourse. Springer, Netherlands, pp 75–85
16. Walker M, Kamm C, Litman D (2000) Towards developing general models of usability with PARADISE. J Natural Lang Eng 6(3–4):363–377

Dialogue Systems for e-health

Authoring Negotiation Content and Programming Simulated Patients

Volha Petukhova, Firuza Sharifullaeva, and Dietrich Klakow

Abstract Competent negotiation with simulated patients can save expenses in medical education and improve outcomes for all parties involved. The use of simulated agents is beneficial for a study of human behaviour and cognition due to the possibility to create and manage a wide range of specific social situations. Building plausible cognitive models underlying the agent's intelligent behaviour from scratch is challenging and costly. Interaction designers and cognitive engineers require sufficient background knowledge to decide which domain information, resources and activities are important. Domain experts require sufficient understanding of human interaction and social cognition. All may lack advanced software development skills and an access to sufficient amount of authentic data. This paper presents a methodology to author cognitive agents and interactions with them. Authors can easily encode agents' knowledge and equip them with different sets of preferences and decision making strategies. This offers abundant opportunities for various social simulations: to create and control situations in which doctor's decision making and negotiation skills can be applied and assessed; employ and relate specific action patterns to various strategies and sociopragmatic variables of interactional power, social distance and degree of imposition; predict outcomes and explain why the choices made lead to what specific outcomes. The proposed approach also enables efficient collection of significant amount of annotated dialogue data and can be applied to model various medical and not medical negotiation scenarios.

V. Petukhova (✉) · D. Klakow
Spoken Language Systems Group, Saarland University, Saarbrücken, Germany
e-mail: v.petukhova@lsv.uni-saarland.de

D. Klakow
e-mail: dietrich.klakow@lsv.uni-saarland.de

F. Sharifullaeva
Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
e-mail: firuza.sharifullaeva@gmail.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_12

1 Introduction

Currently, large amounts of data became available to advance the state of the art in many research fields. Researchers in social disciplines, however, struggle with an issue how to aggregate appropriate data to answer their research questions. The analysis of *authentic interactions* occurred in real social environments is expected to deliver the most satisfactory degree of understanding of natural human behaviour, but researchers do not always have an access to the authentic sites and real participants to be able to collect sufficient number of instances of various phenomena. Authentic data collection in some domains is not always possible for ethical reasons and is even not always desirable due to a loose experimental control. For some use cases, restrictions need to be imposed to be able to investigate a controlled set of communicative activities and related phenomena without having to deal with unrelated details. Therefore, researchers may opt for various forms of specifically arranged interaction such as *elicited interactions*, open and closed *role plays* and *simulations*. Such data collection methods can be effective for eliciting and examining authentic interactive behaviour [3, 20]. The role-playing method is frequently used in interactive data collection efforts [7], and underpins simulations of communicative situations featuring many real-life scenarios.

In medical encounters, focus of this study, the ability to communicate can be one of the greatest assets of health care providers, but also one of their greatest liabilities [6]. Regular practice in efficient, professional and socially competent dialogue with patients often takes place in a patient-simulated setting, where Simulated Patients (SPs) are involved to portray a particular set of symptoms or roles [19, 21]. Simulation with human actors or lay persons are at rather high cost, high fidelity, but may be poorly accessible. Alternatively, artificial agents are used for training various human skills and are successfully integrated into intelligent tutoring systems and intelligent narratives [12, 33, 34]. It has been showed that even very simple agents can exhibit complex emergent behavioural patterns [15]. Advanced agents are able to produce detailed simulation of human learning, prediction, adaptation and decision making [28, 39]. They are also perfectly capable to play the role of a believable human-like agent in various human-agent settings [27, 30].

Cognitive models underlying agent's intelligent behaviour represent rather complex research objects, and despite of their acknowledged potential are not very often integrated into interactive systems. The main impediment is rooted in their ease of the authoring. Creation of plausible cognitive models from scratch is challenging, time-consuming and requires considerable multidisciplinary expertise. For instance, doctor-patient communication is an activity that, in the first place, involves an understanding of behaviour motivated by certain therapy related goal(-s) and medical task(-s). Medical experts are the only ones who have sufficient background knowledge to decide which information, resources and activities are important for which settings. Medical experts can share this information with technical experts or can better directly program agents, however lack skills to do this. Since the success of interactive processes often depends not only on the medical competence of the doctor, but also on his/her linguistic, social and cultural competences [25, 43], the design requires expertise in social interaction and cognition.

In this paper, we propose a solution which produces variable, robust and plausible agents with little efforts. We designed an authoring tool that facilitates an easy (co-)creation of multiple simulated actors for various scenarios and contexts. Agents can be built using limited interactive data: they are supplied with initial authored state-action templates encoding domain knowledge, the agent's preferences concerning issues under discussion and expected outcomes, and decision making strategies. The agent collects interactive experiences and learns from them. An example showcased in this paper demonstrates therapy planning negotiations training. However, the method can be applied in a wide range of other negotiation scenarios, e.g. negotiation of employment terms and conditions, or a mobile deal.

The paper is structured as follows. Section 2 discusses the related work performed in the authoring of dialogue systems and content. In Sect. 3, key characteristics of medical asymmetric negotiations are specified (negotiation content). Subsequently, in Sect. 4, we discuss the design aspects related to the human-agent interaction giving a global outline of a set of negotiation tasks with increasing scenario complexity and performed interactive actions, and decision making strategies (negotiation logic). Section 5 presents the authoring tool and interface to program baseline cognitive interaction agents which simulate patients exhibiting various negotiation behaviour in settings of various complexity. In Sect. 6, we present the simulated dialogue data and demonstrate how the tool can be used to author agents for new negotiation domains. Finally, we summarize our findings and outline directions for the future research and development.

2 Related Work: Authoring Dialogue Exchanges

In few past decades, many toolkits and authoring environments have been developed to build and experiment with dialogue systems—see Table 1 for an overview of the main paradigms in dialogue modelling and available authoring toolkits.

Many existing dialogue systems represent a set of possible dialogue state transitions for a given dialogue task. Dialogue states are often defined in terms of dialogue actions, e.g. question, reply, inform, and slot filling goals. Finite State Machines (FSMs) are applied to represent states and transitions, and are attractive in that they can be easily visualized, the flow is easy to understand and adjust. A toolkit to build, research and experiment with FSMs is CSLU tool [44]. Another FSM based formalism for defining complex, reactive, event-driven systems is based on statecharts [14], e.g. Deal system [8] and IrisTK system [41].

In frame-based dialogue systems, the dialogue manager extracts the necessary information from the user response and fill out the necessary slots while remembering not to ask questions for slots already filled out. Authoring a frame-based dialogue system typically involves authoring/generating or collecting (web-based) templates. CMU Communicator [36] is a toolkit to design frame-based and agenda based dialogue systems. VoiceXML [47] became the standard language used for developing interactive frame-based speech applications.

Conversationally plausible dialogue models are based on rich representations of dialogue context for flexible dialogue management, e.g. information-state updates (ISU, [10, 45]. Several ISU development environments are available, such as TrindiKit [22] and Dipper [5].

Other approaches to dialogue modelling are built as full models of rational agency accounting for planning and plan recognition. RavenClaw [4] is a dialogue architecture where designers can specify hierarchical domain tasks. ViewGen [2] is a system for modelling agents, their beliefs and their goals as part of a dialogue system, which uses a planner to simulate agents' plans.

Certain robustness has been achieved when applying statistical methods to dialogue modelling. OpenDial toolkit [26] relies on an information-state architecture where the dialogue state is represented as a Bayesian network and acts as a shared memory for all system modules. Alex Dialogue Systems Framework (ADSF, [18]) provides a modular platform for experimenting with statistical methods, e.g. based on Partially Observable Markov Decision Processes (POMDP, [49]), in the area of spoken dialogue systems. Similarly, PyDial toolkit [46] provides implementations of statistical approaches for all dialogue system modules.

Recently, deep neural networks have gained a lot of attention. PyOpenDial [16], a Python-based domain-independent, open-source toolkit for spoken dialogue systems design, re-implements OpenDial in Python and provides Python bindings to interface with popular deep learning frameworks such as Tensorflow or PyTorch, for neural dialogue state tracking and action planning.
Although the above mentioned toolkits and architectures have been successfully used for building multiple dialogue systems, using them requires considerable knowledge in the dialogue theories, expertise in software development and dialogue systems design. There have been efforts in the area of question-answering, proposing authoring tools which can be used by non-experts for rapidly building a dialogue system, e.g. NPCEditor [24]. Designers were allowed to author questions and the corresponding answers. However, the approach suffers from the inability to maintain coherence over large turn sequences. There are a dozen authoring tools to help non-experts to design dialogue exchanges with chatbots. All have graphical interfaces, and most of them require no programming. They, however, offer a rather limited set of dialogue actions, are stateless and not able to provide guarantees about content coherence. They are problematic in managing task-oriented interactions.

To author coherent dialogue interactions, methods have been proposed in the area of tutoring dialogue, interactive storytelling and games. For example, TuTalk [17] is an authoring tool which allows educational researchers to rapidly prototype dialogue systems. In digital storytelling, methodologies have been proposed to (semi-)automatically generate coherent dialogue exchanges exploiting a small base of annotated human-authored dialogue exchanges, e.g. combinatorial dialogue authoring [37]. In game design, where branching dialogue is the dominant approach to implement Non-Player Character (NPC) conversations [13], authoring interactions are largely based on scripting in-game behaviour, e.g. authoring (multi)branching dialogue as

Table 1 Toolkits and authoring environments for various dialogue modelling approaches

Dialogue modelling approach	Example task	Toolkit/Authoring environment
Finite state machines	Long-distance calling	CSLU [44]
Statecharts	Virtual receptionist	SCXML [9]; IrisTK [41]
Frame-based	Getting travel information	CMU Communicator [36]; VoiceXML [47]
Information state update	Human-robot interaction	TRINDI [22]; Dipper [5]
Plan-based	Medical diagnosis	RavenClaw [4]
Agent-based	Collaborative planning and acting	ViewGen [2]
Probabilistic approaches	Car driving assistant	OpenDial [26]
	Various information-seeking tasks	Alex DSF [18]; PyDial [46]
Neural approaches	Negotiations	PyOpenDial [16]
Chat-oriented	Retail 'chat commerce'	AIML [48]
Interactive pattern-matching	Personal assistant	Facebook: Botsify, Chatfuel, Chatsuite, etc.
Information-retrieval techniques	Question-answering	NPCEditor [24]

a tree or directed graph. The approach guarantees content coherence, but authoring becomes complex and costly; many states may not be anticipated at authoring time. The construction of conversational threads for NPCs using pattern matching and employing transition graph representations as the main interface for authoring has been proposed [42]. In [38], a fully procedural alternative to branching dialogue is presented. Rather than specifying not easy manageable directed graphs, an author composes individual lines of dialogue and annotates them with respect to the central selection policy: once a conversation turn is allocated to an NPC, an ISU-based dialogue manager requests for a line of a dialogue from the authored content that performs a targeted dialogue move or addresses a targeted topic.

Numerous studies show that participants of real-life dialogues happen to get involved in rather dynamic non-linear interactions where past and future events are out of chronological order, several parallel including disrupted or disjointed lines in the cause of events and no causality/dependency patterns between events are observed, strict directionality disappears revealing large jumps forward or backward in achieving dialogue (sub-)tasks. Good example of a non-linear interaction is negotiation, in particularly multi-issue bargaining. Negotiators may delay making complete agreements, previously reached agreements can be cancelled. Parties have the possibility to simultaneously bargain over several goods and attributes. They also may revise their past offers, accept or decline any standing offer, make counter-offers, etc. The agenda, i.e. order in which the issues are negotiated, might influence on the overall outcome. We offer a tool for authoring dialogue content and programming cognitive agents that are involved in non-linear interactions. Domain experts, e.g. medical professionals, author dialogue content and select the type of agent they would like to interact with.

3 Negotiation Content

3.1 Asymmetries in Interactions

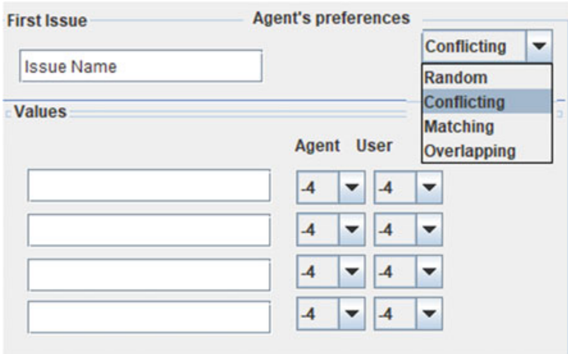
Many real-life interactive situations are characterised by a clear division of roles and an asymmetric distribution of knowledge and interactional power. e.g. at home between parents and children, in school between teachers and students, at work between supervisors and subordinates. Asymmetries are consistently observed in participant's individual attitudes towards behaviour and its outcomes. Large-scale asymmetries are detected in institutional settings where interactions are structured through institution-specific tasks and goals, which make certain institutional roles, topics, and actions available and impose constraints on others. Compared to mundane conversation, institutional discourse is often more predictable and therefore offers abundant opportunities for authoring of various social simulations.

In medical encounters, doctors exert control over the concerns expressed by their patients, and patients defer to the authority of the doctor by refraining from battling for control themselves. This may hinder patient participation in joint medical decision making as regards taking more informed decisions, but also leads to a decrease in patients' therapy adherence [50]. Therefore, the form of the interaction such as negotiation plays an important role, and doctors who show convincing persuasion and negotiation skills achieve better results for their patients [40]. In order to reach an efficient agreement, doctor should propose regimen that are feasible to follow, show an appropriate understanding of patient's desires, expectations and fears, and exercise the right influence on patient's beliefs taking patient's social, cognitive and economic constraints into account [11]. In [30], an integrative bargaining model for shared decision making in medical consultations is proposed in terms of a balancing of values as the patient's best interest and patient autonomy. The patient's best interest is modelled by taking the professional (doctor's) view on a patient's best interest. The patient's autonomy is respected based on an assessment of whether the patient is willing and able adhere to the treatment. Parties reason about the interests (preferences) of each other and negotiate the best possible *mutual agreement*.

3.2 Authoring Negotiation Profiles

Individual preferences involve participant's beliefs about perceived importance and desires concerning certain behaviour and its outcomes (attitudes) and participant's

Fig. 1 Authoring negotiation values and setting a preference profile



beliefs about his abilities to perform this behaviour (self-efficacy beliefs). Using the designed tool, preferences can be specified by an author (e.g. domain expert), human participant and/or generated automatically by the system dependent what type of partner the human participant wants to negotiate with. A graphical user interface was designed where an author can specify negotiation options, their preferences and select partner's preference profile, either identical, conflicting, matching or overlapping defined as (Fig. 1): (1) *identical*: negotiators' preferences are completely identical; (2) *conflicting*: negotiators' preferences are completely the opposite to each other; (3) *matching*: preferences are of the same polarity, but different in strength; and (4) *overlapping*: some preferences are of the same polarity and strength.

3.3 Use Case

The use case domain selected concerns diabetes. The patient-doctor negotiation scenario was designed based on the recommendations for patients who have diabetes of Type 2 of the International Diabetes Federation (IDF, 2017) addressing four issues: (1) medication, (2) diet, (3) activity and (4) exercise recommendations. Each of these issues involves four important negotiation *options* with preferences assigned representing parties negotiation positions, i.e. preference profiles. Preferences are weighted in order of importance (strength) and defined as the participant's beliefs about *attitudes* towards certain behaviour and *abilities* to perform this behaviour. The goal of each partner is to find out preferences of each other and to search for the best possible mutual agreement.

Differences in preferences result in four scenarios of various complexity. The preferences strength is communicated to the human negotiator through colours, see Fig. 2. The human participant—doctor—negotiates with various agents who simulates various types of patients, selecting one option per issue. Further, simulated patients have different preferences and are equipped with a basic set of negotiation and decision-making strategies, see next section.

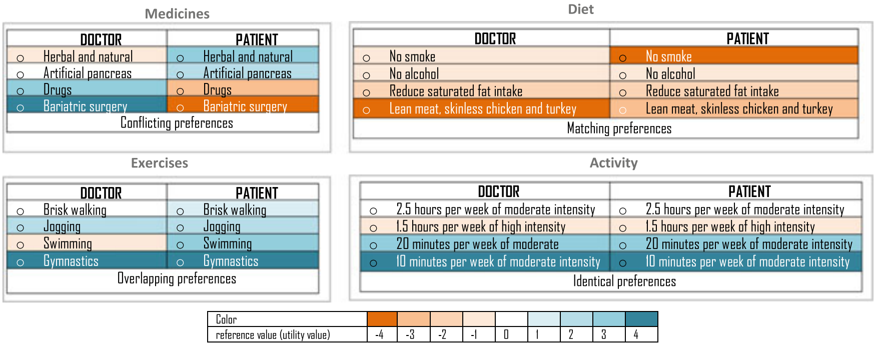


Fig. 2 Example of a participants preference profiles

4 Negotiation Logic

In integrative and problem solving negotiations, the better all possible actions and parties experiences are explored and discussed, the better agreements are reached. Interacting with our simulated patients, doctor is prompted to elicit description of preferable actions, to encourage patient to share his experiences, to match those with his professional expertise, and to adjust his behaviour accordingly. Doing this, doctors train their Theory of Mind skills [32].

4.1 Negotiation Actions

In negotiations, parties typically exchange offers expressing different levels of commitments, see [29]. Parties may propose trade-offs across issues in order for both sides to be satisfied with the outcome. Parties can give up more on one issue, but can receive in exchange for a larger share on another. They can postpone making an agreement or make a partial agreement on one issue, until the agreement on the second one is secured. They may exit agreements during the interaction and revise their past offers, accept or decline any standing offer, make counter-offers.

The successful medical negotiation involves adequate disclosure by both parties indicating their values as well as other relevant matters. It is enabled that participants express the importance, desires and abilities concerning the certain behaviour and its outcomes, i.e. global attitudes and self-efficacy assessments in their preference profiles as discussed above, and 'modalising' their reactions and counter-offers with respect to dynamically changed preferences, abilities, necessity and acquiescence performing, see Fig. 3.

In any medical interaction, many acts are produced not so much for the purpose to exchange information or influence each other's behaviour, but to establish a certain bond between the dialogue participants. Successful partnership building actions

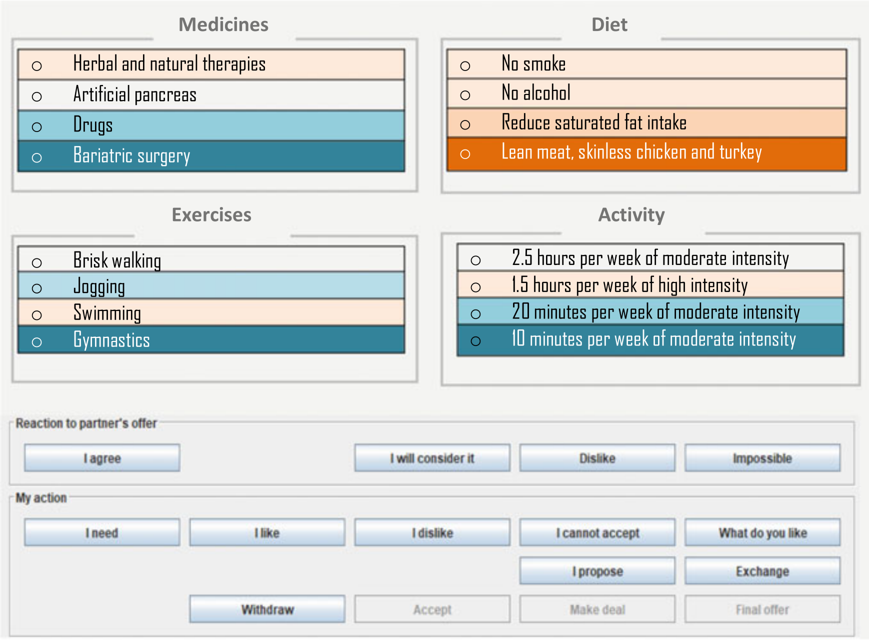


Fig. 3 Example of a preference profile for a 'diabetes' scenario and an action selection menu

promote better cooperation. Patients are often anxious about their medical condition, express concerns or worry, and seek reassurance or special attention. In the future, we will incorporate Interpersonal Relations Management acts, but also affected state and personality related aspects added to the agent's profile, based on the Roter Interaction Analysis System (RIAS, [35]), widely used for analysis of medical interaction. Table 2 provides an overview of actions used by the implemented agents and the categories proposed for future extensions (marked *).

After the agent has made a decision to perform a certain dialogue act, a corresponding linguistic pattern is selected from a database. Patterns are extracted from a comparable negotiation corpus—Multi-Issue Bargaining Corpus (MIB, [31]).[1] From the MIB corpus, lexicalised patterns tagged with communicative functions were extracted and stored as templates with the variable fields for modalised semantic content as slots values. In total, 679 communicative functions patterns were extracted and 64 slot values specified. Although an initial set of patterns was rather small, the combination of sentence patterns with the ability to change individual values, allows the generation of utterances, broader than the target corpus (5,781 MIB utterances vs 43,453 automatically generated utterances).

[1]The MIB corpus is released in the Linguistic Data Consortium catalogue under the reference number LDC2017S11 https://catalog.ldc.upenn.edu/LDC2017S11.

Table 2 Taxonomy of the agent's actions. Adapted from the ISO 24617-2 dialogue act taxonomy and the RIAS categories proposed for future extensions (*)

Interpersonal Relations Management (*)	Task	Semantic content			Global affect/ personality (*)
		Modality	Negotiation move	Issue (options)	
compliment	(open-ended) set question	preference	(final) offer	Figure 2 related to:	uncertainty
empathy	(forced) choice question	ability	exchange	therapeutic regimen	anxiety
compassion	propositional questions	necessity	concession	lifestyle recommendations	dominance
concern/worry	check questions	acquiescence	deal		attentiveness
reassure/ encourage	inform/answer		withdraw		engagement
legitimize	(dis-) agreement				friendliness
self-disclosure	suggest				anger
criticism	request/instruct				openness
	offer				

4.2 Negotiation Strategies

Negotiators apply different strategies to reach acceptable outcomes. We experimented with two basic strategies: cooperative and non-cooperative. We consider negotiators as *cooperative* if they share information about their preferences with their opponents and prefer options that have the highest shared utility. If not enough information is available to make this determination, a cooperative negotiator will elicit this information from his opponent. He will not hold on to a fixed set of preferences regardless the interests of others, instead, he will attempt to find issues where a compromise is possible. *Competitive* negotiators prefer to assert their own preferred positions rather than exploring the space of possible agreements. They ignore partner's interests and information requests, instead insist on their own ideal positions making the opponent concede. The competitive negotiator will threaten to break the negotiation if he cannot gain a significant number of points from it.

Table 3 Instance definition. Adopted from [30]

Slot	Value	Explanation
Strategy	e.g. cooperative	The strategy associated with the instance
Agent-move-value-agent	$[-4, 4]$	The number of points the agent's gets from his own move
Partner-move-value-agent	$[-4, 4]$	The number of points the partner's move brings to the agent
Partner-move-greater	$[true \mid false]$	True if the partner's move brings at least as much as the agent's one, otherwise—false
Next-move-value-agent	$[-4, 4]$	The number of points that the next best move can bring to the agent
Utility	$[0, 17]$	How valuable are the partner's suggestions made by now
Shared utility	$[0, 1]$	How valuable are the partner's suggestions for both negotiators
Agent-move	$(M_1, \ldots, M_n)$	The move that the agent should make in this context
Partner-move	$(M_1, \ldots, M_n)$	The move that the agent believes the partner should make in this context
Compensation	$[1, 4]$	If the agent's move is of the concession or exchange type, what is the minimum utility that the agent should look for choosing an alternative option

5 Simulated Patients

5.1 Agent's Knowledge and Memory

The agents are designed using the ACT-R cognitive architecture.[2] Agent's knowledge is encoded in instances (Table 3) stored in an ACT-R declarative memory which is represented as traces of experiences.

The ACT-R mechanisms account for the effects of recency—more recent memory traces are more likely to be retrieved, and frequency—if a memory trace has been created or retrieved more often in the past it has a higher likelihood of being retrieved.

[2]A Java Simulation and Development Environment for the ACT-R was used http://cog.cs.drexel.edu/act-r/about.php.

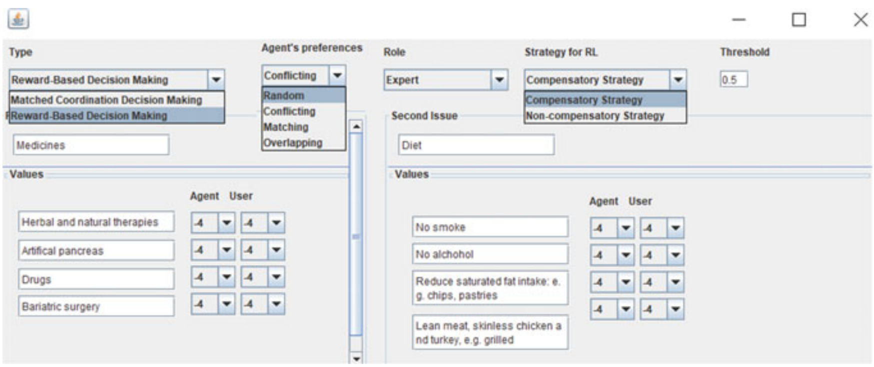


Fig. 4 Decision making strategies selection

The agent is also able to retrieve past instances even when a particular situation has not been encountered before, 'blending' is proposed as a generalization of the retrieval mechanism, allowing to retrieve values from multiple instances. An instance does not have to be a perfect match to a retrieval request to be activated. ACT-R can reduce its activation to compute partial matching, see [1, 23].

To simulate repetitive negotiations assessing doctor's behaviour over time,[3] cross-rounds agent's performance is modelled taking multi-dialogue negotiation history into account. The agent is however able to forget about his experiences, the memory decay rate can be set using the interface (0.5 set as default).

5.2 Agent's Decision Making Strategies

Authors can choose from three decision-making strategies (Fig. 4): *matching coordination*, *compensatory*, and *non-compensatory*. All three strategies are proven plausible to simulate situations where different alternatives will be selected by the agent in a certain context in order to achieve acceptable outcomes [30].

The *matching coordination* agent mirrors decision making behaviour of their opponents: the agent is competitive if it experiences the partner as competitive; it switches to a cooperative mode, if it thinks the partner is cooperative. The *compensatory* agent compensates for his losses: the agent starts with his highest offers, it continues in a competitive mode until it collects enough points (threshold is adjustable and default set at 0.5 standing for the half of the maximum possible scores) switching after to a cooperative mode; acting cooperatively, if the agent starts loosing too much, so that its utility score becomes lower than pre-defined threshold, it will switch to the competitive mode. To simulate *non-compensatory* decision-making, the agent insists on the options beneficial for it: the agent sticks to his preferred options with

[3]In real life, doctors and patient often do not meet only once, but share certain interaction history with each other.

Table 4 Comparison of human-human and human-agent negotiation performance

Evaluation criteria	Human vs human	Human vs agent
Number of dialogues	25	75
Collection time (in min/per dialogue)	9:40	3:50
Annotation time (in min/per min of dialogue)	25	0.0
Mean dialogue duration (in #turns)	23.0	21.3
Number of offers/per round	16.0	14.3
Dialogue Acts (# unique acts)	29	10
Vocabulary size (# unique tokens)	1864	517
Agreements (in %)	78.0	86.3
Pareto efficient agreement (in %)	82.4	90.3
Negative deals (in %)	21.0	34.3
Cooperativeness rate (in %)	39.0	61.9

the hope that partner concedes (position bargaining); if they are not accepted by the partner, the agent breaks the negotiation proposing the final offer playing ‘take-or-leave-it’ strategy. The time (number of moves) until the agent breaks the negotiation is configurable via GUI.

6 Data Collection and Domain Shift

Using the tool we constructed the **LICA** dialogue corpus.[4] Human-human and human-agent dialogues were collected. In human-human setting, one participant was randomly assigned the role of a doctor, the other participant the role of a patient. The tool is used to automatically generate preference profiles for scenarios of various complexity. The goal of each partner is to find out the preference of each other and to search for the best possible mutual agreement. In human-agent negotiations, each human trainee in the doctor’s role negotiated with the simulated patient (agent) who has different attitudes (preferences) and exhibit either cooperative and non-cooperative behaviour and uses different decision-making strategies. In total, we collected 25 human-human negotiations comprising about 575 speaking turns, and 75 human-agent negotiations were collected comprising 2049 turns. Table 4 summarizes the core corpora properties.

[4]LICA stands for Learning Interactive Cognitive Agents.

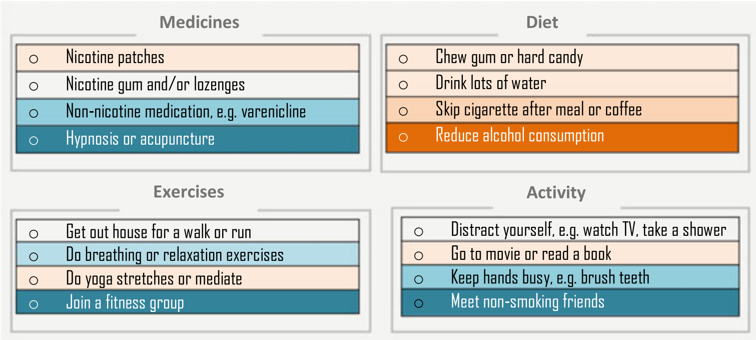


Fig. 5 Example of a set participant's preference profile and action selection menu for 'ceasing smoking' scenario

Automatically generated dialogues are not rated as highly as human-human ones: human-agent dialogues do not have as rich vocabulary as human-human ones and the agents still cannot deliver human-like interactive behaviour delivering a rather scarce repertoire of dialogue acts; human use different tactic which agent do not yet possess, namely they tend to justify most of their offers explaining why it is important to accept a certain option. Agents, however, show task-related negotiation and decision making behaviour comparable to humans in terms of the number of agreements reached and their Pareto efficiency, the number of the accepted negative deals and the cooperativeness rate. Provided with a set of agent's profile parameters and database of behavioural (linguistic) patterns, however relatively small, the tool instantly generates many full exchanges that are semantically annotated and evaluated. In Appendix, there are human-human and human-agent dialogue examples provided, see Table 5.

We envision immediate practical use of our method for a study of social cognition and for collection and exploration of behavioural and functional data. For example, dependencies between pragmalingustic forms, strategies and socio-pragmatic variability and their role for the efficient decision making process can be assessed in a systematic controlled way. Our primary observations showed that human participants facing different types of agents used different negotiation tactics which resulted in different outcomes: delayed making complete agreements; frequently revised their past offers; vary the order in which the issues are negotiated; adjust the alacrity to reveal or hide their preferences. It has been also noticed that not only asymmetries in preferences and participant's status may influence the decision making process, but participants of different gender and personality, and in different emotional state may adopt divergent strategies under identical conditions. We see that our simple agents equipped with various decision making strategies offer plenty of opportunities to investigate relationships between participant's intrinsic characteristics and various dependent variables.

To test the transferability across domains, the tool was used to encode domain knowledge for ‘stop smoking’ negotiation scenario. For this, the recommendations presented on WebMD[5] were extracted addressing the same four issues: (1) medication, (2) diet, (3) activity and (4) exercise recommendations, but involving different four negotiation *options* each, Fig. 5. The doctor’s task was then to help a patient to develop positive attitudes towards ceasing smoking, and strengthen his self-efficacy beliefs that he is capable to maintain this behaviour. Parties negotiate about what is desirable, possible and mandatory.

7 Conclusions and Future Work

In this study, we presented an approach to author simple cognitive agents that produce plausible simulations of human decision making and negotiation performance acting as believable agents in human-agent interactive learning setting. Agents are not restricted to a specific domain, but can be authored to be situated in wide range of negotiation scenarios. An author who is either a domain expert, a human negotiator/trainee, a cognitive model or dialogue system designer, can easily create various agents proving them with minimal domain knowledge, setting preferences, specifying the scenario complexity and choosing appropriate decision making strategy.

The designed baseline agents as well the authoring functionality will be extended for a comprehensive analysis and well-founded computational models of adaptive decision making behaviour in asymmetric patient-doctor interactions while accounting for the interwoven relationship between multimodal language-specific schemes and emotional, social and cultural determinants. For this, participant-specific models capturing characteristics of prototypical and personalised settings will be built. The patients’ age, gender, cultural background, personality and their level of knowledge/expertise aspects will be incorporated into the agents’ knowledge. This will allow for the creation of patient agents with different personality profiles, increasing the realism and complexity of the interaction. Further, baseline and participant-specific models will be extended to enable the agents to exhibit plausible context-aware human behaviour. For example, doctors are often situated in so called ‘unfavourable environments’ characterised by multi-tasking, time pressure, distractions, physical and psychological exhaustion. By simulating such environments, we will be able to create unique social interactive and training situations.

[5] https://www.webmd.com/.

Appendix

Table 5 Human-human and human-agent dialogue example. (D stands for doctor; HP—for human patient; and SP—for simulated patient)

ID	Human vs human	ID	Human vs agent
D1	Hallo		
D2	My name is John Smith and I am your doctor		
D3	Today I would like to discuss our treatment plan for your diabetes of Type 2		
D4	So, are you aware of the seriousness of this disease?		
HP1	No		
HP2	Tell me more		
D5	tell more …		
D6	This is very serious and if we do not act now it increases the risk of heart disease, stroke, high blood pressure and narrowing of blood vessels		
HP3	oh, no		
D7	So, I'd like to discuss a few things with you	SP1	What medicines will work the best for me?
D8	A treatment concerns medicine, diet, exercise and activities	SP2	What can you offer?
D9	For instance, I suggest we start with medicines		
D10	The most efficient therapy from my experience are drugs	D1	We could try drugs
D11	Metformin shows 99% recovery rate		
HP4	well …		

(*continued*)

Table 5 (*continued*)

ID	Human vs human	ID	Human vs agent
HP5	This sounds like a great idea. I would go for it	SP3	I agree, let's do this
D12	Moving further, I would like to talk about is your diet		
D13	We need to make several adjustments		
D14	I would propose to reduce saturated fat intake	D2	For diet, you need to reduce saturated fat intake
D15	That's the most effective strategy fighting diabetes		
HP6	It doesn't sound like a good idea	SP4	I don't like to reduce saturated fat intake
HP7	I really like all fatty stuff and it will be hard to resist		
D16	OK		
D17	But you need to make sacrifices, otherwise we will be not successful You need to do something with your diet		
HP8	OK		
HP9	What do you think if I give up smoking or alcohol, would this be enough?	SP5	I think no smoke or no alcohol is more efficient
D18	Well …		
D19	This is a good start		
D20	Maybe you can give up smoking and spend time on physical activities which can be of varied intensity	D3	OK, let's do no smoke and one and half hour per week of high intensity
D21	I am afraid that in your case there is no other option than one and half hour per week of high intensity		
HP10	well …		

(*continued*)

Table 5 (*continued*)

ID	Human vs human	ID	Human vs agent
HP11	I think it's a doable thing If I split this in a couple of days, it seems doable	SP6	That would be nice
D22	Exercising can help manage your blood sugar and weight		
D23	My personal suggestion would be brisk walking	D4	How about brisk walking?
D24	I recommend brisk walking often to my patients, it's …		
HP12	Actually I do not have any objection against it	SP7	I think I can live with that
HP13	Swimming would be a problem, but brisk walking I can do		
D25	OK …	D5	So, to summarize, we agreed on drugs, to reduce fat intake, 20 min per day of moderate intensity and brisk walking
D26	In 5 month we …		
HP14	Tell me Doctor, am I going survive? and still enjoy life with my wife and kids *laugh*	SP8	No, we agreed on drugs, to reduce fat intake, one and half hour per week of high intensity and brisk walking
D27	Of course *laugh* just follow my recommendations	D6	Right, that's a deal

References

1. Anderson J, Bothell D, Byrne M, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111(4):1036
2. Ballim A, Wilks Y (1991) Beliefs, stereotypes and dynamic agent modeling. User Model User-Adap Inter 1(1):33–65
3. Bardovi-Harlig K, Hartford BS (2005) Interlanguage pragmatics: exploring institutional talk. Routledge, Abingdon
4. Bohus D, Rudnicky AI (2009) The RavenClaw dialog management framework: architecture and systems. Comput Speech Lang 23(3):332–361

5. Bos J, Klein E, Lemon O, Oka T (2003) DIPPER: description and formalisation of an information-state update dialogue system architecture. In: Proceedings of the 4th SIGdial workshop on discourse and dialogue, pp 115–124
6. Brindley P et al (2014) Improving medical communication: skills for a complex (and multilingual) clinical world. Can Respir J 21(2):89–91
7. Brône G, Oben B (2015) InSight interaction: a multimodal and multifocal dialogue corpus. Lang Resour Eval 49(1):195–214
8. Brusk J et al (2007) DEAL: dialogue management in SCXML for believable game characters. In: Proceedings of the 2007 conference on future play. ACM, pp 137–144
9. Brusk J, Lager T (2007) Developing natural language enabled games in (Extended) SCXML. In: Proceedings from the international symposium on intelligence techniques in computer games and simulations (Pre-GAMEON-ASIA and Pre-ASTEC), Shiga, Japan, March, pp 1–3
10. Bunt H (1999) Dynamic interpretation and dialogue theory. In: Taylor M, Neel F, Bouwhuis D (eds) The structure of multimodal dialogue II. John Benjamins, Amsterdam, pp 139–166
11. Burgoon J et al (1987) Relational communication, satisfaction, compliance-gaining strategies, and compliance in communication between physicians and patients. Commun Monogr 54(3):307–324
12. de Weerd H, Broers E, Verbrugge R (2015) Savvy software agents can encourage the use of second-order theory of mind by negotiators. In: CogSci
13. Freed A (2014) Branching conversation systems and the working writer
14. Harel D (1987) Statecharts: a visual formalism for complex systems. Sci Comput Program 8(3):231–274
15. Hegselmann R, Krause U et al (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. J Artif Soc Soc Simul 5(3):1–33
16. Jang Y, Lee J, Park J, Lee K-H, Lison P, Kim K-E (2019) PyOpenDial: a python-based domain-independent toolkit for developing spoken dialogue systems with probabilistic rules. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): system demonstrations, pp 187–192
17. Jordan PW, Vanlehn K (2007) Rapidly creating a tutorial dialogue system using the TuTalk tool suite. In: Proceedings of the 2007 conference on artificial intelligence in education: building technology rich learning contexts that work. IOS Press, p 728
18. Jurčíček F, Dušek O, Plátek O, Žilka L (2014) Alex: a statistical dialogue systems framework. In: International conference on text, speech, and dialogue. Springer, pp 587–594
19. Kaplonyi J, Bowles K-A, Nestel D, Kiegaldie D, Maloney S, Haines T, Williams C (2017) Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 51(12):1209–1219
20. Kasper G (2000) Data collection in pragmatics research. In: Culturally speaking: managing rapport through talk across cultures, pp 316–341
21. Ker J, Bradley P (2013) Simulation in medical education. In: Understanding medical education: evidence, theory and practice, pp 175–192
22. Larsson S, Traum D (2000) Information state and dialogue management in the Trindi dialogue move engine toolkit. Nat Lan Eng 6(3–4):323–340
23. Lebiere C, Wallach D, West R.L. (2000) A memory-based account of the prisoner's dilemma and other 2x2 games. In: Proceedings of international conference on cognitive modeling, pp 185–193
24. Leuski A, Traum D (2011) NPCEditor: creating virtual human dialogue using information retrieval techniques. AI Mag 32(2):42–56
25. Lindemann K (2015) Emotionen in medizinischer Kommunikation. Handbuch Sprache in der Medizin 11:154
26. Lison P, Kennington C (2016) OpenDial: a toolkit for developing spoken dialogue systems with probabilistic rules. In: Proceedings of ACL 2016 system demonstrations, pp 67–72
27. Malchanau A, Petukhova V, Bunt H (2019) Towards integration of cognitive models in dialogue management: designing the virtual negotiation coach application. Dialogue Discourse 9(2):35–79

28. Marewski J, Link D (2014) Strategy selection: an introduction to the modeling challenge. Wiley Interdiscip Rev Cogn Sci 5(1):39–59
29. Petukhova V, Bunt H, Malchanau A (2017) Computing negotiation update semantics in multi-issue bargaining dialogues. In: Proceedings of the SemDial 2017 (SaarDial) workshop on the semantics and pragmatics of dialogue, Saarbrücken, Germany
30. Petukhova V, Sharifullaeva F, Klakow D (2019) Modelling shared decision making in medical negotiations: interactive training with cognitive agents. In: International conference on principles and practice of multi-agent systems. Springer, pp 251–270
31. Petukhova V, Stevens CA, de Weerd H, Taatgen N, Cnossen F, Malchanau A (2016) Modelling multi-issue bargaining dialogues: data collection, annotation design and corpus. In: LREC
32. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(04):515–526
33. Riedl M, Stern A (2006) Believable agents and intelligent story adaptation for interactive storytelling. In: International conference on technologies for interactive digital storytelling and entertainment. Springer, pp 1–12
34. Ritter S, Anderson J, Koedinger K, Corbett A (2007) Cognitive tutor: applied research in mathematics education. Psychon Bull Rev 14(2):249–255
35. Roter D, Larson S (2002) The roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. Patient Educ Couns 46(4):243–251
36. Rudnicky AI, Bennett C, Black AW, Chotomongcol A, Lenzo K, Oh A, Singh R (2000) Task and domain specific modelling in the Carnegie Mellon Communicator system. Technical report, CMU, Pittsburgh PA School of Computer Science
37. Ryan JO, Barackman C, Kontje N, Owen-Milner T, Walker MA, Mateas M, Wardrip-Fruin N (2014) Combinatorial dialogue authoring. In: International conference on interactive digital storytelling. Springer, pp 13–24
38. Ryan JO, Mateas M, Wardrip-Fruin N (2016) A lightweight videogame dialogue manager. In: DiGRA/FDG
39. Salvucci D, Taatgen N (2008) Threaded cognition: an integrated theory of concurrent multitasking. Psychol Rev 115(1):101
40. Sandman L (2009) The concept of negotiation in shared decision making. Health Care Anal 17(3):236–243
41. Skantze G, Al Moubayed S (2012) IrisTKfs: a statechart-based toolkit for multi-party face-to-face interaction. In: Proceedings of the 14th ACM international conference on Multimodal interaction. ACM, pp 69–76
42. Spierling U, Weiß SA, Müller W (2006) Towards accessible authoring tools for interactive storytelling. In: International conference on technologies for interactive digital storytelling and entertainment. Springer, pp 169–180
43. Suchman AL, Markakis K, Beckman HB, Frankel R (1997) A model of empathic communication in the medical interview. JAMA 277(8):678–682
44. Sutton S, Cole R (1997) The CSLU toolkit: rapid prototyping of spoken language systems. In: Proceedings of the 10th annual ACM symposium on User interface software and technology. ACM, pp 85–86
45. Traum D, Bos J, Cooper R, Larsson S, Lewin I, Matheson C, Poesio M (1999) A model of dialogue moves and information state revision. TRINDI project deliverable D2.1
46. Ultes S, Rojas Barahona LM, Su P-H, Vandyke D, Kim D, Casanueva I, Budzianowski P, Mrkšić N, Wen T-H, Gasic M et al (2017) Pydial: a multi-domain statistical dialogue system toolkit. In: Proceedings of ACL 2017, system demonstrations, pp 73–78
47. VoicaXML (2007) VoiceXML specification
48. Wallace R (2003) The elements of AIML style. Alice AI Foundation, 139
49. Williams J, Young S (2007) Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang 21(2):393–422
50. Wynia MK, Osborn CY (2010) Health literacy and communication quality in health care organizations. J Health Commun 15(S2):102–115

A Spanish Corpus for Talking to the Elderly

Raquel Justo, Leila Ben Letaifa, Javier Mikel Olaso, Asier López-Zorrilla, Mikel Develasco, Alain Vázquez, and M. Inés Torres

Abstract In this work, a Spanish corpus that was developed, within the EMPATHIC project (http://www.empathic-project.eu/) framework, is presented. It was designed for building a dialogue system capable of talking to elderly people and promoting healthy habits, through a coaching model. The corpus, that comprises audio, video an text channels, was acquired by using a Wizard of Oz strategy. It was annotated in terms of different labels according to the different models that are needed in a dialogue system, including an emotion based annotation that will be used to generate empathetic system reactions. The annotation at different levels along with the employed procedure are described and analysed.

1 Introduction

Although the use of conversational systems in our daily life seemed to be science fiction not much time ago. Nowadays they are pretty integrated in our homes (Alexa

R. Justo (✉) · L. B. Letaifa · J. M. Olaso · A. López-Zorrilla · M. Develasco · A. Vázquez · M. I. Torres
Universidad del Pais Vasco UPV/EHU, Sarriena s/n, 48940 Leioa, Spain
e-mail: raquel.justo@ehu.eus

L. B. Letaifa
e-mail: leila.benletaifa@ehu.eus

J. M. Olaso
e-mail: javiermikel.olaso@ehu.eus

A. López-Zorrilla
e-mail: asier.lopezz@ehu.eus

M. Develasco
e-mail: mikel.develasco@ehu.eus

A. Vázquez
e-mail: alain.vazquez@ehu.eus

M. I. Torres
e-mail: manes.torres@ehu.eus

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_13

speaker by Amazon), jobs (Cortana or Siri to manage our agenda) or even in our leisure (Siri or Samsung's Bixby for smartphones). They are becoming useful in more and more different domains ranging from game environments to educational contexts. Some of them can pass the Turing test (e.g., Eugene Goostman[1]). Thus, we can say that the way in which people interact with computers is shifting to the use of natural language.

There are many different systems in the literature built for different purposes and that make use of different technologies [7, 21, 23, 26]. However, one of the most extended categorization of conversational systems is the one that distinguishes among "chatbots" and "dialogue systems" [9, 12, 18]. Although the frontiers among those categories are not always clear, focusing on the differences related to the goal, chatbots are aimed at generating appropriate, relevant, meaningful and human-like utterances and there is not an specific goal to be achieved during the interaction like in the case of dialogue systems. Dialogue systems are often developed for a specific domain, whereas simulated conversational systems [chatbots] are aimed at open domain conversation [13].

In this work we deal with a dialogue system developed within the EMPATHIC project [11, 24, 25] framework. The goal of this project is to design and validate new interaction paradigms for personalized Virtual Coaches to promote healthy and independent aging. Thus, a dialogue system that can talk to the elderly, understand them, empathise with them and promote healthy habits is being developed. This kind of dialogue systems need different modules like automatic speech recognizer, natural language understanding module, dialogue manager, natural language generator, etc. Moreover, a module that tries to detect the emotion of the speaker is also being developed in order to provide a system response that can be empathetic with regard to the user emotional status. The methodologies employed to develop these modules are mainly based on machine learning and data driven approaches. When using these approaches, data are needed to be able to train robust models. Moreover, the data have to be closely related to the specific task, environment, channel, etc. Thus, it is very difficult to get valuable resources when specific tasks, like the one presented in this framework are considered. Furthermore the lack of resources is even more noticeable when we consider other languages (apart from English) like Spanish.

The main contribution of this paper is the description of a Spanish corpus devoted to train different models that will be employed in a dialogue system that tries to talk to the elderly people and promotes healthy habits being aware of the affective component. The corpus was annotated in terms of different labels that will be used by the different modules. The annotation procedures, that will be described in the following sections, were selected to allow the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. Section 2 provides a description of the dialogues that comprise the corpus and the way in which they were acquired. In Sect. 3 the annotation procedure developed to build the modules related to dialogue generation are described (natural language understanding, dialog manager and

[1] http://www.reading.ac.uk/news-archive/press-releases/pr583836.html.

natural language generation). Then in Sect. 4 the annotation carried out to detect emotions in different channels (audio, video and text) is detailed.

2 Dialogues in the EMPATHIC Framework

In order to develop a dialogue system, like the one described above, a data acquisition procedure has to be designed first. In this process we used a Wizard of Oz (WoZ) platform [19, 20] for the acquisition of the database. The WoZ constitutes a prototyping method that uses a human operator (the so-called wizard) to simulate non- or only partly- existing system functions. It was used to make users think that they are interacting with a real automatic dialogue system. In this way, the data acquisition procedure considers human-machine conversations that were carried out in an environment as most realistic as possible.

The dialogues in the EMPATHIC project are leaded according to a coaching model, a GROW coaching model in this case, that tries to get the desired goals related to healthy habits. A GROW coaching dialogue consists on four main phases: Goals or objectives, Reality, Options and Will or action plan. During the first phase, the dialogue aims at establishing explicit objectives that the user wants to achieve, e.g. reduce the amount of salt. During the next phase, taking into account the user's personal context, the dialogue tries to detect potential obstacles that prevent fulfilling the previously established objectives. For the next phase, the goal is to analyse the options the user has in order to face the obstacles and achieve the objectives. In the last phase, the dialogue tries to specify an action plan for the user to carry out in order to advance towards the objectives. The final goal for the EMPATHIC virtual coach is to deal with four different domains: leisure, nutrition, physical activity and social and family relationships [17]. However, in the initial phase described in this paper, not all the scenarios were used; two scenarios were integrated in this platform. A first introductory scenario, which in turn was used to strengthen the user in the interaction with the platform. And a second one to simulate a GROW session on nutrition. These scenarios were designed using the documentation provided by a professional coach. Although different acquisition procedures were carried out in the project for different languages: Spanish, French and Norwegian, in this work we focus on the Spanish dataset.

Making use of the aforementioned WoZ platform, 79 native Spanish users selected among the target population (healthy elderly above 65) interacted with the system. The majority of them participated in the two predefined scenarios, but in some cases, due to different reasons, only one of these sessions was carried out. Thus, 142 dialogues were collected. These include around 4,500 user turns and the same amount of machine turns.

The acquired conversational sessions between elderly people and the simulated virtual coach were recorded in order to have an audio-visual database. Each session takes about 10 min so the total recordings correspond to about 23 h of video. The audio part represents about 30% of the database.

3 Resources for Building the Dialogue

Once the acquisition procedure was finished the data were annotated in order to build the different models involved in the conversational process.

3.1 Speech to Text Annotation

One of the first annotation needed for training robust models to be used in a dialogue system is the transcription of the speech. This is essential for the Automatic Speech Recognizer for instance. Thus after the acquisition procedure, the dialogues were manually transcribed. The vocabulary size resulted to be 5,543 for the user turns and 2,941 for the virtual coach's turns. As for the running words, the corpus contains 72,350 in the user turns and 30,389 in the he virtual coach's turns.

The transcriptions of the acquired dialogues were further annotated in order to facilitate the modeling of the dialogues. The following two sections explain how the turns of both the users and the virtual coach were labeled. The two annotation tasks were carried out by 9 annotators, who were instructed about the structure of the labels, the GROW coaching model, and about the context of the project. Each annotator labeled roughly the same number of dialogues.

3.2 Semantic Annotation

The taxonomy of the labels used to represent each of the users' utterances was designed so as to be usable for the dialogue agent that is being deployed in the EMPATHIC project. Several works have addressed the question of defining dialogue act taxonomies [3, 22]. Among them, the DIT++ taxonomy [2] and the more recent ISO 24617-2 standard [4, 15], which is intended to be a development of the previous one, can be considered the general methodological framework of the taxonomy defined in this section. It is a dialogue-act taxonomy aimed to represent the user utterances in a particular human-machine communication framework, which develops a coaching model aimed at keeping a healthy and independent life of elderly. Thus, the taxonomy allows the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. To fulfill its needs, we employed three different types of labels: the topic, the intent and the name entities. The topic label identifies the general context of utterance, such as nutrition, leisure or family; and also some subtopics. The intent label determines the communicative intention of the user, e.g. greetings, agreement, opinion and so on. Additionally, it also includes information about which stage of the GROW model the user is talking about. Finally, the name entities are tuples containing small segments of the utterance and their category.

Table 1 Most frequent topic, intent and entity labels in the corpus

Topic	Intent	Entities
sport & leisure - travelling	*generic - agreement*	*actions*
sport & leisure - hobbies - type	*GROW - habit - present*	*quantities*
nutrition - regularity - ordered	*generic - opinion - positive*	*places*
sport & leisure - motivation	*generic - disagreement*	*amount of time*
nutrition - quantity	*generic - greetings*	*frequencies*
sport & leisure - music	*GROW - plan*	*hobbies*

They can be very useful for understanding the user but also for enriching the natural language generator. We have included, for example, people's names, places, and books.

The topic and intent labels are hierarchical, i.e., each utterance is labeled with multiple tags that can be ordered from more general to more specific. To make the annotation more consistent, each turn was split into several subsentences if there were more than one topic or intents in that turn. In total, 56 different labels were used for the topic representation, 34 for the communicative intent and 22 types of entities were identified. The complete list of labels is provided in detail in [14]. Since it is too large, Table 1 shows the most frequent labels for the topic and intent, and the most frequent entities.

3.3 Dialogue Act Annotation

Dialogue Act (DA) annotation is the equivalent task to the semantic annotation for the turns of the virtual coach. In this case, the outputs of the coach are labelled considering five criteria: DA, polarity, gender of the user and coach and possible appearance of entities in the responses of the coach. Such annotation is highly related to the Natural Language Generation (NLG), one of the modules included in the dialogue system developed in the EMPATHIC project. The NLG is in charge of generating the responses of the virtual coach to the users through a unit of information which contains a set of labels. The inverse process is made in the annotation: one set of labels is assigned to each turn of the virtual coach contained in the data.

The data was extracted from two different sources: the WoZ sessions and a set of handmade dialogues prepared by a professional coach. In both cases, only the turns of the coach were relevant to build this part of the data. Indeed, each turn can be split in different utterances, where an utterance is considered each element which can be labelled with a different DA. In total, the number of utterances is 8173 where 5985 are from the real session with users and 2188 from the handcrafted conversations.

All these utterances were labelled in terms of the five aforementioned criteria. The DA, which is built for one principal label and sublabel in the case of EMPATHIC, describes the communicative function and the semantic of the coach's sentences. There are 10 different values for the principal label and more than 100 for the sublabel. However, the DAs do not allow all the possible combinations, as each label only can be joined with a reduced group of sublabels. The polarity defines the emotional state of the coach, which can be selected between positive and neutral. The possible values for the genders are male, female and not identifiable, since what is annotated is if the gender of the two participants can be known through the coach turn alone (without any context). Finally, the detection of entities followed the same procedure carried out in the semantic annotation.

In the DAs, we identified three different blocks with the following distribution: the GROW block (19.6%), the Introduction one (24.6%) and General one (55.8%). The first block contains eight of the ten principal labels. These labels are the eight typical questions used in the GROW model. The other blocks, each one only contains one principal label. The Introduction label is used to annotate usual turns in a first session with the user (ask for the name, self-introduction, information of the project, ...). Finally, the General one is used to label all the expression which can be part of any conversation (thanking, greetings, agreement, ...). In terms of the polarity, the positive utterances (63.0%) were almost two times the neutral ones (37.0%). For both user and coach gender, they were not identifiable in almost the 99% of the data. Finally, the most frequent entities in the data were actions, dates and food.

4 Resources for Empathizing with the Elderly

Within the EMPATHIC project framework, the idea of empathising was very important. Thus, we wanted not only to understand what elderly is requesting to the system, but also to know their emotional status when interacting with it. Therefore, an annotation in terms of emotion was carried out by Spanish native annotators. The representation of emotional status is not straightforward and different models can be used according to Affective Computing literature [1, 5, 6, 16]. In this work we employed both a categorical model and a three-dimensional VAD (Valence, Arousal and Dominance) model in order to be able to compare both criteria.

Both data modalities, audio and video, were considered. In order to avoid interference between modalities, only audio (i.e. no images) was provided to the speech annotators and only video (without sound) was used by the video annotators.

In this section, we describe and analyse each modality annotation. For more information about the annotation procedure, refer to [8] and [10].

Finally, at the same time as the semantic annotation was carried out, the polarity of the transcribed utterances was also labeled by the same annotators.

Table 2 Audio annotated segments

	Calm	Sad	Happy	Puzzled	Tense
Annotation 1	7017	17	260	347	12
Annotation 2	7794	19	291	297	24
Annotation 3	7655	21	244	360	20

Table 3 Video annotated segments

	Sad	Annoyed	Surprised	Happy	Pensive	Other	Neutral
Annotation1	0	0	12	234	2032	0	2278
Annotation2	0	1	44	151	2059	3	2258

4.1 Audio Annotation

Only the audio part of the conversations between the virtual coach and elder people (which duration is about 7 h) is concerned by the audio annotation process. A manual labeling procedure from scratch was carried out by three native people. The perceived emotion was labelled in terms of categorical labels and the three-dimensional VAD model labels. The labels assigned to the dimensional VAD model were:

- Valence: Positive, Neither positive nor negative, Negative
- Arousal: Excited, Slightly excited, Neutral
- Dominance: Dominant, Neither dominant nor intimidated, Defensive

The categorical labels were Calm, Sad, Happy, Puzzled and Tense. For each emotion label, the number of segments labeled by each annotator is reported in Table 2. “Calm” is the most frequent label. “Happy” and “Puzzled” are less present but “Sad” and “Tense” are quite absent.

Dealing with the duration of emotion labels, “Calm” occurs in 94% of the audio database size which correspond to more than 6 h. “Happy” and “Puzzled” labels are present in only 4% of the database with respective durations of 9 and 8 min. The negative emotions “Sad” and “Tense” have a total duration lower than one second. This could indicate that the dialog system is user friendly.

4.2 Video Annotation

For the video annotation, all the database is involved and the recordings were labeled by two native people. Six video emotion categories were selected: Sad, Annoyed/Angry, Surprised, Happy/Amused, Pensive and Other. The label Other is assigned to segments containing different emotions that the sub-mentioned ones or including simultaneous emotions. The non annotated parts are considered neutral. For each emotion label, the number of segments labeled by each annotator is reported in Table 3.

With respectively 0, 1 and 3 occurrences, “Sad”, “Annoyed” and “Other” are almost absent. “Pensive” and “Neutral” represent the most frequent labels. Indeed, as more than 14 h are not labeled, the content of the database is mainly neutral. The participants are annotated “Pensive” within a duration of about 2 h. Finally they are sometimes happy or amused (during 5–10 min).

4.3 Text Annotation

Emotions were not only labeled from audio and video (Sects. 4.1 and 4.2) but also from text, that is from the manual transcriptions achieved in Sect. 3.1. It was carried out along with the semantic annotation (Sect. 3.2) providing an emotional annotation for each transcribed utterance.

Although the audio and video has richer annotations, the text annotation includes a very significant one related to polarity labels on a scale of three values: negative, neutral and positive. This might be very useful to be combined with the audio annotation in terms of the VAD model. Specifically, when combining of Valence (audio) and Polarity (text) labels we can get the same annotation for different channels.

Looking at the annotated set it can be concluded that Neutral is the most common polarity, representing the 66.24% of the corpus, then a positive behaviour can be analysed, with a 27.21% of the corpus, and finally, negative polarity is almost absent (with 6.55% of occurrences).

5 Concluding Remarks

In this work a Spanish corpus devoted to the development of a dialogue system, oriented to promoting healthy habits among elderly is presented. The corpus was annotated in terms of different labels in order to obtain robust models for generating coherent system responses according to a coaching model. Moreover, an emotion-based annotation is also provided in order to detect emotional status of the speakers and provide a response adapted to it. The procedure carried out to obtain the annotations along with the obtained results is described.

References

1. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59. https://doi.org/10.1016/0005-7916(94)90063-9. http://www.sciencedirect.com/science/article/pii/0005791694900639
2. Bunt H (2009) The DIT++ taxonomy for functional dialogue markup. In: AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, pp 13–24

3. Bunt H, Alexandersson J, Choe JW, Fang AC, Hasida K, Petukhova V, Popescu-Belis A, Traum DR (2012) ISO 24617-2: a semantically-based standard for dialogue annotation. In: LREC, pp 430–437
4. Bunt H, Petukhova V, Malchanau A, Wijnhoven K, Fang A (2016) The dialogbank. In: Chair NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France
5. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1(1):18–37. https://doi.org/10.1109/T-AFFC.2010.1
6. Calvo RA, Mac Kim S (2013) Emotions in text: dimensional and categorical models. Comput Intell 29(3):527–543. https://doi.org/10.1111/j.1467-8640.2012.00456.x
7. Gao J, Galley M, Li L (2019) Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots
8. Justo R, Letaifa LB, Palmero C, Gonzalez-Fraile E, Johansen A, Vazquez A, Cordasco G, Schlogl S, Fernandez-Ruanova B, Silva M, Escalera S, Velasco MD, Tenorio-Laranga J, Esposito A, Kornes M, Torres M (2020) Analysis of the interaction between elderly people and a simulated virtual coach. Amb Intell Humanized Comput
9. Klüwer T (2011) From Chatbots to Dialog Systems. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. https://doi.org/10.4018/978-1-60960-617-6.ch001
10. Ben Letaifa L, Raquel JTMI (2020) Adding dimensional features for emotion recognition on speech. In: International Conference on advanced technologies for signal and image processing, Tunisia, pp 109–114
11. López Zorrilla A, Velasco Vázquez MD, Irastorza J, Olaso Fernández JM, Justo Blanco R, Torres Barañano MI (2018) Empathic: empathic, expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. Procesamiento del Lenguaje Natural
12. Masche J, Le NT (2018) A review of technologies for conversational systems. In: Advanced Computational Methods for Knowledge Engineering. Springer, Cham, pp 212–225. https://doi.org/10.1007/978-3-319-61911-8_19
13. Mctear M (2004) Spoken Dialogue Technology - Toward the Conversational User Interface. Springer, Heidelberg
14. Montenegro C, López Zorrilla A, Mikel Olaso J, Santana R, Justo R, Lozano JA, Torres MI (2019) A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. Multimodal Technol Interact 3(3):52
15. Petukhova V, Bunt H (2012) The coding and annotation of multimodal dialogue acts. In: LREC, pp 430–437
16. Russell J (2003) Core affect and the psychological construction of emotion. Psychol Rev 110:145–172. https://doi.org/10.1037/0033-295X.110.1.145
17. Sayas S (2018) Dialogues on Leisure and Free Time, Dialogues on Physical Exercise, Dialogues on Nutrition. Technical Report DP1, DP2, DP3, Empathic Project; Internal Documents: Tampere, Finland
18. Scerri D, Dingli A (2013) Dialog systems and their inputs. In: Stephanidis C (ed) HCI International 2013 - Posters' Extended Abstracts. Springer, Heidelberg, pp 601–605. https://doi.org/10.1007/978-3-642-39476-8_121
19. Schlögl S, Doherty G, Karamanis N, Luz S (2010) Webwoz: a wizard of Oz prototyping framework. In: Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing System (EICS 2010), pp 109–114. https://doi.org/10.1145/1822018.1822035
20. Schlögl S, Milhorat P, Chollet G, Boudy J (2014) Designing language technology applications: a wizard of Oz driven prototyping framework. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, pp 85–88. https://doi.org/10.3115/v1/E14-2022

21. Serban I, Sordoni A, Bengio Y, Courville AC, Pineau J (2015) Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI
22. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M (2000) Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist 26(3):339–373
23. Su PH, Gašić M, Young S (2018) Reward estimation for dialogue policy optimisation. Comput Speech Lang 51:24–43. https://doi.org/10.1016/j.csl.2018.02.003
24. Torres MI, Olaso JM, Glackin N, Justo R, Chollet G (2019) A spoken dialogue system for the empathic virtual coach. In: D'Haro LF, Banchs RE, Li H (eds) 9th International Workshop on Spoken Dialogue System Technology, Singapore, pp 259–265
25. Torres MI, Olaso JM, Montenegro C, Santana R, Vázquez A, Justo R, Lozano JA, Schlögl S, Chollet G, Dugan N, Irvine M, Glackin N, Pickard C, Esposito A, Cordasco G, Troncone A, Petrovska-Delacretaz D, Mtibaa A, Hmani MA, Korsnes MS, Martinussen LJ, Escalera S, Cantariño CP, Deroo, O, Gordeeva O, Tenorio-Laranga J, Gonzalez-Fraile E, Fernandez-Ruanova B, Gonzalez-Pinto A (2019) The empathic project: mid-term achievements. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '19. ACM, New York, pp 629–638. https://doi.org/10.1145/3316782.3322764
26. Williams JD, Asadi K, Zweig G (2017) Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: ACL

Analysis of Prosodic Features During Cognitive Load in Patients with Depression

Carmen Martínez, Spyridon Kontaxis, Mar Posadas-de Miguel, Esther García, Sara Siddi, Jordi Aguiló, Josep Maria Haro, Concepción de la Cámara, Raquel Bailón, and Alfonso Ortega

Abstract Major Depressive Disorder (MDD) is a largely extended mental health disorder commonly associated with a hesitant and monotonous speech. This study analyses a speech corpus from a database acquired on 40 MDD patients and 40 matched controls (CT). During the recordings, individuals experienced different levels of cognitive stress when performing Stroop color test that includes three tasks with increasingly level of difficulty. Speech features based on the fundamental frequency (F0), and the speech ratio (SR), which measures the speech to silence ratio, are used for characterising depressive mood and stress responsiveness. Results show that SR is significantly lower in MDD subjects compared to healthy controls for all the tasks, decreasing as the difficulty of the cognitive tasks, and thus the stress level, increases. Moreover F0 related parameters (median and interquartile range) show higher values within the same subject in tasks with increased difficulty level for both groups. It can be concluded that speech features could be used for characterising depressive mood and assessing different levels of stress.

C. Martínez (✉)
University of Zaragoza, Campus Río Ebro, C/María de Luna 1, 50018 Zaragoza, Spain
e-mail: martineza@unizar.es

S. Kontaxis · R. Bailón · A. Ortega
University of Zaragoza, Campus Río Ebro, C/María de Luna 1, 50018 Zaragoza, Spain
e-mail: sikontax@unizar.es; rbailon@unizar.es; ortega@unizar.es

M. Posadas-de Miguel · C. de la Cámara
Hospital Clínico de Zaragoza, Zaragoza, Spain
e-mail: marposadasdemi@yahoo.es; conchidlc@hotmail.com

E. García · J. Aguiló
Autonomous University of Barcelona, Barcelona, Spain
e-mail: esther.garcia@ciber-bbn.es; jordi.aguilo@uab.es

S. Siddi · J. M. Haro
Parc Sanitari Sant Joan de Déu, Barcelona, Spain
e-mail: sara.siddi@pssjd.org; jmharo@pssjd.org

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_14

1 Introduction

Major Depressive Disorder (MDD) is the most prevalent mental illness and the fifth leading cause of global disability in the world [1]. Depression is associated with loss of interest, tiredness, and sleep disruptions [2]. Its diagnosis is currently based on clinical interviews and questionnaires according to Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [3], such as Hamilton Depression Rating Scale (HDRS) or Beck Depression Inventory (BDI). However, these methods may be dependent on the predisposition of the patients to talk about their symptoms.

MDD has been linked to an autonomic nervous system (ANS) dysfunction, which in turn leads to changes in psychomotor activity [4]. As a result, speech in MDD patients has been described as slower, monotonous, paused or even hesitant [5], which can guide to the idea of using speech analysis to characterise depressive mood in a non-invasive and comfortable way.

In fact, speech signals have already been used for diagnosing depression, but different results were obtained [6]. In [7, 8], a significant negative trend between fundamental frequency (F0) and depression severity (measured through HDRS or BDI) was reported. However in [9–12], F0 yielded to no significant results for discriminating depression state. Besides frequency parameters, speech ratio-related features have also been studied in depressed speech. In [8], higher depression severity scores were associated with slower speech, while in [10, 13], longer pauses were found in depression patients.

In the present work, a set of prosodic features are used for discriminating between MDD patients and control (CT) subjects. These parameters will be analysed while subjects perform a Stroop test [14], which implies an homogeneous condition for speech but a cognitive load and stressful situation for ANS that can highlight its dysfunction in MDD.

2 Materials and Methods

2.1 Experimental Protocol

A database of 40 MDD patients (24 women, age 47.33 ± 13.07 years, Body Mass Index (BMI) 27.85 ± 5.63 Kg/m^2), was recorded at the Hospital Clínico Lozano Blesa (Zaragoza) and the Parc Sanitari Sant Joan de Déu network of mental health services (Barcelona). MDD group consists of subjects with Major Depression Disorder according with Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria [3] and the HDRS scale (HDRS > 9). Recordings of 40 CT subjects without clinical history of mental disorders, matched by sex, BMI and age, were also acquired in order to ensure that group differences are attributable to differences in depression status and not due to unbalanced demographic data.

The experimental protocol consists in an exposure to a cognitive stress named Stroop test [14]. The test was adapted to the mother tongue of the subjects, i.e.

Spanish, and it is formed of three parts lasting about 45 sec each. In the first task (T1) the subjects should read the words 'red', 'green', and 'blue' printed in black ink. In the second one (T2), the words are substituted by a set of 'XXX's written in the previous colours, and the participants should say out loud the colour of the ink. Finally, in the third test (T3), the words 'red', 'green', and 'blue' are coloured in inks that do not match the true meaning of the text, so the subject should say the colour of the ink and not the written word. Note that the difficulty of the tasks increases gradually from T1 to T3, being the colours and the words in the first level (T1) congruent, and incongruent in the second and third (T2, T3).

Speech signals were acquired using an AKG CK-80 microphone and a Tascam us-122L recording device at a sampling frequency of 44.1 kHz.

2.2 Speech Processing

At the beginning, speech signals were preprocessed, using Audacity® and FFMPEG. Noisy segments, e.g. coughs or throat cleanings, as well as parts of the recording in which the interviewer interfered with the subject, e.g. correcting errors or providing details related to task, were removed.

Prosodic parameters are obtained in order to analyse differences between depressive and control speech. First, F0 is estimated each 10 msec using the robust algorithm for pitch tracking (RAPT) [15] included in the openSMILE toolkit [16]. Then, median and interquartile range of F0 ($F0_m$ and $F0_{iqr}$) during each task are computed as prosodic features. However, these values can be highly variant among different subjects and thus, the normalised differences $\widetilde{F}0_m$, $\widetilde{F}0_{iqr}$ for T2 and T3 with respect to T1 will be studied for measuring response to increasing stress level in both populations.

Moreover, speech ratio (SR), which measures the proportion between the time that the subject is speaking and the total duration of the recording, is computed using a voice activity detector algorithm (VAD) [17] based on Long-Term Spectral Divergence (LTSD).

2.3 Statistical Analysis

Feature set consists of previously mentioned parameters from speech analysis i.e., $\widetilde{F}0_m$, $\widetilde{F}0_{iqr}$, and SR. Unpaired tests between MDD and CT populations are carried out to study the effect of depression on speech, while paired tests are conducted to study the differences within the same subject on stress response at tasks of different difficulty. Moreover, prosodic indexes are analysed taking into account the differences in the frequency ranges of each gender [18], generating four study groups. Student t-tests or Wilcoxon tests are implemented depending on the distribution of the data (Shapiro-Wilk test), Gaussian or not, for comparing means or medians of the distribution, respectively. In this study, the level of statistical significance is set to $p = 0.05$.

3 Results

Results show that $\widetilde{F}0_m$ (Fig. 1(a–b)) and $\widetilde{F}0_{iqr}$ (Fig. 1(c–d)) do not differ significantly between CT and MDD population (unpaired tests) for both genders. Regarding paired analysis, $\widetilde{F}0_m$ and $\widetilde{F}0_{iqr}$ show increasing trends in almost every group due to the increment of difficulty at T3. Only $\widetilde{F}0_m$ in male subjects with MDD do not exhibit significant differences as the cognitive stress raises.

Moreover, SR values, shown in Fig. 2, were found to be significantly lower in every Stroop task in MDD with respect to CT groups. Note that, in both groups, SR values decrease significantly as the difficulty of the cognitive tasks, and thus the stress level, increases, i.e., from T1 to T3.

4 Discussion

In the present work, an analysis in depressive subjects and paired controls has been conducted by means of speech features for studying differences in stress responsiveness. Stroop Test has been previously used in healthy subjects for measuring cognitive load, i.e., mentally stressful situations, from spoken speech [19–21], showing that vocal frequencies can correlate with cognitive load and be used for its classification.

Results in Fig. 1 show that MDD and healthy subjects exhibit higher values in frequency-related features during T3, thereby suggesting that both groups reacted to stress induction. The values of $\widetilde{F}0_m$ and $\widetilde{F}0_{iqr}$ in females (Fig. 1 (a, c)) increase significantly in both groups as the difficulty of the tasks increases, while in males

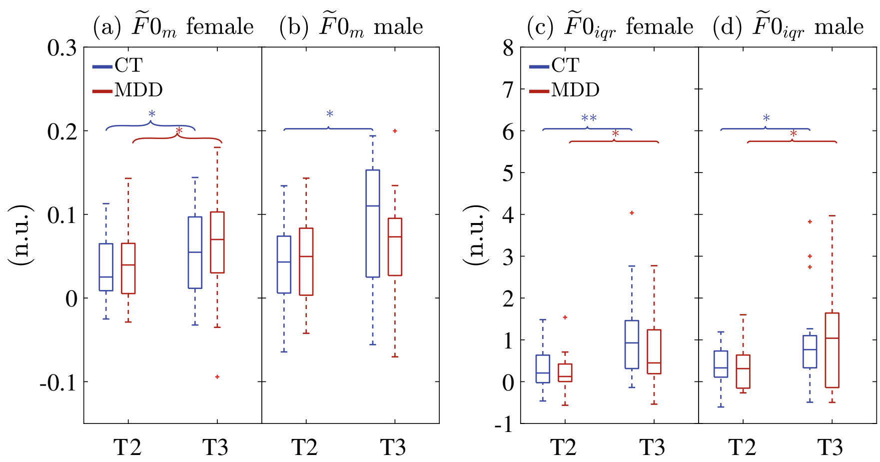


Fig. 1 Boxplot of prosodic features. (a) $\widetilde{F}0_m$ in female subjects, (b) $\widetilde{F}0_m$ in male subjects, (c) $\widetilde{F}0_{iqr}$ in female subjects, (d) $\widetilde{F}0_{iqr}$ in male subjects. CT and MDD groups are marked in blue and red, respectively, while statistically significant differences of paired analysis are marked with one or two colored asterisks for $p < 0.05$ and $p < 0.001$, respectively.

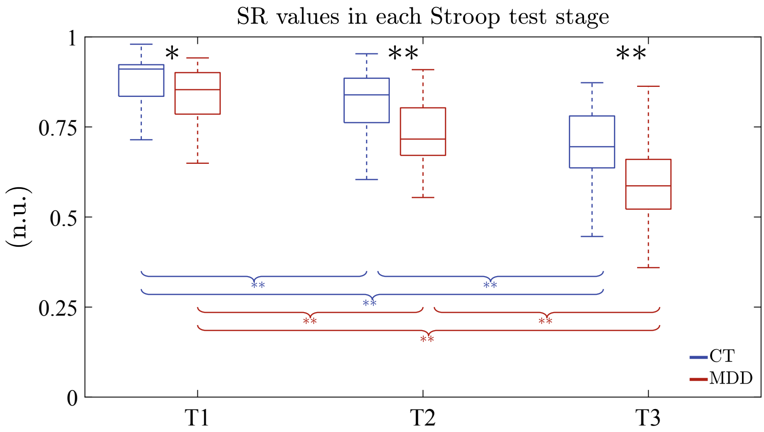


Fig. 2 Boxplot of Speech Ratio (SR). CT and MDD populations are marked in blue and red, respectively. Statistically significant differences for $p < 0.05$ and $p < 0.001$ are marked with one and two asterisks, respectively. Black asterisk correspond to unpaired analysis and colored to paired analysis.

(Fig. 1 (b, d)) only controls show higher values. The absence of significant differences in male subjects with MDD could be either because they do not respond to stressful stimuli, or due to the small number of male subjects (16 out of 40). In [7, 8], a significant negative trend for F0-related features was reported. However, most studies analysed these parameters using non-spontaneous speech, e.g. oral reading [6, 22], instead of performing a cognitive test. Thus, the absence of significant differences in F0-related parameters between MDD and CT group in this study, might be attributed to the stress induced by Stroop Test.

Moreover, SR is the most promising parameter as it shows significant differences between MDD and CT group in all tasks. Results show a decreasing trend with the increment of stress level in both groups. Note that SR speech parameter values are always lower in MDD group, which might be related to cognitive dysfunction in MDD patients [23]. The ability to inhibit cognitive interference in the Stroop Test has been used for measuring cognitive functions such as attention and processing speed among others [24]. Videbech et al. [25], reported that patients with depression had a greater difficulty when inhibiting interference compared to CT subjects, thus leading to lower values of SR in a fixed amount of time. A lower performance of MDD compared to CT subjects, measured as the time required for accomplishing a cognitive task, was reported in [26] for a subset of the present database.

5 Future Work

This study consist in a preliminary work that highlights the importance of the stress response for the monitoring and diagnosis of ambulatory MDD patients. In fact, using speech analysis, depressive mood could be assessed in a non-invasive and comfortable way. Moreover, a joint analysis of additional parameters, such as jitter or shimmer, with other physiological signals, such as HRV features, can be conducted using different classifiers.

6 Conclusion

A preliminar study about speech analysis during a cognitive stress in a database recorded on MDD patients and matched controls has been presented. The analysis of speech parameters during Stroop Test tasks has revealed significantly decreased speech ratio values in MDD patients with respect to matched controls and differences in fundamental frequency related parameters within the same subject among different tasks. Thus, it can be concluded that the analysis of speech can be used for the objective diagnosis of MDD patients in a non-invasive, straightforward, and comfortable way. In conclusion, SR, not only because of its simplicity but also due to its robustness to inter-session variations due to recording conditions, can be thought as a suitable feature for distinguishing between populations in non-controlled environments, as it has shown the best performance among the analysed features.

Acknowledgements This work has been supported by AEI and FEDER under the projects RTI2018-097723-B-I00 and 2014–2020 "Building Europe from Aragón", by CIBER de Bioingeniería, Biomateriales y Nanomedicina, and CIBERSAM, through Instituto de Salud Carlos III, by LMP44-18, BSICoS group (T39-20R), ViVoLab group (T36-20R) and a personal grant to S. Kontaxis funded by Gobierno de Aragón; and by Spanish Ministry of Economy and Competitiveness and the European Social Fund (TIN2017-85854-C4-1-R). The computation was performed by the ICTS 'NANBIOSIS', more specifically by the High Performance Computing Unit of the CIBER in Bioengineering, Biomaterials & Nanomedicne (CIBERBBN).

References

1. Vos T (2017) Global, regional, and national incidence, prevalence, and years lived with disability for 328 disease and injuries, 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390:1211–1259
2. World Health Organisation (WHO) (2011) Depression: let's talk. In: Website of World Health Association. Disorders Management, Depression. https://www.who.int/news-room/detail/30-03-2017--depression-let-s-talk-says-who-as-depression-tops-list-of-causes-of-ill-health. Accessed Jan 2020
3. American Psychiatric Association (1994) Diagnosis and Statistical Manual of Mental Disorders (DSM). 4th edn. Washington DC
4. Sperry SH, Kwapil TR, Eddington KM et al (2018) Psychopathology, everyday behaviours, and autonomic activity in daily life: An ambulatory impedance cardiography study of depression, anxiety, and hypomaniac traits. Int J Psychophysiol 129:67–75
5. Kräpelin E (1921) Manic-depressive insanity and paranoia, 2nd edn. Livingstone, Edinburgh
6. Cummins N, Scherer S, Krajewski J et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49

7. Hönig F et al (2014) Automatic modelling of depressed speech: relevant features and relevance of gender. In: 15th Proceedings of Interspeech, Singapore, 14–18 September 2014
8. Cannizzaro M, Harel B, Reilly N et al (2004) Automatic modelling of depressed speech: voice acoustical measurement of the severity of major depression. Brain Cogn 56:30–35
9. France DJ, Shiavi RG, Silverman S et al (2000) Acoustical properties of speech as indicator of depression and suicidal risk. IEEE T Bio Med Eng 47:309–319
10. Mundt JC, Snyder PJ, Cannizzaro MS et al (2007) Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguist 20:50–64
11. Taguchi T, Tachikawa H, Nemoto K, Suzuki M et al (2017) Major depressive disorder discrimination using vocal acoustic features. J Affect Disorders 225:214–220
12. Quatieri TF et al (2012) Vocal-source biomarkers for depression: a link to psychomotor activity. In: 13th Proceedings of Interspeech, Portland, OR, USA, 9–13 September 2012
13. Mundt JC, Vogel AP, Feltner DE et al (2012) Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 72:580–587
14. Stroop JR (1992) Studies of interference in serial verbal reactions. J Exp Psychol 121:15–23
15. Resch B, Nilsson M, Ekman A et al (2007) Estimation of the Instantaneous Pitch of Speech. IEEE T Audio Speech 15:813–822
16. Eyben F, Wöllmer M, Schuller B (2010) openSMILE - the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia, Firenze, Italy, 25–29 October 2010
17. Ramírez J, Górriz JM, Segura JC (2007) Voice activity detection. Fundamentals and speech recognition system robustness. In: Grimm M, Kroschel K (eds) Robust speech recognition and understanding. InTech
18. Klatt DH, Klatt LC (1990) Analysis, synthesis and perception of voice quality variations among female and male talkers. J Acoust Soc Am 87:820–857
19. Schuller B et al (2014) The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. In: 15th Proceedings of Interspeech, Singapore, 14–18 September 2014
20. Yin B et al (2008) Speech-based cognitive load monitoring system. In: 2008 IEEE international conference on acoustics, speech, and signal processing, Las Vegas, NV, USA, 31 March–4 April 2008
21. Yap TF, Epps J, Ambikairajah E et al (2001) Formant frequencies under cognitive load: effects and classification. EURASIP J Adv Sig Pr
22. Williamson JR et al (2014) Vocal and facial biomarkers of depression based on motor incoordination and timing. In: AVEC 2014 Proceedings of the 4th international workshop on audio/visual emotion challenge, Orlando, Florida, USA, November 2014
23. Lam RW, Kennedy SH, McIntyre RS et al (2014) Cognitive dysfunction in major depressive disorder: effects on psychosocial functioning and implications for treatment. Can J Psychiatry 59:614–654
24. Scarpina F, Tagini S (2017) The stroop color and word test. Front Psychol 8:557
25. Videbech P, Ravnkilde B, Gammelgaard L et al (2014) The danish PET/depression project: performance on Stroop's test linked to white matter lesions in the brain. Psychiatry Res 130:117–130
26. Kontaxis S, Orini M, Gil E, Posadas-de Miguel M, Bernal ML, Aguiló J, de la Cámara C, Laguna P, Bailón R (2018) Heart rate variability analysis guided by respiration in major depressive disorder. In: 45th International conference of computing in cardiology, Maastricht, The Netherlands, 23–26 September 2018

Co-creating Requirements and Assessing End-User Acceptability of a Voice-Based Chatbot to Support Mental Health: A Thematic Analysis of a Living Lab Workshop

Antonio Benítez-Guijarro, Raymond Bond, Frederick Booth, Zoraida Callejas, Edel Ennis, Anna Esposito, Matthias Kraus, Gavin McConvey, Michael McTear, Maurice Mulvenna, Courtney Potts, Louisa Pragst, Robin Turkington, Nicolas Wagner, and Huiru Zheng

A. Benítez-Guijarro · Z. Callejas (✉)
University of Granada, Periodista Daniel Saucedo Aranda sn, Granada, Spain
e-mail: ajbenitez@ugr.es, zoraida@ugr.es

R. Bond · F. Booth · E. Ennis · M. McTear · M. Mulvenna · C. Potts · R. Turkington · H. Zheng
Ulster University, Shore Road, Newtownabbey, Northern Ireland
e-mail: rb.bond@ulster.ac.uk

F. Booth
e-mail: Booth-F@ulster.ac.uk

E. Ennis
e-mail: e.ennis@ulster.ac.uk

M. McTear
e-mail: mf.mctear@ulster.ac.uk

M. Mulvenna
e-mail: md.mulvenna@ulster.ac.uk

C. Potts
e-mail: c.potts@ulster.ac.uk

R. Turkington
e-mail: turkington-R@ulster.ac.uk

H. Zheng
e-mail: h.zheng@ulster.ac.uk

A. Esposito
Università degli Studi della Campania Luigi Vanvitelli, Viale Ellittico 31, Caserta, Italy
e-mail: iiass.annaesp@tin.it

M. Kraus · L. Pragst · N. Wagner
University of Ulm, Albert-Einstein-Allee 43, Ulm, Germany
e-mail: matthias.kraus@uni-ulm.de, louisa.pragst@uni-ulm.de, nicolas.wagner@uni-ulm.de

G. McConvey
Action Mental Health, 27 Jubilee Road, Newtownards, Northern Ireland
e-mail: gmcconvey@amh.org.uk

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_15

Abstract Mental health and mental wellbeing have become an important factor to many citizens navigating their way through their environment and in the work place. New technology solutions such as chatbots are potential channels for supporting and coaching users to maintain a good state of mental wellbeing. Chatbots have the added value of providing social conversations and coaching 24/7 outside from conventional mental health services. However, little is known about the acceptability and user led requirements of this technology. This paper uses a living lab approach to elicit requirements, opinions and attitudes towards the use of chatbots for supporting mental health. The data collected was acquired from people living with anxiety or mild depression in a workshop setting. The audio of the workshop was recorded and a thematic analysis was carried out. The results are the co-created functional requirements and a number of use case scenarios that can be of interest to guide future development of chatbots in the mental health domain.

1 Introduction

Health should be considered in terms of mental wellbeing as well as the absence of mental disorders, with a healthcare focus on the understanding, promotion and protection of wellbeing [15]. With treatment, many people do experience recovery following an episode of mental illness. However, others may be described as "living with" mental ill health, and managing their own symptoms using coping strategies and their own support networks together with the formal support services and therapies such as counselling and medication. Thus, identification of relapse signs and symptoms of decline are important.

The use of digital health interventions, whether it is mobile or web-based applications, have the potential to improve access to mental healthcare, providing opportunities to engage and empower patients/users in treating mental ill health. As demand increases for mental health care, to the point where it exceeds available resources of national health care services, there is potential for digital health interventions to reduce waiting times to be seen by healthcare consultants/therapists through offering flexible and immediate means of accessing therapy or help [6].

One form of a digital intervention for mental ill health support and coaching are conversational systems, to which we will refer to as chatbots[1]. Chatbots can provide additional empowerment and support in the self-management of mental ill health by symptom checking at various times of the day and night, as well as rapid and appropriate interventions e.g. signposting to resources, appropriate feedback such as prompts, supportive messages, coping strategies and symptom summaries. Chatbots are also intuitive to interact with since they utilise natural language, which is the most natural form of human communication [10]. It is important to note that chatbots

[1] There are slight differences between these terms, specially in their academic use as explained in [10], but we use the term 'chatbot' because it is more familiar to the end user and thus facilitates co-creation approaches.

should not be considered full “virtual therapists”, but have many advantages and can provide support to augment traditional services. In fact, current evidence shows there is a potential for conversational agents in psychiatric treatment [14].

There are a number of available chatbot systems which have been designed for monitoring of mental ill health symptoms, management of symptoms and guiding users through established therapy methods. For example, *Woebot* is a text-based chatbot technology which provides mood tracking and guides users through Cognitive Behavioural Therapy (CBT), provides exercises and stories to help facilitate improvement of users’ mental wellbeing. Results showed that symptoms of depression and anxiety had reduced significantly over the course of the study period amongst Woebot users [3]. Another example of an existing text-based chatbot system is *Wysa*, which was co-designed by therapists, coaches, users and experts and comprises CBT, Dialetical Behavioural Therapy, mindfulness and various other behavioural reinforcement tools to help users build emotional resilience skills [7]. In addition, the study with the chatbot *Vincent* [9] found that caring for a chatbot (i.e. when the chatbot is the care receiver instead of the caregiver) enhances self-compassion.

While most existing chatbot technologies for mental health purposes operate on a text basis [1], there is interest in developing chatbot technology for mental health that incorporates the use of audio analysis which can infer mood, sentiment, affect or mental illness from speech data. There are particular verbal and non-verbal (prosodic/supra-segmental) speech features which are considered robust markers for depressive states and can allow learning algorithms to discriminate between depressed and healthy people using speech [2]. Based on current evidence, automated classification of speech could be used as an objective measure of depression and other mental illnesses and therefore help improve diagnosis accuracy.

The work described in this paper is part of the H2020 MSCA RISE project MENHIR: *Mental health monitoring through interactive conversations* (ref. no. 823907, https://menhir-project.eu). MENHIR consists of a consortium of eight partners, which includes five academic institutions, two companies and a non-profit organisation. The technology that will be studied and developed in MENHIR is designed together with persons with mental ill health, experts and caregivers. The underlying principle is that nobody knows better about their needs and preferences than users, which is why they play an influential role in the MENHIR project through co-creation activities.

Co-creation refers to involving stakeholders throughout the design. This can include a series of home visits, focus groups, workshops, in-depth interviews, cultural probes and technology biographies. Co-creation provides valuable information for understanding the user needs and collecting requirements. The aim of this paper is to assess the acceptability and co-create requirements for a voice-based chatbot. We have used thematic analysis of data collected at a user-centred workshop involving clients of Action Mental Health, an organisation in Northern Ireland which assists people with mental ill health.

2 Methods

2.1 Study Protocol

As indicated in [11], living labs are "a collection of people, equipment, services and technology to provide a test platform for research and experiments" that may be employed for a number of purposes including sharing ideas, engagement with user communities and other stakeholders, co-creation, data collection and evaluation.

Co-creation can be defined as the collaborative knowledge generation by academics working alongside other stakeholders [4]. As indicated in [8], co-creation allows creating partnerships between researchers and the people for whom the research is ultimately meant to be of use, in our case people with anxiety or mild depression. The authors underline that not using this type of participatory approaches delays the implementation of effective practices, in this case in mental healthcare, which "affects people's health and contributes to the unsustainability of the health system".

This paper outlines the outcomes of a co-creation workshop using the living-lab methodology which involved the potential users of the chatbot. This piece of research involved a workshop to elicit opinions and co-create requirements for a speech-based mental health chatbot service. The workshop was carried out with Action Mental Health (AMH) clients from New Horizons Newtownards on the 27th June 2019.

AMH is an organisation that assists people with mental ill health in order to offer them services and mechanisms that allow them to manage their conditions. AMH has several centers spread across Northern Ireland, each of them offering different activities and services. AMH clients are often referred from clinical mental health centers with different reports and personal information (reports about their diagnosis, mental and personal clinical history, risk factors, traumas, critical background and other sensitive data).

Ethical approval for this workshop was granted by Ulster University and ethics is managed by MENHIR's Ethics Board. The workshop session had a duration of approximately 2 hours. The participants included 9 clients from AMH New Horizon Newtownards, a key worker from AMH New Horizon Newtownards and 3 researchers from the MENHIR project who presented the project and acted as facilitators. All 9 clients gave their consent for the group discussion to be recorded.

From the different models of co-creation described in [4], we used two: experience based co-design and technology co-design, as the goal of the sessions was two-fold: to use the clients' experience as the starting point to design the role of the chatbot, and to develop a technology that is acceptable, fits a purpose and is tailored to its users' capabilities and what matters to them.

The workshop had the following structure:

1. Welcome and brief introduction to the project.
2. Demonstration of a mental health chatbot and smart speaker by workshop facilitator.
3. Completion of consent form and study proforma.
4. Group discussion:
    - Strengths and limitations of this type of chatbot.
    - Who would use this chatbot?
    - Which type of people would the chatbot suit?
    - Could it be used in conjunction with AMH mental health support services? (how?)
    - What are the pros and cons of using chatbot in the context of mental ill health?
    - What features should a mental health chatbot have?
    - What kind of conversations would you want to have with a chatbot?
    - What kind of persona/s should the chatbot have?

To facilitate the interaction, at the beginning of the group discussion, the participants were given green and pink post-it notes where they could write the benefits (green) and limitations (pink) of the technology. The audio of the group discussion was recorded and then interpreted using thematic analysis.

3 Context of Action Mental Health (AMH)

AMH clients are usually at a medium-low risk level. These clients attend a local AMH center and are interviewed. During this interview the workers of the center explain the types of services offered in the center and ask the clients about their hobbies, goals, schedules and other personal information to understand the activities and therapies that may be of interest.

With this information, they generate an action plan to which the client commits. With the action plan, the user identifies their long term goals and set dates to achieve them. The means for achieving the objectives is to engage in vocation and personal development activities.

Once the action plan is established, the centers will offer clients up to two years of assistance. During this period, patients attend the AMH center several times per week to carry out the activities planned (e.g. gardening, handcrafting, computer science, cooking, etc.).

Although the action plan is a concept developed by AMH, it could be easily used in other settings and can be of great interest not only for the chatbot technology developed in MENHIR, but for general use, as it provides a means to ground motivational feedback in the user objectives and hobbies and structured means of achieving them by means of scheduled activities.

4 Results of the Thematic Analysis

There were 4 main topics of discussion within the workshop, (1) challenges faced by people with ill mental health, (2) chatbot functionalities, (3) chatbot characteristics, and (4) use of the chatbot.

The strengths and limitations of chatbots for mental health as reported by the participants in the post-it notes are presented in Table 1, where the statements corresponding to the same person are marked with the same number.

The group discussion was recorded resulting in a 1 h and 18 min audio recording that was then analyzed using thematic analysis. Thematic analysis is one of the most commonly used methods of analysis in qualitative research and focuses on identifying themes within transcribed data, usually resulting from interviews and focus groups [5]. The topics and themes identified have been listed in Table 2. In total, there were 17 themes identified from the topics of the group discussion.

Table 1 Result of the strengths vs. limitations activity reproduced verbatim

Strengths	Limitations
1. Akin to "talking to the dog"	1. Ability to be exploited
1. Anonymous and secure	1. Flaws where chatbots have sent recordings to contacts
1. Avoidance of stigma	
1. Potential to recognize poor mental health when the individual can't	
1. Voice recognition for security	
1. Potential for voice modulation detection (distress, anxious, etc.)	
1. Ability to know when to "intervene" (e.g. emergency)	
2. As a companion when there is nobody to talk to	2. Individual privacy and wider cryptographic protection
2. Potentially provide an answer to a situation	2. An establishment of absolute trust
3. Trusting Alexa making sure you are talking to people with proper identification	3. Conversation cannot flow 3. Making sure you have privacy
4. Someone/thing to talk to in confidence. It is easier at times to talk to someone/thing anonymously. Don't like to off load to family and friends. They don't always understand	
5. Support on hand	5. Robotic voice
5. Unbiased	5. Privacy
6. Can provide reminders to take medication, attend appointments or engage in helpful activities e.g. "have your remembered to medicate today?"	6. The user receives responses without being sure where their responses are being taken from, are they from a reliable source?
7. Takes at the necessity of relying on flawed human memory the chatbot could give automatic reminders for things like doctor's appointments and the need to take medication and even things as simple as a reminder for some self-care e.g. to eat a meal or take a shower	7. A chatbot cannot be the same as genuine one to one human contact. It is unrealistic to expect technology to be able to understand the subtle nuances of emotion, feelings or the intricacies of speech
8. Support and understanding from people that are or have been in the same situation	8. Can it run 24/7? and if we come across an issue can we report it?
9. Helping someone in need of conversation or advice if nothing or no-one is available.	9. Would it report to authorities to prevent suicide or self harm - or harm to others
9. Some people don't like to share their emotions in a group	

Table 2 Topics and themes identified from the workshop

Topics	Themes
Challenges faced by people with mental ill health	1. Isolation
	2. Difficulty for honest disclosure
Chatbot functionalities	3. Symptom recognition
	4. Continuous monitoring
	5. Disclosure facilitation
	6. Companionship/active listening
	7. Risk detection
Chatbot characteristics	8. Personalization
	9. Configurable proactiveness
	10. User access to their information
	11. Explainability
	12. Privacy
	13. Vulnerability
Use of the chatbot	14. Cost of the chatbot
	15. Access to the chatbot
	16. Intention to use of the chatbot
	17. Use of the chatbot as a complement to therapy

5 Functional Requirements

Based on the thematic analysis and the literature, a number of requirements have been discovered to deliver chatbot mental health technologies. These requirements have been outlined as follows.

Chatbots should have mechanisms for allowing to collect different types of data, with the user's permission. Permissions such as recording of the user's voice, collecting user demographics, information on their daily habits and experiences. Chatbots should also allow users to decide if the system can interpret their current state and if the system should gather information about emergency contacts and whether to permit the system to save the data to train a user model to be used for posterior interactions. Chatbots should also provide mechanisms that allow users to decide what information can be shared with their key workers (i.e. counsellors, psychiatrists) and under which conditions. Moreover, they should be able to start a conversation with the user proactively according to an event, programmed frequency or routine. Also they should be able to start/resume a conversation upon the user request, or set to a sleep mode upon user request. The frequency of engagement with the chatbot should be customisable.

An additional requirement is that the chatbot should provide the users a space where they can express their thoughts and feelings, be able to assess the emotional state and track the changes in the user's emotional state over time. As well, the system should be able to detect potential risks within the user's interactions and during the mood tracking.

The chatbot should be aware of the users action plans and preferences, as set by the users and/or their key workers/counsellors. In conjunction with this, the system should remind and motivate the user to engage in all aspects of their action plan. The system must have reminding and motivational functionalities which are bespoke to the users needs. The system should be able to remind and explain to the user of the benefits of adhering to the plan, while also suggesting new activities to engage in.

6 Discussion

Based on the requirements elicitation and thematic analysis, we developed a number of potential use case scenarios for an envisioned mental health chatbot system. We describe the two main ones.

6.1 Use Case Scenario 1: Intelligent Reminders

Users have stated that motivation to keep their daily tasks is key for their recovery. Therefore, a kind of companion is wished for, that is used as a daily planner for activities and tasks and has a motivation and a reminder functionality. Furthermore, these functionalities should be personalised in order to be fully accepted. In MENHIR, we will develop a sophisticated reminder for actively encouraging users to engage in activities related to their action plan. As it may not always be perceived positive to remind, for some users it may be better to let them have the responsibility of remembering their appointments. Therefore, a strategy that considers the type of user (e.g. personality, mood, activity history) will be considered. Among the reminder options, we will investigate explicit reminders that address different motivational aspects (e.g. pragmatic/emotional explanations, why engaging in activities is beneficial or necessary), and we will also explore the possibility to include implicit approaches to stimulate the memory of the user by indirect cues so that users are responsible for remembering their schedules.

A major challenge of our approach will be to find strategies that establish trust between the human user and the chatbot. Otherwise, reminders will be perceived as obstructive or obtrusive and the MENHIR chatbot as a whole will not be accepted. In order to fulfil this requirement, it is essential to develop appropriate timing strategies. Timing will be based on an event, a programmed frequency or a learned routine. Furthermore, the reminder functionality will be configurable and users should be able to put the chatbot in sleep mode. For motivating users, the dialogue systems needs access to individual action plans and preferences. Additionally, tracked user emotion could be used to adapt the motivational strategies.

For illustrating a strategy that considers user-dependent features as well as different kinds of motivation, we provide the following example:

According to her action plan, Anna is to attend to glass painting classes every Wednesday at 10.00 am. To arrive on time, Anna must leave her house at 9.00 am. In this use case, the system typically reminds her at 8.00am according to a predefined frequency. Depending on the type of users different types of reminding are foreseeable:

Option A (Pragmatic Reminder): MENHIR: Hi Anna, remember you have to leave in an hour for your glass class. Today you will learn how to colour your piece.
Option B (Emotional Reminder): MENHIR: Hi Anna, remember you have to leave in an hour for your glass class. You will have a very good time because today you will be colouring your piece. This will be the last step to achieve your objective to make your own lamp. You can be proud of yourself.
Option C (Indirect Reminder): MENHIR: Hi Anna, do you have any activities programmed for today?
Option D: - Not to remind –

6.2 Use Case Scenario 2: User Diary

The analysis of the co-creation workshop shows that users would value a chatbot that offers them companionship. They want to talk about their struggles without being judged or feeling that they are a burden to others. A chatbot can facilitate the disclosure of their thoughts, as it is inanimate and therefore impartial and permanently available without becoming fatigued.

In MENHIR, we will consider a diary-like functionality that allows users to talk about their day while displaying active listening behaviour as a way to address those needs. The main objective of this use case is to get the user talking about their day by open-ended and engaging questions, as well as to keep them talking by appropriate backchanneling and follow-up questions so that enough data is obtained to perform the automatic analysis of their state and progress. If the diary is used on a regular basis, it can be used to attend to the users' need for monitoring their mental state.

An account of daily occurrences is likely to offer sufficient material to perform emotional recognition using linguistic and paralinguistic information as mentioned before. We can also enrich the data obtained with other resources available for the community, such as those of the Audio/Visual Emotion Challenge and Workshop, which has a sub-challenge devoted to detecting depressive states [12].

A thorough understanding of the diary entry of the user is thus not within the scope of MENHIR, and was in fact not a requisite for users. Therefore, the main challenge of this use case is finding appropriate questions that signal engagement without responding to the specific semantic content conveyed by the user.

A dialogue in this use case will typically start with the chatbot proactively approaching the user to ask about their day. The question must be chosen in a way that builds rapport and does not expect a positive or negative answer, so that the user feels comfortable to talk about whatever is on their mind. There must be also variety in the questions used for opening the dialogue, to foster long-term interactions. If the user talks for a longer period of time, appropriate backchanneling shows engagement and supports building a companionship between user and system. For short answers, the system needs to ask follow-up questions that favour a more extensive answer. As a semantic analysis of every possible user utterance is not within the scope of MENHIR, the follow-up questions must be generally applicable and can at most rely on an emotional analysis if the utterance allows for it.

An example dialogue of such a scenario could be:

MENHIR: Hi Anna, how was your day?
Anna: Not so good.
MENHIR: Did something happen?
Anna: Yesterday I had an argument with my mother and I have been feeling bad all day...
MENHIR: I'm sorry to hear that. You can tell me about it if that makes you feel better.
Anna: ...

7 Conclusion

In this paper we have presented the results of a co-creation workshop in which scientists and people with mild anxiety and depression have collaborated to understand the acceptability and requirements for the development of mental health chatbots.

Results from the workshop were analysed using thematic analysis to highlight the key themes of discussion, from which we have obtained a list of 17 themes related to challenges that could be addressed (isolation and honest disclosure), functionalities (symptom recognition and monitoring, companionship, risk detection), characteristics (including personalization, proactiveness, accessibility, and privacy) and usage conditions (cost, access, intention to use). Our results provide interesting insights for the development of a mental health chatbot, including requirements and use case scenarios, from which we have highlighted and described the intelligent reminder and user diary use cases.

For future work, we will develop the dialogues in the scenarios presented. The first challenges that we will address will be the development of personalized dialogue management strategies based on representations of the user actions plans for the intelligent reminder scenario as well as the multimodal analysis of interaction recordings in the user diary scenario. Once the scenarios are designed, they will be validated with key workers and assessed in further co-creation workshops and Wizard-of-Oz evaluation sessions. For these sessions we will consider, among others, the aspects of informational support, emotional support, positive support, skill building and potential negative experiences studied recently in [13].

Acknowledgements This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR: Mental health monitoring through interactive conversations https://menhir-project.eu).

References

1. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M (2019) An overview of the features of chatbots in mental health: a scoping review. Int J Med Inf 132:103978
2. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49
3. Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR Mental Health 4(2):e19
4. Greenhalgh T, Jackson C, Shaw S, Janamian T (2016) Achieving research impact through co-creation in community-based health services: literature review and case study. Milbank Q. 94(2):392–429
5. Guest G, MacQueen KM, Namey EE (2011) Applied Thematic Analysis. SAGE Publications, Thousand Oaks
6. Hollis C, Morriss R, Martin J, Amani S, Cotton R, Denis M, Lewis S (2015) Technological innovations in mental healthcare: harnessing the digital revolution. Brit J Psychiatry J Mental Sci 206(4):263–265
7. Inkster B, Sarda S, Subramanian V (2018) An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods Study. JMIR mHealth uHealth 6(11):e12106
8. Jull J, Giles A, Graham ID (2017) Community-based participatory research and integrated knowledge translation: advancing the co-creation of knowledge. Implementation Sci 12(1):150
9. Lee M, Ackermans S, van As N, Chang H, Lucas E, IJsselsteijn W (2019) Caring for vincent: a chatbot for self-compassion. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Glasgow, Scotland UK , pp 1–13
10. McTear M, Callejas Z, Barres DG (2016) The Conversational Interface: Talking to Smart Devices. Springer, Heidelberg
11. Mulvenna MD, Bergvall-Kåreborn B, Galbraith B, Wallace J, Martin S (2011) Living labs are innovation catalysts. In: Howlett RJ (ed) Innovation through Knowledge Transfer 2010. Springer, Heidelberg, pp 253–264
12. Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, Schmitt M, Alisamir S, Amiriparian S, Messner EM, Song S, Liu S, Zhao Z, Mallol-Ragolta A, Ren Z, Soleymani M, Pantic M (2019) AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of AVEC '19. ACM Press, Nice
13. Ta V, Griffith C, Boatfield C, Wang X, Civitello M, Bader H, DeCero E, Loggarakis A (2020) User experiences of social support from companion chatbots in everyday contexts: thematic analysis. J Med Internet Res 22(3):e16235
14. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB (2019) Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 64(7):456–464
15. World Health Organization, Victorian Health Promotion Foundation, University of Melbourne: Promoting mental health: concepts, emerging evidence and practice. World Health Organization (2004). https://apps.who.int/iris/bitstream/handle/10665/42940/9241591595.pdf

Development of a Dialogue System that Supports Recovery for Patients with Schizophrenia

Chiaki Oshiyama, Shin-ichi Niwa, Kristiina Jokinen, and Takuichi Nishimura

Abstract Schizophrenia is a mental illness characterized by relapsing episodes of psychosis. Schizophrenia is treatable, and treatment with medicines and psychosocial support is effective. However, schizophrenia is one of the most expensive mental illnesses in terms of total medical costs required, including costs for effective treatment and for the continuous support and monitoring that is necessary. It is therefore useful and beneficial to explore how new technology, such as dialogue systems and social robots, can be used to provide help and assistance for care personnel as well as for the patients in the treatment and recovery from the illness. In this paper, we discuss various issues related to the development of a dialogue system that is able to recognize the characteristics of schizophrenia and provide support for schizophrenia patients 24 h a day.

1 Introduction

Schizophrenia is a mental illness characterized by relapsing episodes of psychosis. It is often associated with social stigma, which makes the treatment and the patient's re-introduction into society difficult. Schizophrenia is treatable, but it requires large-

C. Oshiyama · T. Nishimura
Human Augmentation Research Center, National Institute of Advanced Industrial Science and Technology, Chiba, Japan
e-mail: chiaki.oshiyama@aist.go.jp

T. Nishimura
e-mail: takuichi.nishimura@aist.go.jp

S. Niwa
Department of Psychiatry, Aizu Medical Center, Fukushima Medical University, Fukushima, Japan
e-mail: si-niwa@fmu.ac.jp

K. Jokinen (✉)
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
e-mail: kristiina.jokinen@aist.go.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_16

scale medical treatment and ongoing support and monitoring to recognize symptoms of schizophrenia and to provide appropriate treatment in a proactive manner as early as possible.

In this paper, we discuss the use of novel interaction technology to tackle the problems in schizophrenia treatment, focusing on issues related to detecting the early symptoms of schizophrenia, and providing appropriate support for patients so they can live independently outside of caregiving organizations.

The paper is structured as follows. We first define schizophrenia in Sect. 2, and briefly discuss the status and the types of effective treatments used today. One of the most prominent problems in the treatment is the expensive medical cost, for which AI technology may provide useful solutions. In Sect. 3 we discuss the use of spoken dialogue systems as a possible solution to automatically identify the characteristics of schizophrenia and provide 24-h, day-to-day support for the patient. Finally, we discuss future plans and conclude the paper in Sect. 4.

2 Schizophrenia and Its Current Treatment Approach

2.1 Definition of Schizophrenia

What is schizophrenia? Schizophrenia is a mental illness characterized by relapsing episodes of psychosis. It is characterized by distortions in thinking, perception, emotions, language, the sense of self, and behavior [7]. Common experiences include hallucinations (hearing voices or seeing things that are not there) and delusions (fixed, false beliefs). Schizophrenia is caused by brain dysfunction and can be diagnosed based on three main symptoms: positive symptoms, negative symptoms, and symptoms due to cognitive dysfunction. Positive symptoms refer to behavior which is not usually seen in healthy people such as hallucinations and delusions, losing touch with reality, while the negative symptoms refer to the disruption of normal behavior, reduced facial expressions (flat affect) and speaking, difficulties in starting activities, etc., and cognitive dysfunctions concern changes in memory and other cognitive functions, for example. The diagnosis criteria for a schizophrenia case is to have more than two of the five symptoms of (1) delusion, (2) hallucination, (3) disorganized speech, (4) disorganized behavior, and (5) flat affect, poverty of speech, or loss of energy.

Schizophrenia is a chronic and severe mental disorder affecting 20 million people worldwide, with an annual incidence of 0.2.–0.4 persons per 1000 people, with a lifetime prevalence of about 1%. According to strict diagnostic criteria, morbidity and prevalence are almost the same regardless of country or culture [19].

2.1.1 Treatment Costs for Schizophrenia

Schizophrenia is treatable, and treatment involves a combination of medication, psychological treatment, and support from the community for daily life. Treatment with medicines and psychosocial support is effective. The treatment goal is the patients' recovery, which refers to their social participation. Facilitation of assisted living, supported housing, and supported employment are effective management strategies for people with schizophrenia. However, schizophrenic patients use many healthcare services. For instance, in the United States, the prevalence of schizophrenia is approximately 1% of the total population, but annual mental healthcare costs for treating schizophrenia account for more than 2.5% of all US healthcare costs, including preventive interventions. Schizophrenia is one of the most expensive mental illnesses in terms of total medical costs.

2.1.2 Effective Technology for Recovery

A review by Green [7] concluded that neurocognitive abilities, rather than clinical symptoms, hinder social participation by people with schizophrenia. Characteristic profile in schizophrenia is shown in Fig. 1 [8] and it includes psychological and behavioral features due to cognitive dysfunction in schizophrenia: attention dysfunction, memory dysfunction (verbal memory, working memory), executive dysfunction (goal-setting, planning, execution), slow processing speed, limited comprehension, and learning. Improving these cognitive functions promotes social participation in schizophrenia patients.

Cognitive training with schizophrenia patients has been conducted, see e.g. [18]. However, the ramifications of training improvements have been poor [3]. In addition, when training was discontinued, the patient's performance often returned to the original level [5]. One of the most difficult aspects for schizophrenia patients in skills training is finding strategies and using them. To remedy this difficulty, the rehabilitation program by Wykes [20] uses a technique called "gradual treatment". This method focuses on problematic aspects of the task and creates steps that can slowly but surely lead to improvement, thereby creating a stable structure to solve the problem. In recent years, various programs which aim at improving cognitive functions of schizophrenia patients to improve their social skills have been developed, and the effects have been verified.

3 Dialogue Systems for Schizophrenia Recovery and Treatment

As mentioned, the medical cost of schizophrenia treatment is high. Novel AI technology may provide solutions that alleviate the situation, especially in the mid and later

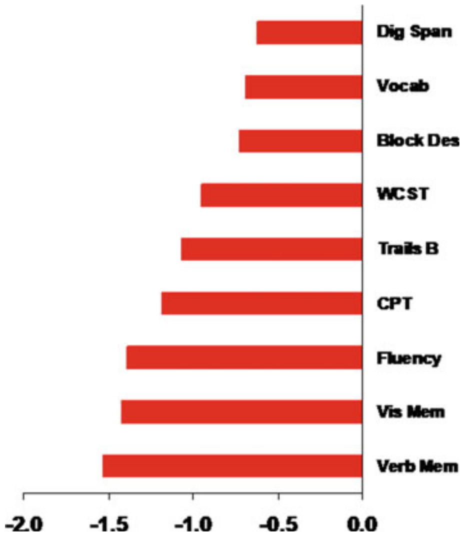


Fig. 1 Maximal impairment in memory, attention, and executive function; relative preservation of old learning and visual perceptual skills (modified from [8]). WCST = Winsconsin Card Sorting Test, Trails B = Trail making Test B, CPT = Cognitive Performance Test

phases of the treatment, when the patient's progress in daily life is more stabilized. In the later phases, it may then be possible to benefit from the use of interactive and attentive dialogue systems which can monitor and support the progress in daily life by interactively providing information of the current situation, reminders of important tasks, recommendations for various exercises, and training in cognitive tasks.

However, the design of automatic dialogue systems for the recovery and treatment of schizophrenia patients requires that several issues are considered. The main goal is not only to provide monitoring or treatment information in an effective and efficient manner to reduce labor costs, but to develop useful, user-friendly, and attentive interactive systems that can be accepted by the patients as a convenient and helpful assistant as well as judged by the professionals as a practical and beneficial "co-worker." Such goals call for thorough consideration of the technical issues concerning the development of an appropriate dialogue system and a suitable robot agent, as well as a careful survey of the requirements and needs for the interaction design from the point of view of medical staff and patients. Also ethical and privacy issues need to be carefully considered. The practical task for designing and developing suitable dialogue systems for rehabilitation thus assumes collaboration between interdisciplinary teams where designers, technicians, and engineers work together with doctors, nurses, and patients. Furthermore, successful development requires collaboration with medical institutions and governmental officials, and the support of the local communities.

3.1 Merits of Using Interactive Robots

Given the complexity of the task and the required effort and resources, development of assistive robot agents is challenging. However, it is envisaged that the benefits of such assistive social robots would greatly exceed the cost of investment, in addition to supporting health and well-being in society through the use of new technology. In fact, with technological advances, such agents and applications may become more feasible and more widespread in future (see discussion of *boundary-crossing robots* in [12]). Moreover, there are some clear merits in using interactive assistive robots to support recovery of schizophrenia patients and to assist their everyday life, which we will discuss below.

3.1.1 Tailor-Made Support

Schizophrenic patients have their own specific needs for support and the variations of their individual habits. Care-givers and supporters have to adapt to the individual ways in which the patients are supported, but there is a limit to the number of supporters and the time they can provide support in the patient's daily life. An interactive robot, which is constantly near the patient providing necessary albeit simple monitoring of the daily tasks, can effectively help the schizophrenic patient to receive sufficient support for their everyday life. Although human-level support may not be possible, the robot's support tasks can vary from reminders for regular activities (e.g., taking medicine, getting up, doing exercises, etc.) to providing factual information and peer-like support as well as offering simple chatting on everyday topics when necessary.

3.1.2 Assistance in Every-Day Situations

Interactive robots can sustain healthy and independent life by providing tools to evaluate daily activities and to advise and suggest alternative ways to act based on this data. For instance, patients with schizophrenia have trouble recognizing the emotions and intentions of others, and they thus choose inappropriate actions and also need a lot of time for systematic and comprehensive assessment of various social and practical situations. If an interactive robot agent can support the patient's evaluation of the important communicative aspects in these situations, it can provide constructive help for the patients to act appropriately in these situations. Moreover, since patients often feel less confident of what they can do themselves, such support can also boost their self-confidence and ability to act independently in various situations. Furthermore, if the robot can perform evaluation in daily clinical practice, this can be used to formulate an appropriate support policy based on these observations.

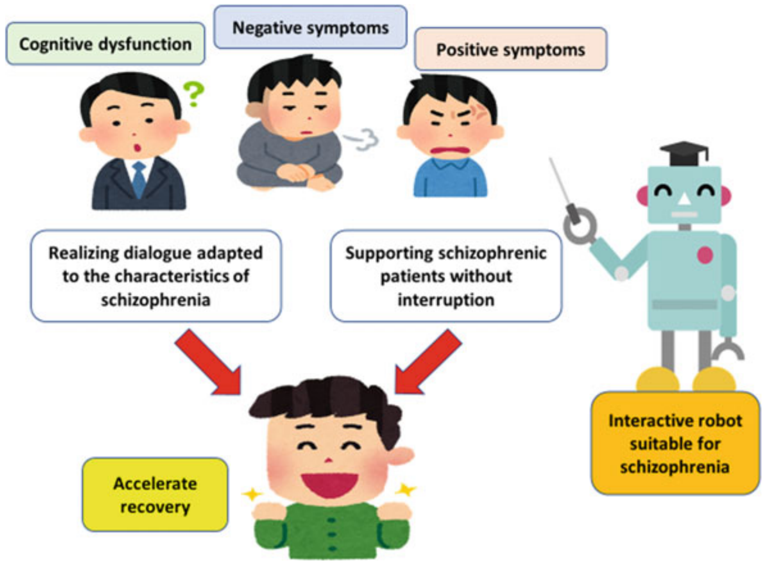


Fig. 2 Proposed work schema

3.1.3 Operation Without Human Bias

Comprehensive assessment of the patient's behaviour traits, such as the ability to recognize the emotions and intentions of others, select appropriate actions in individual situations, and the degree of loss of self-confidence ("I can't do this in any case"), takes time and also requires competence and experience of the assessor. Therefore such assesments are limited to special research situations only. However, such situations may be biased because the assessment is also human-human interaction: the patient's behavior may be influenced by the stress or tacit expectations of the assesment, wheras the assessor's own individual knowledge and subjective judgment may induce a bias in the assessment. It can be assumed that interactive robots, which have less complex interaction capabilities and a less human-like appearance, provide a partial solution to the issues concerning the assessor's lack of experience, subjective bias, and lack of time. For instance, machine learning has already been used to assess bias in clinical trials [15], while research concerning the use of robots with autistic children and in training for children's self-management of diabetes have shown that the robots motivate and engage children in a manner that often overcomes their unwillingness to interact socially with human therapists [2, 4, 13]. Therefore, we propose research and development on a dialogue system that can recognize the characteristics of cognitive dysfunction in schizophrenia and provide appropriate support for schizophrenia patients in their everyday activities. Figure 2 depicts the proposed work schema.

3.2 Towards a Dialogue System for Schizophrenia Recovery

In the healthcare domain, natural language dialogue systems have been studied in many subactivitites, e.g. instructing novice caregivers in basic care-giving tasks [11], motivational interviews [1, 16], diabetes self-management training [4], and providing support for therapists [17]. In the EU project Menhir (https://menhir-project.eu/), the focus is on conversational technologies to promote mental health and assist people with mental ill health problems (depression and anxiety) to manage their conditions. The connectionist natural language processing system DISCERN has been used to complement clinical research on the illness mechanisms proposed to underlie schizophrenia, and to formulate testable hypotheses on schizophrenic symptoms as exhibited in human story processing [6].

Our plan deals with a conversational system for schizophrenia patients, and the ambitious goal is to encourage patients in their independent living by providing support based on observations of the environment and on suitable reaction plans for their situational needs. The system can assist schizophrenia patients in various social interactive situations by recognizing the characteristics of schizophrenia, assessing the progress of the patient's recovery process, monitoring the success of treatment, and providing appropriate support for their daily life.

In general, healthcare systems are based on two complementary tasks: monitoring an ongoing session (i.e., categorization of the observed symptoms) and forecasting the patients' state from the symptoms (i.e., predicting the recovery level, or the patient's needs). We plan to use a robot platform (although other mobile and virtual agent platforms are also possible) with basic dialogue capability; see [9] for a list of requirements for the flexible dialogue functionality of interactive service robots. From the interaction point of view, we also consider the system's decision-making ability, that is, the type and level of reasoning that the system should be able to conduct to support and accelerate the users' recovery. For this application, an important aspect of the interaction design is the system's ability to explain its functioning: why it arrived at a particular solution (especially if the system is based on machine learning technologies rather than clear, simple rules). While the system may exhibit a certain level of independent and autonomous behavior, its dialogue behavior (the information and recommendations it provides) should be clear and transparent to enable the user's easy understanding of the situation and, more importantly, to create trust in and verify the reliability of the system's behavior.

Our proposal for such a dialogue system contains three general functional levels that link to the expected autonomy of the system's behaviour and the dialogue capability that the system should exhibit. These levels are dependent on the limitations of individual technologies.

Information Providing QA System. The system will provide information about the recognized symptoms and can also explain their meaning, functional correlations, role in the therapy, etc. The system can be enabled by a straightforward question-answering system with an explanatory facility. From the point of view of dialogue technology, the system may be text-based such as an online chatbot-type implementation (cf. Alexa Prize), or it can use speech like IBM Watson, or even emotional and personality aspects like Facebook's BlenderBot. For the task-oriented system needed to support dialogues with schizophrenia patients, flexible natural language generation is important to address the patient's individual needs (e.g., vocabulary and sentence length) and to adapt to any requirements in the patient's communication strategy. If the application also includes sensors and multimodal technologies, the basic interaction capability needs to be expanded by taking the multimodal information into account. This means that more complex modelling is required, and the system architecture also needs to integrate all the necessary components that deal with the detection, interpretation, and fusion of the multimodal information.

Assistant System for Skill Enhancement. The system can also instruct and train the schizophrenia patient for suitable activities and action sequences for the purpose of improving the patient's cognitive skills (memory, attention) and enhancing their ability to live independently outside of a hospital. It is expected that the interaction is conducted in speech, so appropriate speech technology is necessary, in addition to the natural language and possible multimodal technologies discussed above. While the interactive sessions between the patient and the robot agent may be independent and autonomous including various games, exercises, and short tasks, the selection of suitable exercises may be determined by caregiving staff or by the patients themselves. The intended interaction capability is typical for dialogue systems that address the requirements for flexible QA interaction, yet allow the human caregivers and the patients to have control over the intended action. The interaction model requires a minimum understanding of the dialogue context and its impact on the interaction, e.g. choosing the right assessment protocols after a training, or recognizing a returning user to the training programme.

Diagnostic Support for Progress Assessment. The system provides diagnostic support to evaluate the patient's progress and recommends suitable tasks and exercises to further accelerate the patient's recovery based on the observed symptoms. This level of functioning requires advanced technology in signal processing, natural language processing, and dialogue modelling, in order to integrate multimodal signals and the system's reasoning capabilities into natural interaction capabilities. Research on the integration and application of knowledge in dialogue systems is actively being conducted, although deeper investigations on knowledge, reasoning, and AI techniques are necessary to fully understand the context and its correlation with the various signals. Aspects of knowledge management, such as the use of knowledge graphs, semantic categories and ontologies in the interpretation and generation of language expressions as well as in the interaction management as a whole, are important steps in the development of more natural and interactive support systems. AI technology also offers possibilities for learning new concepts and contextual constraints interactively with human partners, as well as imitation learning for plausible action sequences. Although complete realization of such intelligent and learning dialogue agents may not be possible as a near-future solution due to the complexity of the tasks involved, new frameworks related to cognitive architectures and context-aware dialogue models [9] pave the way towards technologically advanced solutions that enable the development of more intelligent and interactive support systems.

We believe that explorations in the use of new interaction technology to support schizophrenia patients' recovery is a positive step forward in the treatment of schizophrenia. The benefits will not only include financial issues related to reduction of treatment costs, but also possibilities to harness technology for assisting and monitoring the patient's constructive development. This can provide novel solutions that support the patient's independent living, and consequently, address challenges for building a society that is good for all.

3.3 Ethical Consideration of Supporting Dialogue Systems

The design and development of interactive systems that can provide support and assistance as discussed above, also presupposes consideration of ethical and privacy issues. These are currently under extensive discussion in scientific and legal communities, focussing on data protection, privacy, and responsibility issues. Ethical issues in assisting systems, especially in age-approriate systems are discussed in [14].

Important issues related to dialogue systems concern truthfulness of the presented information and the user's trust in the system's capabilities (such as giving appropriate treatment suggestions, understanding the user's utterances, showing reliable accuracy in reminders, etc.). Equally important is the transparency of the data handling and its manipulation, and clarity of the ownership of the data: who has access to private and sensitive data, how the access rights are determined and who can change them. Moreover, as the support system also assumes monitoring of the person's every day activities, such monitoring may be regarded as intimidating and interfering with one's personal sphere, and the consent for using such autonomous systems is crucial. On the other hand, as discussed above, such monitoring dialogue systems may also be regarded as effective and reliable by the patient because they may appear non-judgmental and provide more consistent feedback. For further discussion on ethical and privacy considerations of dialogue systems see [10].

4 Conclusions and Future Work

In this paper, we have discussed several issues concerning the development of a dialogue system for schizophrenia treatment and proactive support. We discussed various schizophrenia characteristics and treatment forms, and sketched requirements and dialogue competence levels for the development of such assistive systems. In the future, the system design will continue both from the point of view of technical requirements of the robot (its physical and technical properties) and the point of view of the users' needs and expectations (specification of the desired interaction capabilities). We will elaborate the plans in collaboration with colleagues in the medical, social, and engineering domains, and focus on the multifaceted questions concerning the acceptability and adoption of social robots as a reliable member of the care staff and as a helpful companion for the patient. Finally, we aim to collect feedback from the patients and staff to better understand the pertinent issues in the development of such systems to support schizophrenia treatment and recovery processes and sustain the patient's ability to lead a normal independent life.

Acknowledgements The work is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

1. Baer JS, Wells EA, Rosengren DB, Hartzler B, Beadnell B, Dunn C (2009) Agency context and tailored training in technology transfer: a pilot evaluation of motivational interviewing training for community counselors. J Subst Abuse Treat 37(2):191–202
2. Billard A, Robins B, Nadel J, Dautenhahn K (2006) Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. RESNA Assist Technol J
3. Brenner HD et al (1996) Integrated Psychological Therapy for Schizophrenic Patients. Hogrefe and Huber, Gottingen
4. Coninx A et al (2015) Towards long-term social child-robot interaction: using multi-activity switching to engage young users. J Hum-Rob Interact
5. Goldberg TE et al (1987) Further evidence for dementia of the prefrontal type in schizophrenia? a controlled study of teaching the Wisconsin Card Sorting test. Arch Gen Psychol 44:1008–1014
6. Grasemann U, Hoffman R, Miikkulainen R (2011) Modeling acute and compensated language disturbance in schizophrenia. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society. http://www.cs.utexas.edu/users/ai-lab?grasemann:cogsci11
7. Green MF (1996) What are the functional consequences of neurocognitive deficits in schizophrenia? Am J Psychiatry 153:321–30
8. Heinrichs RW, Zakzanis KK (1998) Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. Neuropsychology 12(3):426–445
9. Jokinen K (2018) Dialogue models for socially intelligent robots. In: Ge S, Cabibihan J, Salichs M, Broadbent E, He H,Wagner A, Castro-González A (eds) Social Robotics. ICSR 2018, pp 127–138. Springer, Heidelberg
10. Jokinen K, Fukuda K, Iino N, Nishimura S, Nishimura T, Oota Y, Watanabe K, Yoshida Y (2019) Ethical and privacy issues in interactive service applications for elder people concerning dialogues and social robots. In: 7th Serviceology Conference, Tokyo
11. Jokinen K, Nishimura S, Watanabe K, Nishimura T (2019) Human-robot dialogues for explaining activities. In: Fernando D'Haro L, Banchs R, Li H (eds) 9th International Workshop on Spoken Dialogue Systems, pp 239–251. Springer, Heidelberg
12. Jokinen K, Watanabe K (2019) Boundary-crossing robots: societal impact of interactions with socially capable autonomous agents. In: Salichs H, Ge SS, Barakova EI, Cabibihan JJ, Wagner AR, Castro-González Á, He H (eds) Social Robotics, pp 3–13. Springer, Heidelberg
13. Kruijff-Korbayová I et al (2015) Young users' perception of a social robot displaying familiarity and eliciting disclosure. In: Tapus A et al (eds) Proceedings of ICSR 2015, LNAI 9388, pp 380–389. Springer, Heidelberg
14. Manzeschke A, Weber K, Rother E, Fangerau H (2015) Ethical questions in the area of age appropriate assisting systems
15. Marshall IJ, Kuiper J, Wallace BC (2016) RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Neuropsychology 23(1):193–201

16. Pérez-Rosas V, Mihalcea R, Resnicow K, Singh S, An L, Goggin KJ, Catley D (2017) Predicting counselor behaviors in motivational interviewing encounters. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol 1, pp 1128–1137
17. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V (2016) A comparison of natural language processing methods for automated coding of motivational interviewing. J Subst Abuse Treat 65:43–50
18. Wexler B, Hawkins K, Rounsaville B, Anderson M, Sernyak M, Green M (1997) Normal neurocognitive performance after extended practice in patients with schizophrenia. Schizophr. Res. 26(2–3):173–180
19. WHO: Schizophrenia (2017). https://www.who.int/news-room/fact-sheets/detail/schizophrenia
20. Wykes T (2000) Cognitive rehabilitation and remediation in schizophrenia. In: Sharma T, Harvey PD (eds) Cognition in schizophrenia: Impairments, importance, and treatment strategies. Oxford University Press, Oxford

Human-Robot Interaction

Caption Generation of Robot Behaviors Based on Unsupervised Learning of Action Segments

Koichiro Yoshino, Kohei Wakimoto, Yuta Nishimura, and Satoshi Nakamura

Abstract Bridging robot action sequences and their natural language captions is an important task to increase explainability of human assisting robots in their recently evolving field. In this paper, we propose a system for generating natural language captions that describe behaviors of human assisting robots. The system describes robot actions by using robot observations; histories from actuator systems and cameras, toward end-to-end bridging between robot actions and natural language captions. Two reasons make it challenging to apply existing sequence-to-sequence models to this mapping: (1) it is hard to prepare a large-scale dataset for any kinds of robots and their environment, and (2) there is a gap between the number of samples obtained from robot action observations and generated word sequences of captions. We introduced unsupervised segmentation based on K-means clustering to unify typical robot observation patterns into a class. This method makes it possible for the network to learn the relationship from a small amount of data. Moreover, we utilized a chunking method based on byte-pair encoding (BPE) to fill in the gap between the number of samples of robot action observations and words in a caption. We also applied an attention mechanism to the segmentation task. Experimental results show that the proposed model based on unsupervised learning can generate better descriptions than other methods. We also show that the attention mechanism did not work well in our low-resource setting.

1 Introduction

In the recent advance of human assisting robot technologies, the ability to explain robot behaviors is becoming an important task. Robot actions are often generated from uninterpretable systems, such as deep neural networks [8]. However, robots

K. Yoshino (✉) · K. Wakimoto · Y. Nishimura · S. Nakamura
Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, Nara 6300192, Japan
e-mail: koichiro@is.naist.jp

K. Yoshino · S. Nakamura
Center for Advanced Intelligence Project (AIP), RIKEN, 8916-5, Takayama, Ikoma, Nara 6300192, Japan

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_17

are expected to be explainable for their behavior because they will cooperate with humans in daily life areas. Most of the existing works for bridging robot actions and natural language captions are oriented to generated robot actions given natural language commands [6, 12, 21, 25]. They tackled the problem in a pipeline process by using manually designed base action units, i.e., small basic robot actions, which are easy to recognize and generate. Both robot actions and captions were converted to sequences of base action units and then bridged each other.

In contrast, end-to-end approaches have been widely used to date by recent advances in the neural network field [3, 4, 23]. These approaches directly train mappings between two sequences. Such end-to-end mapping is applied to robot behavior captioning, which generates a natural language caption given a raw robot action sequence [20, 24, 26, 29].

It is not easy for human assisting robots to generate natural captions from their observations, sequences of actions, and observation from robot sensors, because robots encounter diverse environments and tasks. Collecting large-scale dataset for any different robots or environment is not realistic. Furthermore, one robot observation sequence has more samples than words in a caption due to sampling rates of observation devices. However, this gap makes it challenging for neural networks to learn relationships between robot observations and natural language captions. This problem is especially true if we apply neural network-based approaches, in particular recurrent neural networks, which are generally used for sequence-to-sequence learning.

Segmenting robot actions to base action units will contribute to making the training process more straightforward. In the work we performed, we focused on using unsupervised segmentation methods [15–17] for speeding up and stabilizing sequence-to-sequence training, rather than using manually designed base robot action units [6]. We used K-means clustering and chunking based on byte-pair encoding (BPE) to learn the unsupervised base action segments of the robot. K-means clustering can unify similar robot observation into one class. Chunking based on BPE can wrap a typical sequence of robot actions into one class.

We also used an attention mechanism in sequence-to-sequence learning [13, 27], which explicitly learns the correlation between input and output sequences. Attention mechanism calculates correlation weights between each input sample and each output sample. It is known that the attention mechanism can solve a problem of sample number gaps in the area of speech recognition and synthesis [5, 28]. We expect that the training results obtained for the attention can be incorporated into a robot action unit vocabulary that corresponds to actual words in sentences. We combined both methods; unsupervised action segment learning and attention mechanism, to generate more natural captions since we expected that each method would provide different contributions.

In experiments, we recorded videos of robots with their observation: actuator sequences and videos recorded by a first-person viewpoint camera. We used these videos to collect robot behavior captions through crowdsourcing, which corresponds to robot action sequences. However, we limited the number of collected samples toward a realistic situation to apply the method. We conducted two experiments:

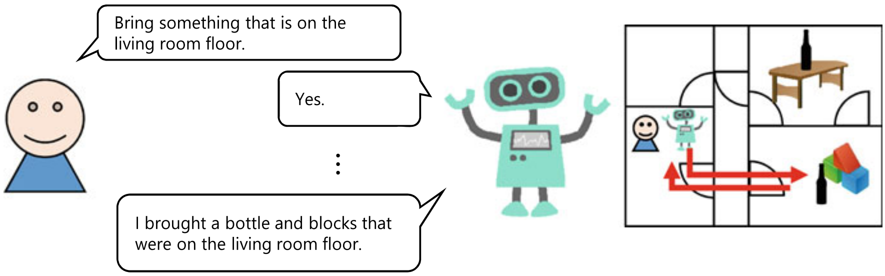


Fig. 1 Example human robot communication obtained through natural language commands/descriptions

an automatic evaluation based on references in the test-set and human subjective evaluation. These results showed that the proposed three types of base action units contribute to improving better robot behavior captions in natural language. In particular, robot action units acquired by the unsupervised learning based on clustering and chunking got the best performance in human subjective evaluation.

2 Natural Language Captions for Robot Action Sequences

2.1 Problem Setting and Related Work

For daily life scenarios, various robot tasks have been designed as human assisting robots have evolved. In particular, these tasks include human assisting tasks, such as moving objects in rooms and recording videos of actual places. In these situations, it is required for robots to communicate with users about their behavior in natural language. There are two important directions. The first direction is deciding robot actions according to the natural language commands by the user. Another direction is providing robot behavior captions in natural language. Both directions are important to realize cooperative human assisting robots. In these tasks, there are sequences of robot actions and natural language captions/descriptions, and the robot needs to learn the mapping between them. Figure 1 shows an example of the task in a particular situation. In this example, the user requests the robot to "Bring something that is on the living room floor," and the robot does so. After this, the robot uses natural language to tell the user what was done, e.g., "I brought a bottle and blocks that were on the living room floor". In this paper, we focus on the latter problem, generating a natural language caption given a robot action sequence. We define the problem as generating a word sequence Y given a sampled robot observation sequence S.

As the robot action sequence S, most reported existing works have been done using the pipeline process, such as handcrafting base action units, recognizing all robot observation sequence portions with these base action units and constructing

Fig. 2 HSR robot on SIGVerse environment, conducting a task in the setting of WRS

complicated actions [12]. However, these approaches for handcrafting base action units are costly.

Based on the advance of technologies based neural networks, some works tried to build sequence-to-sequence learning system between robot observations and natural language captions. Takano et al. [24] used bi-grams for end-to-end learning between natural language captions and robot action sequences, to separate the semantics and the language expressions in captions. Yamada et al. [29] utilized recursive autoencoders for the learning so as to minimize the distance between the latent variables of autoencoders for robot actions and natural language captions, toward the end-to-end bridging. In another study, a simple encoder-decoder was applied to learn the relation [20]. However, these methods require large-scale training data. It is not realistic to prepare such large-scale training data whenever the robot is applied to a new environment, or when the robot functions are updated. The physical environments of robots themselves will be different if a new robot is developed.

To prevent problems caused by a lack of training data, in this paper, we applied two robot action segmentation types in an unsupervised manner to make it possible for the system to learn the relationship from a small data amount. Some related studies have tried to define base action units with unsupervised learning [15–17] or adaptively [10]. These studies focused on structuring robot actions from data. Such clustering-based approaches have the potential to be used for reducing training data with simplifying symbols contained in training data.

2.2 Robot Task, Environment and Data in Simulator

In this section, we describe collected dataset for the training of captioning given robot observations. We used the service category environment defined in the World Robot Summit (WRS) [11] as the robot task for our action description generation. We used Human Support Robot (HSR) [30], a human assisting robot in a domestic environment, and also used a robot simulator SIGVerse [9]. We followed WRS

Partner Robot Challenge (Virtual Space) instruction to build our environment.[1] The HSR robot in SIGVerse environment is shown in Fig. 2. The robot conducted tasks in this environment and described its actions with natural language behavior captions. We generated robot actions in the simulator and added their captions, by using crowdsourcing, as our training data.

The input S contains nine-dimensional motor actuation (rotations) and a three-dimensional robot moving (two directions and one rotation) that is sampled every 0.3 s. The robot also uses 160×120-pixel images from a first person viewpoint camera mounted on robot arm's tip. The image feature is embedded into a ten-dimensional vector by using convolutional autoencoder (CAE) [14]. The robot action actuation and the image features are concatenated into one vector, which will be used as an input of encoder at each time-step.

We used crowdsourcing to add a caption Y for a robot action sequence S. We made videos of robot action sequences from third-person viewpoint camera as shown in Fig. 2. We showed the video to crowd-workers and then requested them to give their captions that describe behaviors of the robot in natural language. [2] We made 50 videos and added 20 captions for each. Robot actions in the dataset contained "bring", "put", "pick up", "drop" and "go to see" actions and 10 different videos for each action. Note that, we did not give crowd-workers any instructions on how they should include these verbs in descriptions. This means that, for example, they might use "get" as an alternative to "bring" in accordance with their language sense. In total, we collected 1000 sentences in Japanese, which were segmented by KyTea [18]. Annotated captions are simple sentences, which consist of single predicate and several arguments, i.e., "Drop the toy dog from the table" or "Pick up the teapot on the table".

2.3 Bridging Based on Encoder-Decoder

We show the architecture of end-to-end relationship learning between the robot action sequence and the word sequence based on an encoder-decoder, which had been proposed previously [20]. The overall architecture in a time sequence is shown in Fig. 3. As mentioned in the problem setting, the system gets an input robot observation sequence S to a generate natural language caption Y. Here, s_i indicates robot observations, the raw trajectory of robot action and a frame recorded by cameras on the robot, at time i and y_j indicates the jth word in the natural language description. The encoder embeds the robot action s_k in time k to the hidden layer h_k as

$$h_i = \sigma(W_{sh}s_i + W_{hh}h_{i-1} + b_h), \tag{1}$$

[1] https://worldrobotsummit.org/download/rulebook-en/rulebook-simulation_league_partner_robot_challnege.pdf.

[2] Dataset including captions will be distributed when we publish the paper.

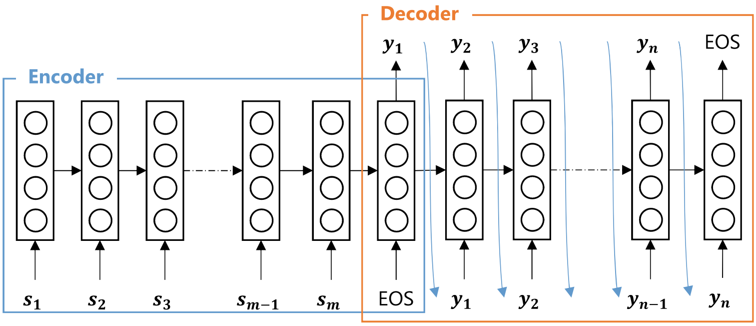


Fig. 3 Basic encoder-decoder architecture. In our problem, the encoder receives a sample of robot observation s_i in time i according to its sampling rate. Once the network receives the end-of-sequence symbol (EOS), the decoder start to generate a description word sequence

where, W_{sh} and W_{hh} are weight matrices for conversion and b_h is a bias. σ is the activation function. The decoder starts generating word w_j as soon as the encoder receives the end-of-sequence (EOS) symbol and continues to generate as it generates w_{n+1}. The network is updated by

$$h_i = \sigma(W_{yh}w_{i-1} + W_{hh}h_{i-1} + b_h), \quad (2)$$

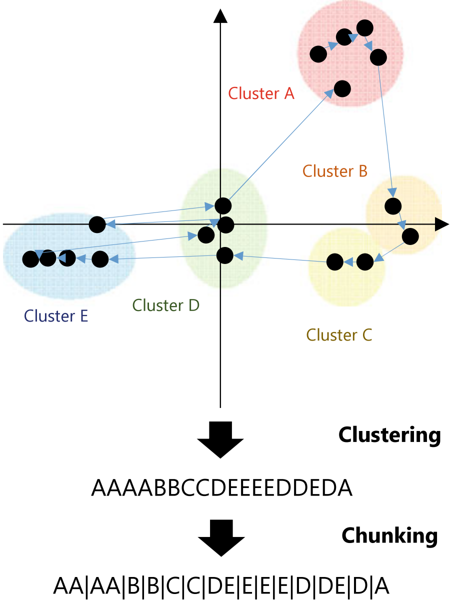
$$y_i = softmax(W_{hy}h_i + b_y), \quad (3)$$

where W_{yh}, W_{hh} and W_{hy} are decoder weight matrices, and b_h and b_y are bias. $softmax$ is the activation function to predict words. This architecture was proposed in the task of machine translation and is widely used on a variety of mapping tasks between sequences, such as conversational response generation. The difference from machine translation or conversational response generation tasks is that the robot action sequence S will be very long in accordance with the robot trajectory sampling rates. This property increases the risk of that vanishing gradients [1]. In the area of speech recognition or synthesis, attention mechanism is used to solve the problem [5, 28]. The problem can be mitigated if there is sufficient training data; however, this means that large-scale new data will be required for each new domain or each robot task.

3 Robot Action Segmentation

The dataset we used had only 1000 samples, corresponding to 50 actions, thus making it difficult to learn a good mapping with a vanilla encoder-decoder. However, as we mentioned before, collecting a large-scale dataset for all domains and robots is not a realistic option. Thus, we trained robot action segments as base action units by using unsupervised learning: clustering and chunking. We also incorporated the attention

Fig. 4 Robot action segmentation process: clustering and chunking



mechanism into the encoder-decoder, since we expect the attention weights could learn the action classes.

3.1 Robot Action Segmentation Based on Clustering and Chunking

Our system samples raw robot actions every 0.3 s. Figure 4 showing a clustering and chunking example. This graph shows robot actions in only two dimensions for an easy description. Robot action vectors are quantized by k-means clustering [7], which decides what the centroids should be for the defined number of classes. In this example, the number of samples is 18 and the number of class is set as 5. We used the Elbow method [2] to determine the class number should be 150.

After the clustering, we extracted some chunks by using byte-pair encoding (BPE). BPE extracts frequent symbol sequences and replace them to another symbol as maximizing the compression of a given raw sequence. The robot action sequence quantized by clustering is still longer than the word sequence; however, we can shorten the sequence by applying BPE. The BPE units can be used as base action units. In the Fig. 4 example, AA and DE are extracted as vocabularies. In our experiments, we set the BPE vocabulary size as 200 and trained the vocabulary from the training data. We refer to this method as "explicit segmentation" in the following sections.

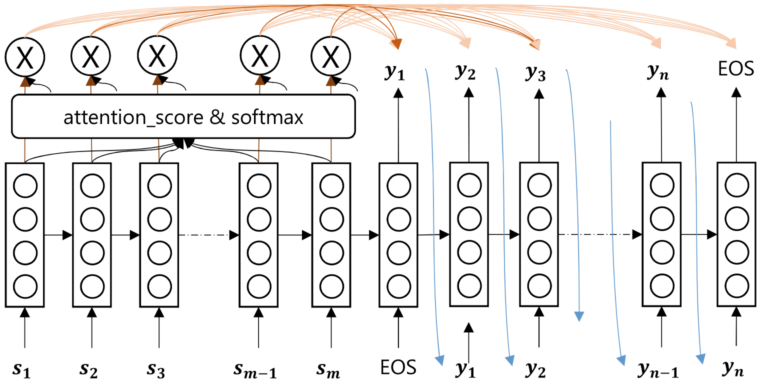


Fig. 5 Encoder-decoder with attention mechanism

3.2 Segmentation Based on Attention Mechanism

We conducted explicit segmentation as a pre-processing of relation learning by using encoder-decoder neural networks. In contrast, the attention mechanism can consider input chunks in attention weight learning implicitly. The attention mechanism learns the mapping from an input sample to an output sample by using a gating mechanism. In our situation, the number of input samples is larger than the number of output samples; thus input samples associated with the same output sample can be interpreted that they belong to the same class.

Attention mapping is calculated between the hidden layer of encoder $h_{e,i}$ and the hidden layer of decoder $h_{d,j}$ as,

$$a_{i,j} = h_{e,i}^T W_a h_{d,j}. \tag{4}$$

An attention weight is trained as a part attention vector a_j in each decoding step. Input samples that have similar attention weights for an actual output sample are strongly used in the decoding of the associated output. We expect that this property of attention mechanism works as segmentation. We show an encoder-decoder that has an attention mechanism in Fig. 5. In this example, strong attention weights are indicated in a deep color. States s_2 and s_3 have higher attention weights to y_3 in the output, and states s_{m-1} and s_m have higher attention weights to y_1. We expect that such weights in the attention mechanism work as base action units. We hereafter call this method "implicit segmentation".

Note that, it is known that the attention mechanism contributes to improve the entropy of the generated sentence; however, it also often causes an overfitting and dull generations. In other words, sentence naturalness will be improved by the attention mechanism; however, the system often generate similar sentences to different input sequence.

3.3 Hybrid Method of Implicit and Explicit

Explicit segmentation segments sequences by focusing on input sequence information, in contrast to that of implicit segmentation, which focuses on the relationship between the input sequence and the output sequence for segmenting the input sequence. In other words, these segmentations use different types of information. Thus, we integrated these segmentations into a single method as "hybrid segmentation" to make the best of them both. In the hybrid segmentation, we used explicit segmentation to segment the data as pre-processing, and then fed the data into the encoder-decoder with the attention mechanism to combine with the implicit segmentation method.

4 Experiments

In our experiments, we used the data collected in Sect. 2.2 to train an encoder-decoder that generates a word sequence of natural language caption T given a raw robot action sequence S. As the baseline, we prepared a vanilla encoder-decoder, because the focus of this paper is investigating a good robot action segmentation, which makes it possible to train the encoder-decoder from a small amount of data. We compared three methods described in Sect. 3; explicit segmentation, implicit segmentation, and hybrid segmentation, with the vanilla encoder-decoder. Experimental details follow.

4.1 Experimental Settings

We separated our 50 robot actions into 40/5/5 as training/validation/testing datasets and conducted ten cross-fold validation to evaluate all the data. The number of decriptions associated with one robot action was 20; thus training/validation/testing datasets cotains 800/100/100 pairs of sequences. We used a one-layered LSTM that has 160 units as our encoder-decoder model. The batch size was 64, the dropout rate was 0.5, the learning rate was 0.001, and the weight decay was 1e-0.6. We decided the number of training epochs by using the validation dataset.

4.2 Automatic Evaluation with BLEU

We used automatic evaluation metrics BLEU-2, 3 and 4 [19] to evaluate the generated captions, by using references annotated to videos in the testing set. BLEU-n calculates the ratio of matched n-grams with smaller numbers of n, between the genereation and the reference. Each video has 20 reference sentences; thus we used multi-reference

Table 1 Evaluation with BLEU scores. Note that the range is from 0 to 100

Model	BLEU-2	BLEU-3	BLEU-4
Vanilla	0.1	0.0	0.0
Explicit	18.0	11.7	8.4
Implicit	21.4	14.7	10.4
Hybrid	23.4	16.2	11.7

Table 2 Subjective evaluation result distribution for each method.

Model	a	b	c	d	e	a–c
Vanilla	0.0%	0.0%	0.0%	0.0%	**100.0**%	0.0%
Explicit	**10.6**%	**17.2**%	**37.2**%	28.3%	6.7%	**65.0**%
Implicit	3.9%	5.0%	35.0%	**56.1**%	0.0%	43.9%
Hybrid	7.2%	8.9%	33.3%	43.9%	6.7%	49.4%

setting in the automatic evaluation. We calculated BLEU-n scores for each reference and selected a reference that has the best BLEU-n score for the evaluation.

Evaluation scores, BLEU-2, 3 and 4, for each model are shown in Table 1. Vanilla encoder-decoder scores were deficient and scores are improved by introducing segmentation methods. Both the explicit segmentation and the implicit segmentation were effective in improving scores. The hybrid segmentation achieved the best score in each metric; it indicates the explicit segmentation and the implicit segmentation made different contributions to each other. However, scores are slightly low even if the task is simple. We will investigate the reason on the next human subjective evaluation and generation analysis.

4.3 Human Subjective Evaluation

The automatic evaluation with BLEU calculates distance from references to generated sentences. The results have moderate correlations to the quality of generation; however, there is still a gap between the result of evaluation with BLEU and human impression to the quality. For example, it is challenging to evaluate variations of word selection by only using BLEU. Thus, we conducted a human subjective evaluation to investigate the qualities of generated sentences from each method.

We showed both robot action videos and their associated captions generated by models to human subjects for scores these generated captions. In the experiment, each subject evaluated 20 sentences for each method, 80 sentences in total (for 4 methods). Following evaluation criteria are shown to human subjects.

(a). The sentence is exactly describing the robot action in the video.
(b). The sentence is mostly describing the robot action in the video; even it contains some minor errors.

Table 3 Generation examples. The "score distribution" indicates numbers of evaluators who gave a, b, c, d, and e

Method	Generation results (Translation)	Score distribution
Reference	inu no omocha wo tehburu kara otoshite (Drop the toy dog from the table.)	–
Vanilla	te te te te te te te te te te te te te (Repeating function words.)	e:9
Explicit	ringo wo otoshite (Drop the apple.)	b:2, c:7
Implicit	tehburu no ue no sohsu wo totte kite (Bring the sauce on the table.)	d:9
Hybrid	tehburu no ue no koppu wo totte kite (Bring the cup on the table.)	d:9
Reference	tehburu no ue wo nozoki konde (Look at the top of the table.)	–
Vanilla	te te te te te te te no no no no no no (Repeating function words.)	e:9
Explicit	shinshitsu no yousu wo mite (Look at the bed room.)	b:3, c:5, d:1
Implicit	tehburu no ue no sohsu wo totte kite (Bring the sause on the table.)	a:1, d:7, e:1
Hybrid	heya no ue no yousu wo totte (Bring the state of the upper of the room.)	d:2, e:7
Reference	yuka no thipotto wo hirotte (Pick up the teapot on the floor.)	–
Vanilla	te te te te te te te te te te te te te (Repeating function words.)	e:9
Explicit	yuka no sohsu wo totte (Pick up the sauce on the floor.)	a:7, b:1, c:1
Implicit	tehburu no ue no sohsu wo totte (Pick up the sause on the table.)	b:1, c:7, d:1
Hybrid	yuka ni aru sohsu wo totte (Pick up the sauce placed on the floor.)	a:7, c:2
Reference	tehburu no koppu wo katadukete (Clear away the cup on the table.)	–
Vanilla	te te te te te te te te te te te te te (Repeating function words.)	e:9
Explicit	tehburu no ue no okimono wo totte kite (Bring the ornament on the table.)	a:1, b:5, c:3
Implicit	tehburu no ue no sohsu wo totte kite (Bring the sause on the table.)	b:1, c:8
Hybrid	tehburu no ue no okimono wo totte kite (Bring the ornament on the table.)	a:1, b:4, c:3, d:1
Reference	tehburu no thipotto wo kauntah ni oite (Bring the teapot from the counter to the table)	–
Vanilla	te te te te te te te no no no no no no (Repeating function words.)	e:9
Explicit	kicchin no ue ni aru kan wo totte kite (Bring the can on the kitchen.)	b:1, c:3, d:5
Implicit	tehburu no ue no sohsu wo totte kite (Bring the sause on the table.)	a:6, b:3
Hybrid	kicchin no ue ni aru ka mite kite (Look at the upper of kitchen to check.)	d:9

(c). The sentence is moderately describing the robot action, but the description still has some major errors on target objects or environments.
(d). The sentence is grammatically correct; however, the action name and the object name are different.
(e). The sentence has wrong grammar and is meaningless.

(a)–(c) are prepared to check semantics of generated captions in different levels, compared with (d) and (e) focus more on sentence syntax. Human subjects selected one of them for the presented pair of a robot action video and its generated captions. The number of subjects was 9; the total number of evaluated example for each method was 180. Table 2 shows results.

From the results, we can know that the vanilla encoder-decoder did not generate any meaningful sentences; this result indicates that it is challenging to learn the mapping between robot action sequences and caption word sequences from a small portion of data. In contrast, tried segmentation methods make it possible the model to generate many meaningful sentences; in particular, the explicit segmentation method generated meaningful sentence a lot. Human subjects judged that 65% of generated sentences from the explicit segmentation method is meaningful (a, b and c); this result indicates the usefulness of our proposed segmentation method. Overall score is still slightly poor because of the number of training samples, as a result, the majority in the evaluation result was c. This result indicates that the explicit segmentation method successfully learned predicates from robot action trajectory; however, failed to learn about objects names from camera observations in most cases. The implicit segmentation method also improved the score from the vanilla system score; however, the ratio of meaningful generated sentences was 43.9%; it was less than the result of the explicit segmentation method. This result indicates that it is challenging to learn the relation by using only the attention mechanism, even if the effect is high. The implicit segmentation method generated no ungrammatical sentence; this result indicates the potential of using attention mechanism, even if its score is less than the explicit segmentation method. This is probably because the attention mechanism is more focusing on fluency than content words. The score of the hybrid segmentation method was slightly less than the explicit segmentation method, even if the method achieved the best BLEU score in the automatic evaluation. This problem is probably caused by overfitting of word prediction in the softmax cross entropy loss calculation of encoder-decoder. However, the explicit segmentation method lightened the problem of overfitting than other methods. In summary, the explicit segmentation method, which uses clustering and chunking, achieved the best score for captioning robot actions in an end-to-end manner, in particular, generating a caption of actions (predicates) from the robot action trajectory. However, we still have a problem with generation details, i.e., object names in descriptions.

4.4 Discussing Generated Sentences

Table 3 shows examples generated from each method with their scores on human subjective evaluation. First, it can be seen that the vanilla method generated ungrammatical sentences in most cases. This supports our assumption that it is tough for a vanilla encoder-decoder to train the mapping from a small data-set. In contrast, the explicit segmentation method generated grammatical sentences. In particular, verbs in sentences were correct in most cases. This indicates that our method successfully learned the mapping between robot trajectories and verbs.

Generations from the implicit segmentation method was fluent; however, it sometimes contain similar sentences even if robot actions are different. For example, if we look at the first and second examples, only the explicit segmentation method can generate verbs “drop” and “look”, compared with implicit and hybrid generate “bring”, which is the most frequent verb in our training data. As observed elements, the implicit segmentation method often generated the word “sauce”, because the word was frequent in the training data. This is probably caused by the overfitting of neural networks and generating dull generations in many cases. The explicit segmentation method reduced the problem in some cases.

The other problem that indicates the overfitting of the system is generating “sauce” and “table” as objects. The implicit segmentation method generated these words frequently than other methods. These words are more frequent than other objects; however, the implicit method generated these words in most cases, even if the other two methods generated different object and place names. This indicates the difficulty of our task, learning a mapping between robot observations and natural language captions from a small amount of data, and the contribution of clustering and chunking with explicit methods.

In many cases, robots do not successfully identify objects. This is probably since the system misunderstood visual information because our method did not have an object recognition module. Pre-training based on object labels is a possible solution solve this problem.

5 Summary

This paper presented a method that generates robot behavior captions in natural language setting from robot action sequences, trajectories, and images, toward a cooperative human assisting robot that has explainability. The method uses two different segmentation types for robot action actuation for training from small-size data: explicit segmentation based on clustering and chunking and implicit segmentation based on the attention mechanism of neural networks. We also described the hybrid segmentation that integrates both types. Experimental results indicated that the proposed methods improved generated caption quality, especially for generating verbs related to the robot trajectories.

We are still trying to improve the method’s ability to describe objects, and so we recognize that a subject for future works will be to use an object recognition system as a module. We applied simple clustering and chunking methods; thus, other segmentation methods or different types of robot action input to the network (e.g., VideoBERT [22]) have the potential to improve scores. Unsupervised learning on generated data on the simulator is also a prospective approach to improve scores. Another subject for future work will be to apply our method to generate robot action sequences when it receives natural language commands from users.

Acknowledgement Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101.

References

1. Bengio Y, Simard P, Frasconi P et al (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Networks 5(2):157–166
2. Bholowalia P, Kumar A (2014) Ebk-means: a clustering technique based on elbow method and k-means in WSN. Int J Comput Appl **105**(9)
3. Chiu CC, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, Kannan A, Weiss RJ, Rao K, Gonina E et al (2018) State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4774–4778
4. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1724–1734
5. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Advances in neural information processing systems, pp 577–585
6. Fasola J, Mataric MJ (2013) Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In: 2013 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 143–150
7. Hartigan JA, Wong MA (1979) Algorithm as 136: a k-means clustering algorithm. J Roy Stat Soc Ser C (Appl Stat) 28(1):100–108 (1979)
8. Hatori J, Kikuchi Y, Kobayashi S, Takahashi K, Tsuboi Y, Unno Y, Ko W, Tan J (2018) Interactively picking real-world objects with unconstrained spoken language instructions. In: 2018 IEEE international conference on robotics and automation (ICRA). IEEE, pp 3774–3781
9. Inamura T, Shibata T, Sena H, Hashimoto T, Kawai N, Miyashita T, Sakurai Y, Shimizu M, Otake M, Hosoda K, et al (2010) Simulator platform that enables social interaction simulation–sigverse: Sociointelligenesis simulator. In: 2010 IEEE/SICE international symposium on system integration. IEEE, pp 212–217
10. Iqbal T. Shah A, Riek L (2018) Toward a real-time activity segmentation method for human-robot teaming. In: Proceedings of the robotics: science and systems (RSS), towards a framework for joint action: what about theory of mind workshop
11. Kimura T, Okugawa M, Oogane K, Ohtsubo Y, Shimizu M, Takahashi T, Tadokoro S (2017) Competition task development for response robot innovation in world robot summit. In: 2017 IEEE international symposium on safety, security and rescue robotics (SSRR). IEEE, pp 129–130
12. Kollar T, Tellex S, Roy D, Roy N (2010) Toward understanding natural language directions. In: Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction. IEEE Press, pp 259–266
13. Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1412–1421
14. Masci J, Meier U, Cireşan D, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. In: International conference on artificial neural networks. Springer, pp 52–59
15. Nakamura T, Iwata K, Nagai T, Mochihashi D, Kobayashi I, Asoh H, Kaneko M (2016) Continuous motion segmentation based on reference point dependent GP-HSMM. In: Proceedings of the IROS workshop on machine learning methods for high-level cognitive capabilities in robotics

16. Nakamura T, Nagai T, Iwahashi N (2009) Grounding of word meanings in multimodal concepts using LDA. In: 2009 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 3943–3948
17. Nakamura T, Nagai T, Mochihashi D, Kobayashi I, Asoh H, Kaneko M (2017) Segmenting continuous motions with hidden semi-Markov models and Gaussian processes. Frontiers Neurorobotics 11:67
18. Neubig G, Nakata Y, Mori S (2011) Pointwise prediction for robust, adaptable Japanese morphological analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers-volume 2. Association for Computational Linguistics, pp 529–533
19. Papineni K, Roukos, S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 311–318
20. Plappert M, Mandery C, Asfour T (2018) Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Rob Autonom Syst 109:13–26
21. Sugiura K, Iwahashi N, Kashioka H, Nakamura S (2011) Learning, generation and recognition of motions by reference-point-dependent probabilistic models. Adv Robot 25(6–7):825–848
22. Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) Videobert: a joint model for video and language representation learning. In: Proceedings ICCV
23. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112
24. Takano W, Nakamura Y (2015) Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. Int J Rob Res 34(10):1314–1328
25. Tellex S, Kollar T, Dickerson S, Walter MR, Banerjee AG, Teller S, Roy N (2011) Understanding natural language commands for robotic navigation and mobile manipulation. In: Twenty-fifth AAAI conference on artificial intelligence
26. Thomason J, Padmakumar A, Sinapov J, Walker N, Jiang Y, Yedidsion H, Hart J, Stone P, Mooney RJ (2019) Improving grounded natural language understanding through human-robot dialog. In: Proceedings ICRA
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
28. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S et al (2017) Tacotron: towards end-to-end speech synthesis. Proc Interspeech 2017:4006–4010
29. Yamada T, Matsunaga H, Ogata T (2018) Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Rob Autom Lett 3(4):3441–3448
30. Yamaguchi U, Saito F, Ikeda K, Yamamoto T (2015) HSR, human support robot as research and development platform. In: The Abstracts of the international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM 2015, vol 6. The Japan Society of Mechanical Engineers, pp 39–40

Towards a Natural Human-Robot Interaction in an Industrial Environment

Ander González-Docasal, Cristina Aceta, Haritz Arzelus, Aitor Álvarez, Izaskun Fernández, and Johan Kildal

Abstract Nowadays, modern industry has adopted robots as part of their processes. In many scenarios, such machines collaborate with humans to perform specific tasks in their same environment or simply guide them in a natural, safe and efficient way. Our approach improves a previously conducted work on a multi-modal human-robot interaction system with different audio acquisition and speech recognition modules for a more natural communication. The semantic interpreter, with the aid of a knowledge manager, parses the resulting transcription and, using contextual information, selects the order that the operator has uttered and sends it to the robot to be executed. This setup is evaluated in a real manufacture scenario in a laboratory environment with a large set of end users both quantitatively and qualitatively. The gathered results reveal that the system behaves robustly and that the assignment was also considered by the end users as manageable, whilst the system in overall was received with a high level of trust and usability.

A. González-Docasal (✉) · H. Arzelus · A. Álvarez
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia – San Sebastián, Spain
e-mail: agonzalezd@vicomtech.org

H. Arzelus
e-mail: harzelus@vicomtech.org

A. Álvarez
e-mail: aalvarez@vicomtech.org

C. Aceta · I. Fernández · J. Kildal
Tekniker, Basque Research and Technology Alliance (BRTA),
Parke Teknologikoa Iñaki Goenaga 5, Eibar, Spain
e-mail: cristina.aceta@tekniker.es

I. Fernández
e-mail: izaskun.fernandez@tekniker.es

J. Kildal
e-mail: johan.kildal@tekniker.es

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_18

1 Introduction

One of the main pending challenges in the so called Industry 4.0 relies on the natural communication between robot and human agents within manufacturing scenarios as a way to increase productivity [16]. With the aid of this human-robot interaction, the machines should be able to perform complex tasks in unstructured environments whilst maintaining a natural interaction between human operators and guaranteeing their safety. This challenge is particularly demanding in situations with noisy background conditions and poor or changing illumination, which are quite common in industrial environments.

This work is presented as an evolution of a previous system for multi modal interaction between humans and robots described in [16]. This system featured vision- and voice-based interaction via speech and gesture recognition for the processing of operators' requests. Additionally, it was complemented by a semantic knowledge database with understanding of the characteristics and capabilities of the robot. In this work, we put the focus firstly on the voice channel, proposing an improvement for a more robust natural interaction interface, with the evaluation of two Automatic Speech Recognition (ASR) systems that feature robustness on noisy audio conditions. Besides, the previous semantic interpreter was enhanced including a more complex analysis of the context of the working space before commanding any instruction to the robot, ensuring the feasibility and coherence of its execution. The whole system was evaluated by speakers with different accents, featuring bilingual users with a pronunciation strongly influenced by the Basque language speaking in Spanish. The recordings were performed using two different devices and in a real manufacturing scenario in laboratory environment.

The rest of the paper is structured as it follows. Section 2 introduces related work. Section 3 specifies the architecture of the proposed system and the detailed descriptions of its main modules. In Sect. 4, a precise evaluation of the overall system and its parts is presented. Section 5 presents a summary of the evaluations and the conclusions that can be drawn from them. Finally, Sect. 6 concludes the paper and presents future work.

2 Related Work

Several approaches in the literature combine multiple sources of information for human-robot interactions, given that they may help overcome the environmental difficulties or detect contradictions in the commands directed to the robot. For instance, the use of image technologies may help ensuring a safe collaboration between robots and humans using, for example, speed and distance measures [3], providing environmental information [14] or identifying elements in the scene.

One of the key elements of a natural language communication between humans and robots in an industrial environment is the ability of the system to understand voice commands. In the work proposed by [12], different commercial alternatives to child speech recognition for interaction between humans and robots are compared. Although the paradigm does not correspond to industrial environments, they focus on the importance of the quality of the transcription in order to achieve a rich communication between both agents. In order to assure the correct interpretation for a given command, natural language technologies are used in several solutions in the literature [3, 16]. The authors in [3] emphasise the role of ontologies, since they define in detail the domain and reduce ambiguity between the agents. As mentioned in [13], it is important to provide robots with the capacity of communicating in natural language in order to generate acceptation from humans.

The previous work [16] presented a multi modal approach for human-robot collaboration in industrial environments, in which speech and gestures were used to ensure a natural and safe communication between the two agents. Similar implementations can be found in other works, such as [9, 14]. In [9], multi modality was achieved in a robotic arm using speech technologies, visual technologies for element situation and natural language processing. The architecture presented in [14] made use of natural language processing (for both spoken and written language) along with 3D visual technologies for object modelling and recognition in order to design an operating system for industrial robots.

In the current work, we improved the interaction process between the human and the robot by (1) enhancing the robustness of the ASR system against noisy audio conditions and (2) improving and extending the previous semantic interpreter with the integration of a more sophisticated analysis of the context before sending an instruction to the robot.

3 Description of the Main System

This work contributed to create a safe work-space environment that features a natural human-robot interaction. The communication between both parts is thus based on voice commands that can be uttered in a way the operator feels natural. Some examples of this interaction could be asking the robot to complete certain task, to stop the execution for manual intervention, or to resume a previously paused task once the human operator has finished.

In order to achieve such communication between the agents, we propose a system consisting in different modules: (1) the *Audio Acquisition* module for gathering the uttered voice commands; (2) the *Automatic Speech Recognition* (ASR) system that transcribes the audio previously collected; (3) the *Semantic Interpreter* that parses the key elements of the transcription and converts them to a representation understandable by the robot with the help of (4) a *Knowledge Manager* (KM); (5) a *Fusion Engine* that fuses the interpretations of the different channels if they exist (out of the scope of this paper) and checks the feasibility and coherence of its execution supported on the real time gathered in the *Knowledge Manager*; and lastly (6) the

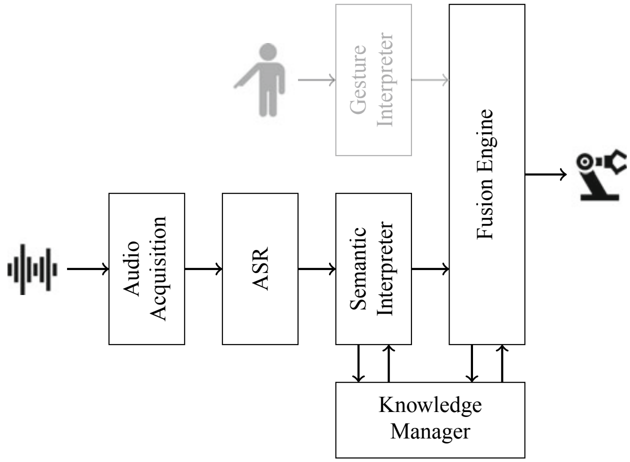


Fig. 1 Diagram of the proposed system. The gesture interpreter is out of the scope of this work.

Robot itself as an agent that will perform the task finally computed. A diagram of the modules is shown in Fig. 1.

In the following subsections, the concatenated modules that compose this human-machine interaction system are described.

3.1 Audio Acquisition

The audio acquisition module was built using two different models of mobile phones, the Xiaomi Mi A2 and the Motorola Moto G5, through an Android application. The output format of the audios was single channel in 3GPP format with AAC codec, a bit rate of 96 Kbps and a sampling rate of 16 kHz.

3.2 Automatic Speech Recognition

Two different ASR systems were integrated and compared, one based on the Kaldi toolkit [21] and the other following the Baidu's Deep Speech 2 architecture [1].

3.2.1 Chain DNN-HMM Model

The hybrid chain Deep Neural Network-Hidden Markov Model (DNN-HMM) was trained with Kaldi, an open-source toolkit for ASR systems construction that provides powerful submodules for feature extraction and acoustic and language modelling based on Finite State Transducers (FSTs). As deep neural net, a Time-Delay Neural

Network (TDNN) was employed [19, 20], which consists of a feed forward network with a lighter computation of the fed context window. The parametrization of the input audio signal was performed using the Mel Frequency Cepstral Coefficients (MFCC) which were augmented by delta- and delta-delta-coefficients and then processed by Linear Discriminant Analysis (LDA). The final feature vector dimension was 40. The model was trained through Maximum Likelihood Linear Transformation (MLLT) [10], Speaker Adaptive Training (SAT) [2] and discriminative non-linear feature transformation [22], methods to improve the recognition accuracy by adapting to unknown and changing noise conditions. The training was enriched with speed- and volume-perturbation techniques as explained in [20].

The training acoustic corpora was composed by a total of 283 h of transcribed audio in Spanish consisting of 132 h from the corpus Savas [23], 21 h from the local Basque television EiTB (Euskal Irrati Telebista), 103 h from national Spanish television RTVE (Radio Televisión Española) from the Albayzin challenge 2018 liberated corpus [15], 20 h from Mozilla's Common Voice project and 7 h from two smaller corpora, Albayzin [7] and Multext [6].

Regarding the language model, it consisted of modified Kneser-Ney smoothed n-grams (with $n = 3$) estimated with the KenLM toolkit [11] that was trained with 20 M words in 860 K sentences from general news domain.

3.2.2 End-to-End Model

The End-to-End (E2E) model was an evolution of the system presented by the authors in the work described in [4]. The model was originally trained using Mel-scale based spectrograms and corresponded to a mixed model trained with both media and telephone audio contents subsequently fine-tuned with the media domain data. This baseline model for Spanish was first built for 25 epochs and a batch size of 20. The fine-tuning was performed for 5 more epochs. In this work, this model was then evolved with data augmentation techniques of Spectrogram Augmentation [18] for 20 more epochs, time-domain dilation of the input signal by a factor of 1.1 and 0.9, and noise injection to the original audio signal from different sources. For this training a batch size of 10 and a learning rate of $5 \cdot 10^{-5}$ annealed by 1.12 was employed.

All these training processes were realised using the 3-fold augmented 132 h in the corpus Savas [23] in both media and telephone domains, summing up a total of 795 h.

The output of the E2E model was then improved through a rescoring process using the same previously specified n-gram (with $n = 5$) as external language model, and the values of $\alpha = 1.5$ and $\beta = 1$ on a beam search of 1 000 for computing the final probability of the recognition.

3.3 Knowledge Manager

The Knowledge Manager (KM) used an ontology (defined in OWL) to model the environment and the robot capabilities as well as the relationships between the elements in the model, which can be understood as implicit rules that the reasoner exploits to infer new information.

The ontology included individuals like `Start`, `Stop`, `Resume` and `Move` belonging to the class `BasicAction` (a subclass of `TemporalThing` class) that were used to define the commands that could activate the different type of the robot's methods like `ExecutionMethod`, `StopMethod` and so on in the scenario. Those individuals included in the tag `dataProperty` the natural expressions that could be used to mention them.

The knowledge manager also included information represented semantically according to the ontology about the current status of the scenario: which operation was being executed, if it involved any robot, if the robot was in the middle of an action or waiting for the start of a program, etc. All this information supported both the semantic text interpretation and the fusion engine to decide the feasibility of a certain action in a specific moment.

3.4 Semantic Text Interpreter

Given as input a human request in which a person indicates the desired action via voice, the purpose of this module was to understand exactly what the person wanted the robot to do and, if the information was complete, to generate the corresponding command for the robot. For such an interpretation, the module followed two main steps: an initial rule-based step for the extraction of key elements from the transcribed text and a second step matching the key elements and the tasks that were feasible for the robot, defined in the KM module.

For the first step, Natural Language Processing (NLP) techniques were used. The main idea was to use syntactic information by means of rules for the extraction of key elements. In this work, the Spanish version of the FreeLing [17] library was used for this structural analysis. With regard to the definition of rules, FreeLing was employed for the morpho-syntactic analysis and dependency parsing of a set of request examples obtained from different users. The complete information was manually revised and the most frequent morpho-syntactic patterns that were relevant for extracting the key elements were afterwards identified. Such patterns were implemented as rules.

Once the key elements were extracted, it was necessary to identify which one of the tasks that the robot was able to perform suited best the request. This last step was overcome by making use of the KM information described above. First, it was required to verify if the identified actions were among the feasible tasks described in the KM, accessing the action's data property tag in the KM using the semantic query language SPARQL. The final output from the semantic text interpreter consisted of various frames, one for each potential robot action candidate.

3.5 Fusion Engine

In the scope of this work, where the focus is on the voice channel and additional inputs have not been considered, the aim of the Fusion Engine is to check the coherence and feasibility of the frames coming from the Semantic Interpreter, when more than one checked in order of relevance, considering the current situation of the scenario, and send the proper action to the robot if it is appropriate.

Given an action, its coherence and feasibility checking in a certain situation is supported by the information in the Knowledge Manager. More specifically, if there is any operation running, the fusion engine consults the Knowledge Manager to check the compatibility between the status (running, stopped, ...) and the request (stop, resume, ...). Otherwise, when there is no operation running, it checks the possible next operation(s) to run. If the request coming from the voice channel is compatible and all the necessary information is available, it sends the action to the robot. When the request does not become an order for the robot due to incompatibilities found in this step, the user is notified why the request has not been successfully sent to the robot (no action identified, incompatibility between the request and the situation ...).

4 Evaluation and Results

An experimentation workshop was scheduled for the developed system evaluation with the participation of students from the IMH Advanced Manufacturing Centre in Elgoibar (Gipuzkoa, Spain), and it was performed in a real manufacturing scenario in laboratory environment, with a high-level of background noise coming from manufacturing machines.

The real manufacturing scenario in laboratory environment (Tekniker's facilities) approached in this work covered a collaborative process where an operator had to work with an industrial robot (see Fig. 2) to assemble a piece. The robot took care of the manipulation of the different parts of the final piece, while the operator screwed both parts. The operator should use natural voice expressions to start the process which triggered a program execution in the robot that manipulated the parts of the final piece. The operator could stop and resume through the same channel, the voice, the activity of the robot whenever considered necessary (some procedure was not completed, something in the environment could trigger a malfunction ...). When the robot concluded the first part of the whole manipulation process, it would notify to the operator through a message that their collaboration was required. Once performed the screwing of the parts, again via voice, the operator asked the robot to continue with the process, until completion.

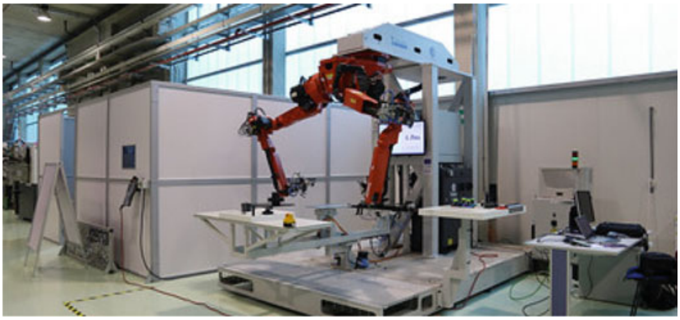


Fig. 2 Industrial collaborative robot used for the evaluation workshop in Tekniker's facilities.

Table 1 WER results for each ASR system and each recording device

	Kaldi	E2E
Xiaomi Mi A2	14.29%	14.52%
Motorola Moto G5	3.13%	8.86%

The group of testers, composed by 25 students plus 2 of the main developers of the system, was composed by 21 males and 6 females, with ages ranging between 21 and 49 years old and an average of 25.4. The volunteers were first introduced to the experimentation task they had to solve using a natural voice interaction, and then divided into two different groups. The first group tested the Kaldi based recognition system and recorded 53 audio utterances (27 with Motorola Moto G5 and 26 with Xiaomi Mi A2), whilst the second group employed the E2E based ASR system and generated 49 audios (23 with Motorola Moto G5 and 26 with Xiaomi Mi A2). The whole system performance was evaluated from both quantitative and qualitative point of views.

4.1 Quantitative Evaluation

In order to measure the performance of both ASR systems, the Word Error Rate (WER) of each transcription was computed, which sums up the insertion, deletion and substitution errors over the total number of words in the reference transcription:

$$\text{WER} = \frac{\text{insertions} + \text{deletions} + \text{substitutions}}{\text{number of reference words}} \quad (1)$$

The WER results are described the following in Table 1 for each ASR system and audio acquisition device.

It is worth noticing that out of the 102 voice inputs processed by the ASR systems, 81 (79.41%) had no transcription errors at all, resulting in an error rate of 0% for these commands.

Table 2 Number of correct, incorrect or missing orders divided by correct and incorrect transcriptions.

	Correct orders	No orders	Incorrect orders
Correct transcriptions	77	3	1
Incorrect transcriptions	12	0	9
Total	89	3	10

Table 3 Number correct, incorrect and missing orders executed by the robot for each recognition output, ASR system and mobile device.

	Kaldi	Correct orders	No orders	Incorrect orders
Moto G5	Correct transcriptions	24	1	0
	Incorrect transcriptions	1	0	1
Xiaomi Mi A2	Correct transcriptions	17	1	1
	Incorrect transcriptions	5	0	2
	E2E	Correct orders	No orders	Incorrect orders
Moto G5	Correct transcriptions	16	1	0
	Incorrect transcriptions	5	0	1
Xiaomi Mi A2	Correct transcriptions	20	0	0
	Incorrect transcriptions	1	0	5

A coherent method to compute the error of the whole system and the interpretation is to evaluate which of the given voice input commands were correctly performed by the robot. To this end, the spoken voice orders were previously manually mapped to the action that the robot should have performed and then compared with the real action performed during the evaluation.

As presented in Table 2, 89 orders (87.25%) were correctly performed by the robot, 10 (9.80%) were incorrect and for 3 of them (2.94%) the system was not able to get any robot action for the given audio input. Examples of these last orders were *vale* (okay, in English) or *más rápido* (faster).

Going more into detail and evaluating the performance of the semantic text interpreter component considering only the correctly transcribed 81 transcriptions, it can be observed that it increases the performance almost in 8 points, interpreting correctly a total of 77 orders (95.06%), and only interpreting incorrectly 1 (1.23%).

Furthermore, the interpreter was able to deliver the correct action for 12 out of the 21 (57.14%) voice inputs with some transcription error.

It is worthy to mention that the users were informed about the workflow of the process before the experimentation, and thus none of the uttered commands was out of the scope of the evaluation procedure. Therefore, the fusion component sent all the existing interpreted actions to be executed by the robot.

Table 3 shows the correct, incorrect and missing orders executed by the robot for each recognition output, ASR system and mobile device.

As it can be observed in Table 3, the best results were reached through the configuration that combined the Kaldi-based chain DNN-HMM model and the Motorola Moto G5 device, which performs much better than the Xiaomi Mi A2 mobile for this recognition system in this evaluation. A total of 24 voice inputs were correctly ordered, whilst only 2 commands were missing or incorrect. In contrast, for the E2E recognition system, both devices performed similar in terms of final correct orders, although better ASR recognition results were obtained using the Xiaomi Mi A2 device.

4.2 Qualitative Evaluation

Once the experimentation was over, each of the 27 volunteers was given a questionnaire with the aim of evaluating the user experience of the system subjectively. First, participants were asked to rate the difficulty of the task in a scale from 1 (very difficult) to 7 (very easy). The average rating from 26 participants was 6.42, with a standard deviation of 0.76, which suggested that the task was considered easy to execute.

Afterwards, participants were asked to evaluate their trust in the collaborative robot through the Charalanbous scale questionnaire [8]. This test consists of 10 questions in a 5-point scale, so that the maximum achievable trust value corresponds to 50 points. Responses from 27 participants threw a trust index of 41.1 points, placing it towards the higher range of the trust scale.

The usability of the system was assessed with the System Usability Scale (SUS) questionnaire [5], consisting also of 10 questions in a 5-point scale. The results revealed a SUS index of 75.93 for the 27 users, which translates into an acceptable usability level of grade B.

Finally, an additional *ad-hoc* survey was filled in by the participants, consisting of 9 questions (again in a 5-point scale) that queried about other aspects that might affect user experience (UX). Results suggested that the subjective experiences obtained were rather positive. Participants most clearly agreed with the following three statements: “The robot did what I ordered” ($M = 4.41$, $SD = 0.80$); “Working with the robot was a good experience” ($M = 4.22$, $SD = 0.80$); “The voice recogniser understood well what I said” ($M = 4.07$, $SD = 1.00$).

5 Discussion

Regarding the ASR module, a more robust behaviour was achieved with the chain TDNN-HMM based architecture using both audio acquisition devices. First, it should be noticed that the TDNN-HMM acoustic model was trained with more varied data that could have contributed to a more robust acoustic model against the noisy background conditions. Besides, it has shown a better performance for the accent variability and the natural language used by the volunteers. It should also be noticed that in an environment with generally short voice inputs, the WER values increase faster than in other scenarios. This is supported by the fact that 81 out of the 102 voice orders had no transcription errors at all.

Concerning the Semantic Text Interpreter, comparing the evaluation carried out with only correctly transcribed audios (95.06%) respect to the evaluation including also incorrectly transcribed texts (87.25%), it can be observed that, as expected, transcription errors influenced the system negatively. However, in those cases where the transcription error did not alter significantly the meaning of the original voice requests, the system was still able to infer the correct interpretation, since the interpreter guessed the right action for the robot for 12 out of the 21 (57.14%) voice inputs with some transcription error.

In terms of user experience based on the completed surveys, the human operators considered that the assignment had a low level of difficulty, with a 6.42 out of 7 points of confidence. Additionally, the system built a high level of trust and a good point of usability among the volunteers.

6 Conclusions and Further Work

This paper presents a robust approach for natural interaction between human and robots in industrial environment as a continuation of a previous work. The audio processing was enhanced by comparing two state-of-the-art ASR systems and two audio acquisition devices. The Semantic Interpreter extracted the information from the transcribed spoken content by means of NLP techniques and a posterior set of syntactical rules, an ontology containing the achievable final actions and information of the context of the scenery. Finally, an order is sent to the robot if it resolves it is coherent and applicable in the current situation. The evaluation performed on the proposed system involves both quantitative and qualitative measurements, discussed on the Sect. 5.

As a future work, a deployment of the system on a real industrial environment is proposed, instead of on a laboratory. The system should be able to behave correctly in a majority of situations as it has been proven in the results, supported by the high acceptance of the users in the evaluation. The system could also be provided a higher level of complexity for a more complete and fluid human-robot communication in a safe environment, such as a complete dialog system with answers from the robot to the user. This could help to raise both the acceptance and the usability scores achieved in this work. Finally, a set of biometric sensors measuring the physiological parameters of the operators during their interaction with the robot could improve the quality of the evaluation for further conclusions on acceptance and comfort of the proposed system.

Acknowledgements This work was supported by the Department of Economic Development and Competitiveness of the Basque Government via the LANGILEOK project.

References

1. Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J, Chrzanowski M, Coates A, Diamos G, Elsen E, Engel J, Fan L, Fougner C, Han T, Hannun A, Jun B, LeGresley P, Lin L, Narang S, Ng A, Ozair S, Prenger R, Raiman J, Satheesh S, Seetapun D, Sengupta S, Wang Y, Wang Z, Wang C, Xiao B, Yogatama D, Zhan J, Zhu Z (2015) Deep speech 2: end-to-end speech recognition in English and Mandarin
2. Anastasakos T, McDonough J, Schwartz R, Makhoul J (1996) A compact model for speaker-adaptive training. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, vol 2. IEEE, pp 1137–1140
3. Antonelli D, Bruno G (2017) Human-robot collaboration using industrial robots. In: 2017 2nd International Conference on Electrical, Automation and Mechanical Engineering (EAME 2017). Atlantis Press
4. Bernath C, Alvarez A, Arzelus H, Martínez CD (2018) Exploring E2E speech recognition systems for new languages. In: IberSPEECH, pp 102–106
5. Brooke J et al (1996) Sus-a quick and dirty usability scale. Usability Eval Ind 189(194):4–7
6. Campione E, Véronis J (1998) A multilingual prosodic database. In: Fifth International Conference on Spoken Language Processing
7. Casacuberta F, Garcia R, Llisterri J, Nadeu C, Pardo J, Rubio A (1991) Development of Spanish corpora for speech research (ALBAYZIN). In: Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy, pp 26–28
8. Charalambous G, Fletcher S, Webb P (2015) The development of a scale to evaluate trust in industrial human-robot collaboration. Int J Soc Robot 8. https://doi.org/10.1007/s12369-015-0333-8
9. Gnjatović M, Tasevski J, Nikolić M, Mišković D, Borovac B, Delić V (2012) Adaptive multimodal interaction with industrial robot. In: 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics. IEEE, pp 329–333
10. Gopinath RA (1998) Maximum likelihood modeling with gaussian distributions for classification. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol 2. IEEE, pp 661–664
11. Heafield K (1998) KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp 187–197
12. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, Senft E, Belpaeme T (2017) Child speech recognition in human-robot interaction: evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp 82–90
13. Kildal J, Fernández I, Lluvia I, Lázaro I, Aceta C, Vidal N, Susperregi L (2019) Evaluating the UX obtained from a service robot that provides ancillary way-finding support in an industrial environment. In: Advances in Manufacturing Technology XXXIII: Proceedings of the 17th International Conference on Manufacturing Research, Incorporating the 34th National Conference on Manufacturing Research, 10–12 September 2019, Queen's University, Belfast, vol 9. IOS Press, p 61
14. Lin Y, Min H, Zhou H, Chen M (2018) A natural language interaction based automatic operating system for industrial robot. In: International Conference on Intelligent Computing. Springer, pp 111–122
15. Lleida E, Ortega A, Miguel A, Bazán-Gil V, Pérez C, Gómez M, de Prada A (2019) Albayzin 2018 evaluation: the iberspeech-RTVE challenge onspeech technologies for spanish broadcast media. Appl Sci 9(24):5412. https://doi.org/10.3390/app9245412
16. Maurtua I, Fernandez I, Tellaeche A, Kildal J, Susperregi L, Ibarguren A, Sierra B (2017) Natural multimodal communication for human-robot collaboration. Int J Adv Robot Syst 14:1–12. https://doi.org/10.1177/1729881417716043

17. Padró L, Stanilovsky E (2012) Freeling 3.0: towards wider multilinguality. In: LREC2012
18. Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV (2019) Specaugment: a simple data augmentation method for automatic speech recognition. Interspeech 2019. https://doi.org/10.21437/interspeech.2019-2680
19. Peddinti V, Chen G, Manohar V, Ko T, Povey D, Khudanpur S (2015) JHU ASpIRE system: robust LVCSR with TDNNS, iVector adaptation and RNN-LMS. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp 539–546. https://doi.org/10.1109/ASRU.2015.7404842
20. Peddinti V, Povey D, Khudanpur S (2015) A time delay neural network architecture for efficient modeling of long temporal contexts. In: INTERSPEECH
21. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No. CFP11SRW-USB
22. Povey D, Kingsbury B, Mangu L, Saon G, Soltau H, Zweig G (2005) fMPE: discriminatively trained features for speech recognition. In: Proceedings.(ICASSP 2005). IEEE International Conference on Acoustics, Speech, and Signal Processing 2005, vol 1. IEEE, pp I–961
23. Pozo A, Aliprandi C, Álvarez A, Mendes C, Neto J, Paulo S, Piccinini N, Raffaelli M (2014) SAVAS: collecting, annotating and sharing audiovisual language resources for automatic subtitling

Nudges with Conversational Agents and Social Robots: A First Experiment with Children at a Primary School

Hugues Ali Mehenni, Sofiya Kobylyanskaya, Ioana Vasilescu, and Laurence Devillers

Abstract This paper presents an experimental protocol during which human interlocutors interact with a dialog system capable to nudge, i.e. to influence through indirect suggestions which can affect the behaviour and the decision making. This first experiment was undertaken upon a population of young children with ages ranging from 5 to 10 years. The experiment was built to acquire video and audio data highlighting the propensity to nudge of automatic agents, whether they are humanoid robots or conversational agents and to point out potential biases human interlocutors may have when conversing with them. Dialogues carried with three types of agents were compared: a conversational agent (Google Home adapted for the experiment), a social robot (Pepper from Softbank Robotics) and a human. 91 French speaking children participated in this first experiment which took place in a private primary school. Dialogues are manually orthographically transcribed and annotated in terms of mental states (emotion, understanding, interest, etc.), affect bursts and language register, which form altogether what we call a user state. We report on an automatic user states detection experiment based on paralinguistic cues in order to build a future automatic nudging system that adapts to the user. First results highlight that the conversational agent and the robot are more influential in nudging children than the human interlocutor.

H. Ali Mehenni (✉) · S. Kobylyanskaya · I. Vasilescu · L. Devillers
CNRS, LIMSI, Paris-Saclay University, Saint-Aubin, France
e-mail: alimehenni@limsi.fr

S. Kobylyanskaya
e-mail: skobyl@limsi.fr

I. Vasilescu
e-mail: ioana@limsi.fr

L. Devillers
e-mail: devil@limsi.fr

L. Devillers
Sorbonne University, Paris, France

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_19

1 Introduction

The increasing usage of conversational robots in many everyday situations raises the question of their influence on humans [13]. However, this influence is hard to measure and often disregarded. This paper addresses the issue of the influence of the conversational agents on human users, and focus on "nudges": indirect suggestions which can affect the behaviour and the decision making. The notion of "nudging" first came to light in 2008, proposed by Thaler (Nobel Prize in Behavioural Economy, Nov 2017) and Sunstein [15]. They stressed the fact that "nudging" was a tactic to subtly modify a person's behaviour, without restricting that person's choice. Indeed, nudging mainly operates through the affective system or by exploiting common cognitive bias (e.g. attention, memory, laziness).[1]

Nudges could have a large impact on society, both negative and positive. On the one hand, they pose a threat to privacy [1] since people can be incited to leak their personal information. On the other hand, they could be used to improve efficiently and smoothly a vast number of tasks from diverse fields, e.g. education (attention, memory), transportation, health care. For now, "nudging" as a research topic has been covered mostly in behavioural economics [15].

The long term goal of this work is both to build an automatic dialog system able to nudge and to measure the influence of nudges exerted by conversational agents and robots on humans in order to raise the awareness of their use or misuse and open an ethical reflection on their consequences. The experiment was undertaken as part of the project "Bad Nudge Bad Robot" focusing on the modeling of nudging strategies within a spoken dialogue, funded by the DATAIA institute.[2] This project is part of the AI chair HUMAAINE: HUman-MAchine Affective Interaction & Ethics (headed by L. Devillers) at CNRS and DATAIA. The chair team is composed of researchers in computer science, linguists and behavioral economists from Paris-Saclay University.

Here nudges are considered within a new research paradigm, the human vs conversational robots and/or agents verbal interaction. Data are collected trough a dialog protocol built to convey nudges. The aim of the experiment is to determine whether automatic learning allows measuring the influence of verbal nudges on a human engaged in a dialog with an automatic system.

2 Experimental Framework and Methodology

The experiment was conducted in a private primary school in early July 2019 as part of a multidisciplinary project involving researchers in Spoken Language Processing at the LIMSI-CNRS laboratory and in economy (RITM, Paris-Sud/Paris-Saclay University).

---

[1] https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18.

[2] https://dataia.eu/.

The rationale behind this experiment is that human behaviour may be subject to cognitive bias when the interlocutor is a robot or a conversational agent. Indeed, humans tend to anthropomorphize machines [4] and to project emotions on them. This experiment also targets a sensitive population, children, who are more likely to be influenced. However, a preliminary experiment involving adults pointed out that they can also be influenced by social robots [5].

2.1 Experimental Design

The design of the experiment consisted of a child interacting with a conversational partner for approximately 5–10 min. Children's age ranged from 5 to 10 (that is all the primary school levels). To observe the bias a child may display towards robots, volunteers were equally divided into 3 groups, corresponding to a balanced distribution in terms of age and gender. Each group was paired with a different conversational partner: (1) with a humanoid robot (Pepper from Softbank Robotics), (2) with a speaker (Google Home, adapted to the task) and (3) with a human (a PhD student participating in the project, aware of the experiment but working on a different topic than human-machine dialog) (Fig. 1).

The experiment was conducted towards a Wizard of Oz (Woz) procedure. The dialogue was scripted and both the robot and the Google Home speaker were manipulated by a researcher during the interaction with the children, unbeknownst to the latter.

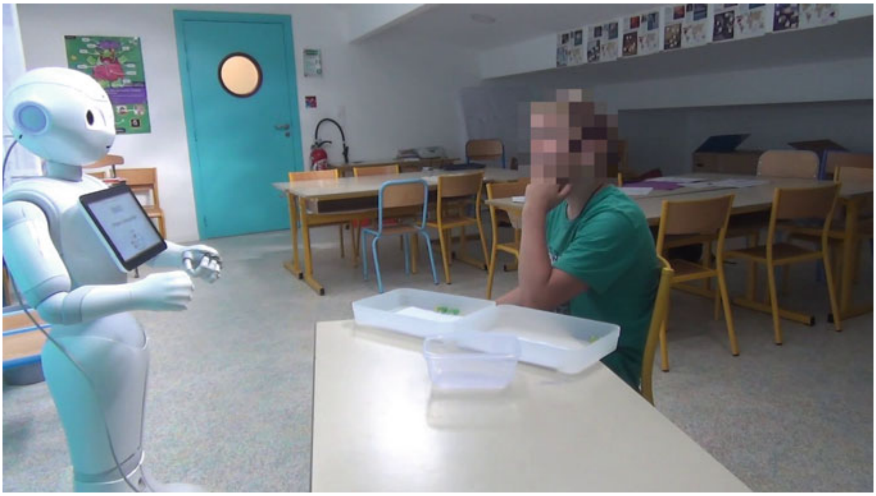


Fig. 1 Child volunteer engaged in a conversation with the robot Pepper

2.2 Structure of the Experiment

The experiment is divided into 3 parts. The rationale behind this structure is to fits several analytic dimensions with respect to the various issues concerning nudges, that is behavioural economics, the propensity of a dialog system to implement nudging strategies and the dialog and emotional specificity of a population of children, known both as vulnerable and increasingly confronted to such systems. Prior to the experiment we requested and obtained the official approval of the ethical comity of Paris-Saclay University.

The three parts are as following:

1. Dictator Game, adapted to children. The game is well-known within the Game Theory [2] and is intended to measure altruism. In the present case, we give to a child a certain amount of marbles (for adults we would use money), and ask him to choose how much he wants to keep for himself and how much he is willing to give to another person (in the current configuration, another classmate, not specifically mentioned). We then try to influence the decision by using anchoring techniques (e.g. peer-effect strategies).
2. Open question, testing the amount of the confidence a child attributes to the discussion partner.
3. Quiz addressed to the children. The selected topic was video games and we investigated in advance that children would held the questions concerning some specific games. Specific dialog strategies (e.g. question repeated at several speaker turn distance) were employed during the quiz, to measure the attention and estimate which nudging strategies are the most efficient to raise children's trust and potentially affection towards a robot.

The experiment was followed by a discussion with all the children, during which we explained how the robot worked in an understandable way. The aim of this last action was to sensitize the young population about the potential misuse of nudges.

In the following sections we will describe the results of the experiments according to the 3 steps described above and propose a first detection system using machine learning techniques and paralinguistic cues associated to reactions (mental states, affect bursts, choice of language register) to nudging strategies. The section bellow focuses (Sect. 3) on a short description of the Dictator game experiment. Although this experiment implies few spoken parts, it provides a first overview of the children behavior as a function of the interlocutor employing nudging techniques. Section 4 focuses on the description of the corpus, the annotation strategy and the kappa calculation. Results are displayed in Sect. 5 and concern both automatic detection of mental states in a broad sense (including also language register and affect bursts), based on paralinguistic features, and the linguistic description of the correlation between nonverbal information (here filled pauses) and annotation labels. We finally propose a discussion and further work objectives in Sect. 6.

3 Towards Quantifying Nudging Strategies: The Dictator Game

The first part of our experiment, the Dictator game, enabled us to get concrete and quantifiable data on how much children were influenced by their interlocutor (cf Table 5). Although verbal data remains limited during this part of the experiment, the results provide insights on the patterns of interaction as a function of the type of interlocutor (robot, conversational agent, human). To start, children were invited to make a first decision on how they would like to distribute a fixed number of marbles. They were then subjected to 2 successive nudges using 2 different anchoring techniques (peer-effect and first-person strategies). We measured on average an altruism of about 45% during the experiment. About 50% of the children were influenced by their interlocutor and changed their initial choice during each nudging attempt. Furthermore, during the first and second nudging attempt, children were more influenced by non-human interlocutors (Google Home speaker or robot) than by the human one (cf Table 5). Indeed, among the children who changed their choice during the second nudge, 16% interacted with the adult whereas 40% and 44% respectively interacted with the robot and the Google Home speaker. This result is consistent with the hypothesis that a non-human interlocutor is more likely to influence. However, in order to generalize the observation, additional data would be necessary and should cover other groups of children as well as older volunteers (adults, elderly) (Table 1).

In order to quantify the nudging effects during the dictator game, a simple metric has been retained that consists of measuring how much (in terms of marbles) the child's choice got closer to the value of the nudge given by the interlocutor.

Nudge metric = (Difference in absolute value between the result before the nudge and the value given during the nudge) − (Difference in absolute value between the result after the nudge and the value given during the nudge)

Although, this metric gives an estimation of the number of marbles modified by a child and the direction of the change (positive or negative result of the metric), it does not show whether the child completely complied with the value given during the nudge or only approached it. To take this aspect into account, the last nudge metric was divided by the difference in absolute value between the result before the nudge and the value given during the nudge.

Normalized Nudge metric = Nudge metric/(Difference in absolute value between the result before the nudge and the value given during the nudge)

The metrics underlined that children are more influenced by the robot and the Google Home speaker than by the adult. During the game, different children behaviours can be observed: amusement, nervousness, doubts, etc. This led us to focus on these aspects in Sect. 5.

Table 1 Data collected during the Dictator game

Nb of children (Age: 5–10)	Adult: 31	Robot: 29	Speaker: 31	Total: 91
Altruism Marbles (mean)	At first: 4.583	1st nudge: 4.7	2nd nudge: 4.427	Mean: 4.57
Nudged children 1st nudge	Adult: 15	Robot: 12	Speaker: 19	Total: 46
Nudged children 2nd nudge	Adult: 7	Robot: 17	Speaker: 19	Total: 43
Nudge metric 1st nudge	Adult: 0.68	Robot: 1.10	Speaker: 1.26	Total: 1.01
Normalized Nudge metric 1st nudge	Adult: 0.12	Robot: 0.30	Speaker: 0.26	Total: 0.23
Nudge metric 2nd nudge	Adult: 0.97	Robot: 1.90	Speaker: 1.48	Total: 1.44
Normalized nudge metric 2nd nudge	Adult: 0.19	Robot: 0.43	Speaker: 0.46	Total: 0.35

4 Corpus

This section focuses on data description, annotation strategy and agreement between the annotators.

4.1 Data Description

The experiment was conducted with 91 children, divided into 3 groups: one interacted with the social robot Pepper, one with the Google Home speaker and a third one with the human researcher. Table 2 below sumps up the distribution of the participants across the three configurations.

4.2 Annotations

The corpus has been annotated with different labels at the speaker turn level in order to further correlate the characteristics of the user state at the turn and paralinguistic features. The labels are used to train classifiers with machine learning techniques in order to automatically detect the audio information (in this experiment, paralinguistic) that would help the dialogue system to better assess the state of the user. Plutchik's

Table 2 Data description

Number of children	91
Number of male children	53
Number of female children	38
Number children with the Robot	29
Number children with the Speaker	31
Number children with the Human	31
Age of the children	5–10
Mean length of dialogues	8 min 10 s
Total corpus duration	12 h

wheel [12] of emotions served as reference and the following six basic emotions were selected: amusement, respect, surprise, irritation, nervousness and intimidation (the 2 last labels describing two different low levels of fear). Furthermore, meta-labels opposed positive to negative emotions, in order to make the classification task easier for the models and to overcome limited data for some classes. "Attention" or "Interest" was also retained as label within the annotation system as such labels may help assessing whether a person is likely to be influenced or not. The label was associated to 2 classes: "Interest" and "Disinterest" (which could be linked to boredom, another emotion on Plutchik's wheel). Along the same lines, we then considered two additional annotation labels involving binary classes allowing to describe the mental state of the child: "Confidence"/"Doubt" and "Understanding"/"Confusion". Therefore, in this annotation system, mental states broadly designate emotion labels and labels relative to "Doubt", "Interest" and "Understanding".

We also annotated affect bursts: "Laughter", "Hesitation" and "Breath"; although, the number of chunks annotated with the label "Breath" are few and do not allow a machine learning classification.

We added two language register labels "Polite"/"Colloquial" as well, so as to further correlate the level of language children would adopt as a function of the interlocutor.

Finally, the label "Other" mainly contains neutral chunks and serves as a default class for the classifiers. Table 3 bellow sums up the annotation labels and the number of speech chunks per label.

The corpus data was annotated (cf Fig. 2) with the software ELAN[3] and the audio was extracted using Praat[4] scripts. The corpus also benefited from an orthographic manual transcription.

[3] https://tla.mpi.nl/tools/tla-tools/elan/.

[4] https://praat.fr.softonic.com/.

Table 3 Description of the annotations and number of speech chunks per class (91 dialogs)

Annotations of mental states, affect bursts and language registers	Number
Positive Emotions (amused, surprised, respectful)	354
Negative Emotions (irritated, nervous, intimidated)	59
Confidence	622
Doubt	286
Interested	330
Disinterested	49
Understanding	68
Confusion	101
Polite	49
Colloquial	117
Hesitation	273
Laughter	52

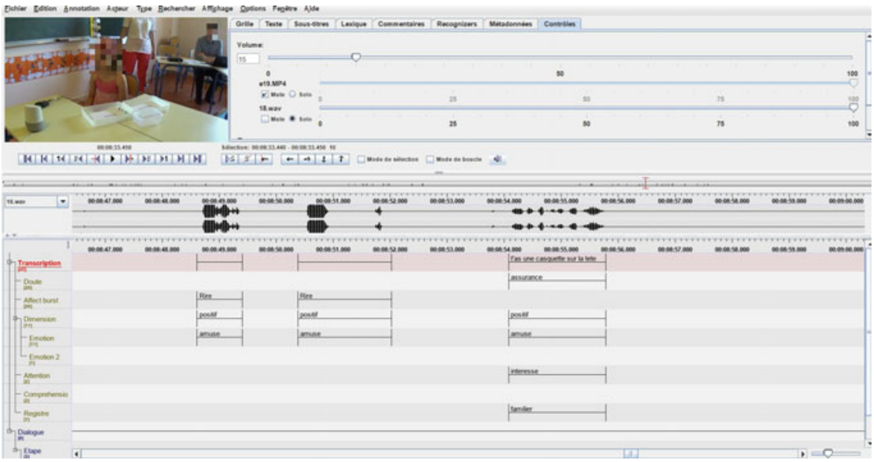


Fig. 2 Annotations of the user state (mental states, affect bursts, language register) with ELAN

4.3 *Kappa: Inter-rater Agreement*

In order to measure the quality of the annotations, we use a control protocol: 5 files of each category (15 files in total) were annotated by two annotators then the Cohen's kappa coefficient was calculated. At this point, the emotion label contained also an "Indeterminate" class to be sure to take into account every potentially unclear emotions. We obtained 0.76 of agreement for the emotion and 0.68 for the doubt labels. Both metrics correspond to a substantial level of agreement. Divergences concern the number of segments annotated by each annotator. To compute the coefficients, we took into account only the segments which were annotated by both annotators.

Table 4 Annotated segments

Annotated by A1 and A2:	Annotated by A1 but not A2:	Annotated by A2 but not A1:	Total
134	205	73	412
0.33	0.49	0.18	1.0

Table 5 Annotation divergences

posit./negat.	posit./indet.	negat./indet.	intimidated/nervous
0.06	0.22	0.20	0.38

We considered that if one segment was annotated by one annotator but was left without annotation by the other, the emotion of this segment was not sufficiently well marked. So, these segments could not have a strong influence on the performance of the further automatic classifiers (Table 4).

Few divergences were observed for opposite categories, for example for negative and positive dimensions. Most of the differences in annotations were observed for pairs such as indeterminate/negative, indeterminate/positive, intimidated/nervous which are more likely to be confused, and may strongly depend on the cultural and personal backgrounds of the annotators. In our future data collection, we will reproduce this control protocol several times during the annotation phase in order to improve the inter-annotator agreement (Table 5).

5 Results: Automatic Detection of Mental States, Affect Bursts and Language Register Based on Paralinguistic Features and Linguistic Analysis of Non-verbal Information

This section focuses on the automatic detection of the user state based on paralinguistic features and on the linguistic analysis of this non-verbal information. Recall that the main focus of the experiment is to automatically detect and classify mental states (emotion, Understanding, Doubt, Interest), affect bursts and language registers through paralinguistic information. The scores obtained from the automatic classifiers are presented in the Sect. 5.1, whereas the Sect. 5.2 focuses on the linguistic description of the correlation between non-verbal information (here filled pauses) and mental states.

Table 6 Mean F1-scores for mental states, affect bursts and language registers classifiers (10-fold nested cross validation) with the emobase2010 features (1581 feat.)

Classes	SVM	R.F.
Positive emotions/Negative emo./Other	**0.56** ± 0.01	0.45 ± 0.03
Interested/Disinterested/Other	**0.56** ± 0.01	0.49 ± 0.01
Doubt/Confidence/Other	**0.59** ± 0.01	0.48 ± 0.01
Understanding/Not Understanding	0.65 ± 0.04	**0.66** ± 0.02
Polite/Colloquial	**0.82** ± 0.01	0.79 ± 0.03
Hesitation/Laughter/Other	**0.71** ± 0.01	0.64 ± 0.01

5.1 Paralinguistic Detection of the User State

This section focuses on the detection of mental states, affect bursts and language register based on paralinguistic information as part of a future automatic dialogue system (which is further described Sect. 6). Paralinguistic studies about emotion with SVM (Support Vector Machine) or other conventional approaches have been conducted in previous works and a specific interest was given to minimalistic acoustic parameter sets, e.g. GeMAPS (the Geneva minimalistic acoustic parameter set) for voice research and affective computing [8] or robust small sets [14]. Here paralinguistic parameter sets were also used to implement classifiers with speech data specific to nudges.

We used the software OpenSMILE[5] to extract relevant features in order to train the models (emobase2010 features of OpenSMILE [9]). The results are provided in Table 6. The models implemented were SVM (Support Vector Machine) and Random Forest.

Each classifier had 2 or 3 classes to classify. The results given for each set of classes are the mean F1-scores (with standard deviation), which evaluate how well a classifier discriminates one class from another. The scores were computed with nested cross validation and the data for each class was weighted in the classification models so as to take into account the data imbalance.

SVM classifiers seem to produce better results than Random Forest classifiers with our dataset.

These results show that some pairs/triplets of "states" are more easily distinguishable than others on the basis of the paralinguistic features. Positive and negative emotions are indeed harder to classify than affect bursts such as hesitation and laughter. The small size of the dataset may explain the relatively high variances of some scores and the classifiers would probably benefit from more data. Another constraint the classifiers had to cope with was the noise in some speech chunks because of the noisy environment of the school. This brings some robustness to the models and more realistic scores.

[5] https://www.audeering.com/opensmile/.

Given the performance of the classifiers, they will be used as a weighted input for the dialogue manager rather than as a discrete input. A vector gathering all of their predictions in real time will then give relevant information about the state of the user to a reinforcement algorithm in the dialogue manager (further discussed Sect. 6).

5.2 Contribution of Non-verbal Information for Mental States Characterisation

This section focuses on the role of non-verbal information such as filled (e.g. in French "euh" and "bah") and silent pauses in mental states characterization. 15 dialogues are considered for the analysis (5 dialogues for each pair infant vs robot/Google Home speaker/human) in order to estimate if the non-verbal information can provide reliable cues about the speaker's state which can further be correlated to the prediction of the classifier. Filled pauses are key elements in dialog construction and management [3, 7]. They carry also salient paralinguistic information and can be language dependent. Besides, it has been shown that the position of a vocalic hesitation within the speaker turn can be correlated with various functions such as keeping the floor, introducing new information or manifesting the intention to close the dialog [16]. Here the filled pauses are considered in three positions within the speaker turn, that is initial, internal and final. As for the silent pauses, we consider so far the speaker turn internal ones. Indeed, the pauses observed at the beginning of a speaker turn can be decoded as moments of latency between the question of the human, robot or Google Home speaker and the child's response and consequently, were analysed separately. The occurrence of filled (hesitations) and empty pauses is considered as functions of different pairs of mental states, that are Doubt/Confidence, Positive/Negative emotions, Intimidated/Nervous negative emotional state, Interested/Disinterested (Table 7). Although previous research pointed out the correlation between hesitations and negative or at least non-neutral mental states [6], the current analysis does not point out a strong correlation between pauses and non-neutral states such as doubt.

The number of hesitations and silent pauses corresponding to speech chunks labeled as "Confidence", "Positive emotions", "Intimidated" and "Interested" is superior to the remaining labels. This observation applies for all the three groups (human, robot, Google Home speaker). The label "Intimidated" which roughly corresponds to a discomfort or stressful attitude of the children in front of the agent, and the quantity of hesitations and pauses are correlated. In the observed examples, if the child seems intimidated, the amount of pauses and hesitations produced increases, however the label "Nervous" does not seem to elicit an increasing number of such vocal items.

Finally, the mean duration of pauses and hesitations (Table 8) corresponding to the speech excerpts labeled "Doubt" are in most of the cases superior to those labeled

Table 7 Number of pauses and hesitations introducing mental states

	Human		Robot		Speaker	
	hesit.	pauses	hesit.	pauses	hesit.	pauses
conf.	46%	50%	23%	23%	50%	60%
doubt	17%	36%	19%	13%	16%	16%
pos.	23%	–	10 %	3%	36%	32%
neg.	–	–	–	3%	–	–
intimid.	20%	–	6%	36%	7%	20%
nervous	–	–	–	–	–	–
interest.	43	36%	26%	44%	41%	32%
disinter.	–	–	10%	5%	2%	4%

Table 8 Mean duration of hesitations and pauses (in s)

	Human		Robot		Speaker	
	hesit.	pauses	hesit.	pauses	hesit.	pauses
conf.	0.86	10.41	0.72	1.53	0.85	2.13
doubt	1.11	13.62	0.80	3.28	1.16	1.78

"Confidence", despite the fact that they are more frequently observed. This observation is true for the interactions with a human and a robot.

The preliminary results above point out the potential correlation between the incidence of non-verbal information and user states and it is promising for further integration within the automatic detection experiments.

6 Towards an Automated Nudging Dialogue System

The next step of this project is then to build an automated dialogue system. The user state detection (described in Sect. 5.1) will thus be used in the Spoken Language Understanding (SLU) part of the dialogue system. It will be built as modular parts and will be coupled with semantics for a better understanding. The different classifiers trained and explained in Sect. 5.1 will be fed with speech chunks during new conversations and produce in real time a probabilistic distribution for each set of classes. These predictions will be then gathered in a vector, which will be given as input for the dialogue manager.

For the dialogue manager model, we are working on a POMDP-based architecture [17] with online learning. Reinforcement learning algorithms will be used in order for our agent to adapt to its interlocutor and learn the most efficient nudging strategies for him. We will thus use a sample-efficient algorithm such as the Kalman Temporal Differences model (KTD) [11], whose advantages were explained for instance in E. Ferreira's thesis [10].

7 Conclusion

The study presented in this paper is part of a larger project whose aim is to measure the influence of nudges exerted by conversational agents and robots on humans, in order to raise the awareness of their use or misuse and to open an ethical reflection on the consequences. Two underlying objectives are to design an automatic dialog system able to nudge and to evaluate its feasibility in realistic conditions. The present paper focuses on these 2 objectives: we describe a preliminary Wizard of Oz experiment built for collecting a corpus of dialogues in real-life situation and a detection system based on paralinguistic features. The corpus consists of dialogues between three types of interlocutors (the robot Pepper, a Google Home speaker used as a conversational agent and a human) and children, recorded in a primary school. The rationale behind this choice is that children are a sensitive population and that they are assigned to increasingly interact which such tools. 91 children with ages ranging from 5 to 10 participated in the experiment. Firstly, a Dictator game is proposed to the volunteers under the form of a marble game. To quantify the nudging effects during the Dictator game, a simple metric is retained that consists of measuring the difference between the value of the nudge proposed by the interlocutor (as amount of marbles) and the choice of the children (the effective amount of marbles selected by the children). Thereafter, the corpus collected is annotated in terms of emotions, attitudes, affect bursts, disfluences and language register in order to provide an input for the detection system aimed to automatically identify mental states in a broad sense, correlated to nudging strategies. Moreover, the contribution of the non-verbal information (filled ans silent pauses) for user states characterisation is estimated and seems a promising lead for improving the automatic detection. The first results highlight that the conversational agent and the robot are more influential in nudging children than the human interlocutor.

The next step of this project is to build an automated dialogue system able to nudge. Future work is also focusing on collecting additional data from both children and other sensitive (elderly) and non sensitive populations. The results will feed into a more general reflection on the nudges and the ethical issues they raise within the activities of the AI research chair HUMAAINE.

Acknowledgements The study was funded by the DATAIA project “Bad Nudge, Bad Robot”.

References

1. Acquisti A et al (2017) Nudges for privacy and security: understanding and assisting users’ choices online. ACM Comput Surv 50(3):1–41 Article 44
2. Benenson JF, Pascoe J, Radmore N (2007) Children’s altruistic behavior in the dictator game. Evol Hum Behav 28(3):168–175
3. Clark HH, Fox Tree JE (2002) Using uh and um in spontaneous speaking. Cognition 84:73–111
4. Reeves B, Nass C (1996) The media equation

5. Devillers L et al (2015) Inference of human beings' emotional states from speech in human-robot interactions. Int J Social Robot 7(4):451–463
6. Devillers L, Vasilescu I, Vidrascu L (2004) Anger versus Fear detection in recorded conversations. In: Proceedings of speech prosody
7. Duez D (2001) Signification des hésitations dans la production et la perception de la parole spontanée. Parole 17/19:113–137
8. Eyben F, Scherer R, Schuller B, Sundberg J, André E, Busso C, Devillers L, Epps J, Laukka P, Narayanan S, Truong K (2015) The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202
9. Eyben F, Weninger F, Gross F, Schuller B (2013) Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the ACM multimedia (MM), Barcelona, Spain, October 2013, pp 835–838. ACM. https://doi.org/10.1145/2502081.2502224. ISBN 978-1-4503-2404-5
10. Ferreira E (2015) Apprentissage automatique en ligne pour un dialogue homme-machine situé. Université d'Avignon
11. Geist M, Pietquin O (2010) Kalman temporal differences. Artif Intell Res 39(1):483–532
12. Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Emotion: theory, research, and experience, vol 1. Theories of emotion. Academic, New York, pp 3–33
13. Sciuto A, Saini A, Forlizzi J, Hong JI (2018) Hey Alexa, what's up? In: Proceedings of the 2018 on designing interactive systems conference 2018 – DIS 2018. https://doi.org/10.1145/3196709.3196772
14. Tahon M, Devillers L (2015) Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE/ACM Trans Audio Speech Lang Process 24(1):16–28
15. Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
16. Vasilescu I, Rosset S, Adda-Decker M (2010) On the role of discourse markers in interactive spoken question answering systems. In: Proceedings of the seventh international conference on language resources and evaluation
17. Young S et al (2013) POMDP-based statistical spoken dialogue systems: a review. IEEE

Exploring Boundaries Among Interactive Robots and Humans

Kristiina Jokinen

Abstract Research and development of social robots have rapidly increased in recent years, and it is expected that usefulness of such agents in society will expand for various tasks, for which embodied natural interaction can make an important difference in interfacing humans and AI applications. This paper discusses challenges concerning the robot agent's skills and knowledge in the development of interactive robot technology, focusing on a novel conceptualization of robot agents which cross boundaries from computational machines to human-like social agents and from human-controlled tools to autonomous co-workers.

1 Interactions and Robot Agents in Future Society

Dialogue interactions with robot agents conducted in natural language are expected to become more common in the coming years, as various applications to assist humans in their everyday life will include dialogue capabilities: speaking agents can e.g. provide useful information, chat about interesting topics, and give instructions for routine everyday tasks. Expectations and research goals for such robots include alignment with the user's intents and emotional needs and dealing with sensitive information in a socially tactful and trustworthy manner. Interactions can also take place in smart environments with other smart devices, and dialogue context is thus extended from face-to-face environments to intangible ones.

The future society can be envisaged in the concept of Society 5.0 [7] which assumes that economic advancement should be balanced with the aim to find solutions to social problems by novel technology that integrates cyberspace and physical space. In such a society, symbiotic relation between humans and robots is important, and the development of such interaction requires the robot's capability through learning and interaction as a basis for further co-evolution.

K. Jokinen (✉)
AI Research Center, AIST Tokyo Waterfront, Tokyo, Japan
e-mail: kristiina.jokinen@aist.go.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_20

To better understand various changes in society, and the needs for human well-being, it is important to reflect how the capabilities and characteristics of current robot agents *can* shape the world and our reality (skills) and how such agents *should* shape the future societies and services (needs). The first topic is related to the robot skills enabled by the current technology, while the second one concerns expectations and potential needs for social robots in various practical tasks and situations. The design and development of applications always include interactions between capabilities and needs, and it is crucial to recognize such interactions also take place in social robotics. In fact, in order to address smooth and sustainable development of society and wellbeing of humans, policy making on the benefits and challenges of social robots requires understanding of the interactions between technological advances and visions of symbiotic relations between humans and robots, besides knowledge and experience of the existing applications.

In this paper, interactions between robot skills and expectations are discussed from the point of view of the robot agent's role in society and the boundaries that lie between generally accepted practices on one hand and the challenges faced by the changing society on the other hand. The focus is on situated social robots supporting various everyday tasks at home and at work, and their ability to cross traditional boundaries that are related to agenthood and companionship.

2 Situated Context-Aware Robot Agents

Dialogue interactions with social robots are *situated* [1]. The robot is situated in its own embodied communication environment, and it processes interaction by integrating action plans and language understanding so that its responses can include both speech acts and physical actions. Interactions also take place in a dynamically changing world that cannot be totally specified in advance.

Social robot interactions thus differ qualitatively from the typical screen-based interfaces because of the robot's embodied presence in the situation and its ability to move and perceive the environment independently without explicit human presence. Moreover, the robot can extend human physical capability by reaching locations that are unknown or inaccessible for the human (e.g. rescue robots, or shop assistant robots which can reach very high or low shelves), and by being able to receive information through communication with other devices (smart environments) and by processing large amounts of data efficiently (big data processing).

However, appropriate actions presuppose that the robot agent is aware of its context and able to create common ground from the observations. Besides recognizing the partner's presence and communicative intentions, the agent must learn to disambiguate and categorize its perceptions as well as plan its actions related to the current task. For the design and development of context-aware social robots, interaction models concerning the robot's contact, perception, understanding, and reaction are thus important.

Further challenges concern cooperation and collaboration with humans, and interleaving task actions with speech acts. Architectures that would capture relations between dialogue modelling, motor control, and action are still insufficiently addressed although more research is directed towards integrating action and dialogue models in mobile robot platforms. A crucial aspect for designing integrated models is to understand how speech and action can be linked in the social agent's internal model of the interaction and in its construction of joint actions and a common ground with the partner in the interactive situation as a whole.

In cognitive science, a new paradigm is being developed as *4E Cognition*, i.e. cognition which is embodied, embedded, enactive and extended [5]. The view is based on recent developments in neuroscience and psychology and regards cognition as being formed and structured through dynamic interactions between the brain, body, and the environment. The insights are also applicable to social robotics to conceptualise the linking of embedded cognition, language and the society

As mentioned in Sect. 1, interaction between technological skills and societal needs is crucial in the development of social robots. Similar principles of interaction that are observed in the individual learning process for cognitive skill development can also play a vital role when considering possibilities of social robots for new services in the future society. Consequently, we can say that, through such interactions a new type of robot agent, a Boundary Crossing Robot, will emerge.

3 Boundaries to Cross

Previous work presented the notion of Boundary Crossing Robot (BCR) [2], which provides an efficient conceptualisation of the development of social robot technology. The concept is based on the assumption that social robot applications modify our views of the world by introducing revisions of the traditional categorisation of what robot agents can do and what is their role in society. The dual characteristics of the robots as sophisticated computer tools and simultaneously as communicating agents [1] makes BCRs as facilitators for the humans to interface with AI technology and to re-categorize the environment by suggesting new ways and novel services to do practical tasks.

Context-aware robot agents resemble human agents and challenge prototypical categorisations of the world into agents and non-agents. Although real interactions with present-day robot agents quickly reveal that the partner lacks full conversational capabilities, they elicit ontological questions related to the robot's agency, intention, volition, and responsibility. The robots' human-like appearance with the perceived autonomous behaviour also aids their categorisation as agents. The uncomfortable feeling (the Uncanny Valley) that people often describe when they encounter geminoid robots for the first time, can be explained by the challenge that interactive human-like robots pose to the prototypical cognitive category of an interactive agent. Eeriness is part of the uncertainty and fear related to the autonomous robot agent crossing the boundary from common to unknown, cf. [4].

The robot agents also tend to cross boundaries at workplaces where they cross the boundary between traditional tools and co-workers. Instead of being a tool controlled and governed by the human user for a particular task, the robot agent is invited to exist as a co-worker and a member of the work environment (cf. robot receptionists at hotels and restaurants). Although such a robot co-worker may not exhibit capability for full natural conversations, it still can extend the notion of a co-worker into a useful and reliable companion which does not get tired or annoyed at monotonous work.

To assess a social robot's acceptability, functionality, and usability for service domains, it is useful to study various types of robot applications. For instance, due to their less social appearance and evaluative behaviour, robot agents are reported to elicit more personal engagement in human partners, such as sharing personal experiences and "true" opinions. The robot's simplified capability to communicate can also support therapy sessions or everyday situations for users who encounter problems with the abundant and rich communication style typical for human interactions. Cultural habits and traditions also affect the way in which people are ready to accept social robots, e.g. willingness to accept recommendations from a robot or expectations on sharing of emotional or personal involvement in professional interactions (see [6]).

Finally, safe and ethical data management also becomes a pertinent issue [3]. The robot agent should provide truthful and reliable information, and its training should be based on unbiased data models. In smart environments with interconnected devices, the robot needs knowledge of the situation, participants, and appropriate behaviour patterns, especially when delivering personal information. The context is to be carefully assessed while the appropriate behaviour may require understanding of the balance between safety and privacy.

4 Co-creation of Interactive Innovations

Innovative services are co-created in interaction with novel technologies by the users. The development of interactive social robots can be conceptualised into the notion of Boundary Crossing Robot, which can facilitate interaction and mutual intelligibility between different perspectives. Such an interactive robot agent is not just a tool but an agent to communicate with, and it can embody interaction that is sensitive to social aspects of communication: the partner's intentions, needs and affective state, natural language interaction, and cooperation on tasks which require coordination and planning of speech and motion. Design and development of such robot agents require integration of complex technology that takes into account the necessary skills and needs for the situation and the evaluation of the expectations and experience of the users.

References

1. Jokinen K (2018) Dialogue models for socially intelligent robots. In: Ge SS, Cabibihan J-J, Salichs MA, Broadbent E, He H, Wagner A, Castro-González Á (eds) Social robotics. 10th International conference, ICSR 2018, Qingdao, China, November 28–30, 2018, Proceedings, Springer LNCS 11357, pp 127–138
2. Jokinen K, Watanabe K (2019) Boundary Crossing Robots: societal impact of interactions with socially capable autonomous agents. In: Salichs MA, Ge SS, Barakova EI, Cabibihan J-J, Wagner AR, Castro-González Á, He H (eds) Social Robotics. 11th International conference, ICSR 2019, Madrid, Spain, November 26-29, 2019, Proceedings, Springer, LNCS 11876, Cham, pp 3-13
3. Jokinen K, Fukuda K, Iino N, Nishimura S, Nishimura T, Oota Y, Watanabe K, Yoshida Y (2019) Privacy and sensor information in the interactive service applications for elder people. In: The 7th national conference of serviceology, Tokyo
4. Moore R (2012) A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Sci Rep 2:864
5. Newen A, De Bruin L, Gallagher S (eds) (2018) The Oxford handbook of 4E cognition. Oxford handbooks online. Oxford University Press, Oxford
6. Rau P, Li Y, Li D (2009) Effects of communication style and culture on ability to accept recommendations from robots. Comput Hum Behav 25(2):587–595
7. Society 5.0. https://www.japanindustrynews.com/2019/04/connected-industries-japans-approach-to-industry-4-0-on-the-way-to-society-5-0

Data Augmentation, Collection and Manipulation

MixOut: A Simple Yet Effective Data Augmentation Scheme for Slot-Filling

Mihir Kale and Aditya Siddhant

Abstract We present a data augmentation strategy for slot-filling in task-oriented dialogue systems. It is simple yet effective and does not rely on external corpora. Lexicons for all slot types are generated from available annotated data. Synthetic, yet realistic utterances are then created by replacing slot values with other values of the same type. The method can also be easily extended to synthesize mixed language utterances for cross-lingual training. Monolingual experiments on 14 datasets across 10 different domains, 4 languages and cross-lingual experiments on 3 language pairs demonstrate the effectiveness of this method.

1 Introduction

Spoken language understanding (SLU) is a critical part of task-oriented dialogue systems, aiming to understand the semantics of user utterances. It is generally decomposed into two sub-tasks - intent detection and slot-filling. Intent detection is generally treated as text classification, while slot-filling is cast as a sequence labelling problem where each word is given a semantic label.

As personal assistants like Siri, Google Assistant, Alexa etc. continue to expand capabilities to new domains and languages, the requirement for labelled data increases tremendously. Scaling labelling efforts in order to have large labelled datasets for each and every domain and language can get prohibitively expensive in terms of time and money. This motivates the need for sample-efficient architectures, better methods to utilize unlabelled data, data augmentation etc. Unsupervised pre-training on task-specific data has emerged as a way to learn powerful representations of text leading to large increases in model accuracy [11]. However, these techniques, along with popular data augmentation methods such as self-training, rely on access to large amounts unlabelled data. Such in-domain data typically becomes

M. Kale and A. Siddhant—Work done as independent researchers.

M. Kale (✉) · A. Siddhant
Mountain View, USA
e-mail: siddhantaditya01@gmail.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_21

available at a later stage, when enough user logs are available and not when a model for a new domain/language needs to be built from scratch. Augmentation without relying on external resources is an attractive option, especially when expanding an assistant's capabilities to brand new domains/languages. Motivated by this setting, we investigate a data augmentation strategy - for the slot-filling task of SLU - that is (1) simple to implement, (2) does not rely on any external sources and (3) works in a variety of domains, languages and can also be employed across languages for cross-lingual training.

2 Related Work

Slot-Filling. [9] demonstrated the effectiveness of deep learning for slot-filling by learning recurrent neural network based models. [16] perform joint training of intent detection and slot-filling, leading to better performance on both tasks. [11] show that unsupervised pre-training leads to more accurate and sample efficient SLU models.

Data Augmentation in NLP. [14] use a technique similar to ours but with applications to cross-domain natural language generation. [6] propose augmentation via language model based word replacement, with applications to text classification. [17] learn a generative latent variable model to synthesize new utterances. [5] use a combination of delexicalization and seq2seq model for synthesizing new utterances. In contrast to both these methods, our approach is extremely simple and does not require training a separate model for augmentation.

Cross-lingual Transfer Learning. [13] use aligned word embeddings for cross-lingual slot-filling and show that such an approach is superior to machine translation based baselines. [10] find that using input representations from neural machine translation encoders perform even better than aligned word embeddings. In this work, we find that using representations from multilingual BERT perform comparably or even better than NMT encoders. This is quite surprising (given that BERT uses no parallel data), but is in line with findings from [15].

3 Approach

3.1 Slot-Filling for Task Oriented Dialogue

In this work, we focus on the slot-filling aspect of spoken language understanding, which is cast as a BIO sequence labeling task [10], where each word is assigned a semantic label. Following prior work [3, 10], our architecture consists of:

1. **Embedding Layer.** This layer embeds the utterance and returns embeddings for each word: these can be either non-contextual like fasttext [1] or contextual like BERT [3].
2. **biLSTM Layer.** The word embeddings are fed into a biLSTM layer. The final hidden state for each word is a concatenation of the forward and backward states.
3. **Dense Layer.** This layer projects each hidden state into the label space, generating logits for each class.
4. **CRF Layer.** The logits are then fed into a Conditional Random Field, which returns the final probability distribution over the classes.

3.2 *Data Augmentation*

The augmentation scheme is decomposed into 3 steps as illustrated in Fig. 1.

1. **Lexicon Generation.** First, for each slot type (e.g.. artist and year), we create lexicons from the training data, which are simply lists of all slot values for a given slot type.
2. **Delexicalization.** Next, each utterance is delexicalized by replacing slot-values with slot-type placeholders.
3. **Relexicalization.** Finally, to generate synthetic samples, we sample a delexicalized utterance and then relexicalize it by sampling slot values from the lexicon of the corresponding slot-type.

The newly generated utterances, while synthetic, are generally realistic. For any given utterance, we can use this method to generate multiple synthetic utterances, upper bounded by the size of the lexicons. We coin the procedure MixOut. This method can easily incorporate external lexicons for slots like movie titles, songs, locations etc., when available.

Fig. 1 Data augmentation procedure

Given that most slot-filling datasets are small (particularly for non-English languages), augmentation via MixOut expands the number of context and slot-value pairs seen during training which helps in improving the generalization capability of the model.

Cross-Lingual MixOut. When training data is available in two languages (e.g.. English and Thai), we can easily extend the augmentation scheme to generate mixed-language synthetic samples. The only modification is that if the delexicalized utterance is in one language say, English, the re-lexicalization step samples slot-values from the other language, Thai. Essentially, for each original English sentence, the synthetic sample will contain Thai slot-values. For each original Thai utterance, the synthetic sample will contain English slot-values. We hypothesize that such an augmentation would further the benefits of cross-lingual transfer learning. It could also be potentially useful for zero-shot generalization to code-switched utterances.

4 Experiments and Results

Following prior slot-filling work [10, 11], we use F1 score as our evaluation criteria. The numbers reported in Tables 2, 3 and 4 are F1 scores averaged across 3 runs for each experiment.

4.1 Model Architecture and Training

Our baseline model concatenates 300 (768) dimensional FastText (BERT) embeddings with a CNN character based representation. The character representation uses 16 dimensional character embeddings and 64 convolutional filters of width three characters, a ReLU activation and by max pooling. The token representation is passed through a biLSTM layer of 100 units, followed by a dense layer to generate logits which are fed to a CRF. The FastText/BERT embeddings are kept frozen during training. We optimize using Adam, with a learning rate of 0.001, and select the best model via early stopping on the basis of development set performance.

Table 1 Dataset description

Domain/Intent	#Train Samples	#Dev Samples	#Test Samples	#Slot- types
English				
Search Cr. Work	954	400	700	7
Weather	1000	400	700	9
Book Restaurant	973	400	700	14
Play Music	1000	400	700	9
Add to Playlist	942	400	700	5
Rate Book	956	400	700	7
Search Sc. Event	959	400	700	2
Thai				
Alarm	777	439	597	2
Reminder	578	336	442	6
Weather	801	460	653	5
Spanish				
Alarm	777	439	1,011	2
Reminder	578	336	1,025	6
Weather	801	460	1,057	5
Turkish				
Flight Booking	400	200	715	120

4.2 Performance Across Domains and Input Representations

For English, we use the Snips dataset [2]. The dataset represents a variety of domains that task oriented dialogue systems need to be able to handle - weather, restaurant booking, media player etc. However, since each intent has only 100 utterances in the test set, we create a new split with 700 samples for testing, 400 for development and the rest $\sim$ 1,000 for training. See Table 1 for more details.

When applying MixOut, for each utterance we sample 4 synthetic utterances and add them to the training data. We conduct two sets of experiments - with fastText and BERT as input embeddings [1, 3]. Contextual word representations have consistently outperformed their context independent counterparts for a large variety of tasks [3]. It has also been noted that improvements from architectural changes or multi-task learning are diminished when context independent representations are replaced with contextual representations such as ELMo [12]. Experimenting with both fastText and BERT lets us study if MixOut is susceptible to the same effect.

Table 2 Results on monolingual experiments. SOTA for Turkish is from [13] while for Spanish and Thai its from [10]

Domain/Intent	Baseline	Baseline +MixOut	Increase (Abs/Rel)	SOTA
English (FastText)				
Search Cr. Work	89.13	90.13	1.00/2.39 %	—
Weather	93.44	94.92	1.48/22.56 %	—
Book Restaurant	92.76	93.29	0.53/7.32 %	—
Play Music	89.94	90.18	0.24/2.39 %	—
Add to Playlist	92.66	93.51	0.85/11.58 %	—
Rate Book	97.40	97.65	0.25/9.62 %	—
Search Sc. Event	96.95	97.04	0.09/2.95 %	—
Average	**93.18**	**93.81**	**0.63/9.3** %	—
English (BERT)				
Search Cr. Work	93.52	94.16	0.64/9.88 %	—
Weather	96.15	96.52	0.37/9.61 %	—
Book Restaurant	94.06	94.82	0.76/12.79 %	—
Play Music	91.41	92.06	0.65/7.57 %	—
Add to Playlist	95.09	96.07	0.98/19.96 %	—
Rate Book	98.22	98.46	0.24/13.48 %	—
Search Sc. Event	98.17	98.55	0.38/20.77 %	—
Average	**95.23**	**95.80**	**0.57/12.04** %	—
Thai (BERT)				
Alarm	88.76	90.36	1.60/14.23 %	—
Reminder	84.61	86.02	1.41/9.16 %	—
Weather	93.65	95.96	1.91/30.07 %	—
Average	**89.56**	**91.23**	**1.73/22.30** %	**90.63**
Spanish (BERT)				
Alarm	83.50	84.71	1.21/7.91 %	—
Reminder	87.44	88.07	0.62/5.25 %	—
Weather	91.42	92.10	0.68/8.61 %	—
Average	**87.47**	**88.31**	**0.83/6.69** %	**81.64**
Turkish (BERT)				
Flight Booking	**77.04**	**79.36**	**2.31/10.10** %	**75.50**

In the case of fastText, as seen in Table 2, augmentation results in relative error reduction of 9.3% averaged across the 7 domains. Replacing fastText with contextual BERT embeddings itself improves the baseline model by a large margin. However, MixOut further improves performance across each of the 7 domains, with an average relative error reduction of 12%. The impressive gains even with BERT indicate that the synthetic data is providing information that complements the contextual signals provided by BERT.

4.3 Applicability to Other Languages

Our motivation for MixOut came after observing patterns in English language datasets. In this section we study the applicability of the method to other languages. Specifically, we choose Thai, Spanish and Turkish. This set lets us study languages with different character sets from English (Thai), languages from the same family as English (Spanish), languages that are linguistically similar (English, Thai and Spanish are SVO) and dissimilar (Turkish is SOV).

For Thai and Spanish, we rely on the multilingual spoken language understanding corpus from [10], which provides data for 3 domains. Since the Spanish datasets are bigger than Thai and to foster a fairer comparison of results across the two languages, we randomly sub-sample the Spanish training and development splits so that they are of the same size for both languages. Note that the test sets are left untouched. For Turkish, we utilize the recently released *multilingual-atis* datasets from [13] which consist of translations of utterances from the original ATIS dataset [7]. Both languages have 600 samples for training. Since the authors do not provide a development split, we randomly split the datasets into 400 / 200 for training/development. Again, the test sets are left untouched. Table 1 presents statistics for the datasets, along with example utterances. Similar to Sect. 4.2, for each utterance, we sample 4 synthetic utterances.

For our embedding layer, we use the multilingual BERT model trained on Wikipedia corpora of 102 languages [3]. From results in Table 2, we observe improvements in all 7 datasets across the 3 languages considered. On average, we note a relative error reduction of over 22% for Thai, 10% for Turkish and 6.7% for Spanish.[1] The gains are higher for typically low resource languages like Thai and Turkish as compared to Spanish. This set of experiments also confirms that MixOut works well for a variety of languages with different linguistic characteristics. Finally, for all languages, we also obtain a new state-of-the-art in the monolingual setting.

4.4 Cross Lingual Transfer Learning

In this section, we study the applicability of MixOut for cross-lingual transfer learning. In many real-world scenarios, there is abundant training data for English, but other languages are severely under resourced. This motivates the need for transfer learning from high resource languages (HRL) to low resource languages (LRL). [13] show that joint training of LRL and HRL can prove to be beneficial for slot-filling in the presence of aligned word embeddings and performs better than translation based alternatives [4]. [10] show that joint training can be further improved by using contextual embeddings from bidirectional machine translation models trained on large amounts of parallel data.

[1] Following [10], average across domains is a weighted average, weighted by the number of samples in the test set.

Table 3 Results on cross-lingual experiments. SOTA for Turkish is from [13] while for Spanish and Thai its from [10]

	Concat	Concat +MixOut	Increase (Abs/Rel)	SOTA
Thai-English (BERT)				
Alarm	89.74	89.62	−0.12/−1.16 %	—
Reminder	87.27	89.64	2.37/18.61 %	—
Weather	93.24	95.26	2.02/29.88 %	—
Average	**90.44**	**91.80**	**1.35/14.19 %**	**91.51**
Spanish-English (BERT)				
Alarm	85.12	85.14	0.01/0.11 %	—
Reminder	88.67	88.48	−0.19/−1.67 %	—
Weather	92.66	92.08	−0.57/−7.28 %	—
Average	**88.84**	**88.58**	**−0.25/−2.23 %**	**83.00**
Turkish-English (BERT)				
Flight Booking	**80.23**	**81.72**	**1.44/7.90 %**	**78.90**

In all our experiments, English is treated as the HRL. The English datasets from [10] have 9,282/14,339/6,900 (alarm/weather/reminder) samples, while the English ATIS dataset consists of 4,478 utterances. Overall, the HRL datasets contain 5–10 more data than LRL. When applying MixOut, we first concatenate the HRL and LRL datasets. Then we generate one mixed-language synthetic sample for each utterance as explained in Sect. 3. In other words, the augmented dataset consists of monolingual utterances from both languages along with the synthetic mixed-language utterances.

For our experiments, we use the multilingual BERT model for the embedding layer. Note that multilingual-BERT does not use any parallel data. Neural machine translation and MUSE based embeddings used in prior work are explicitly trained to align representations of similar words/phrases across languages. On the other hand, the BERT training procedure does not directly encourage such alignment. This can pose significant challenges especially when the HRL and LRL don't share any subwords (English-Thai, English-Hindi etc.). However, since our synthetic samples are mixed language, training on this data would force the model to have similar representations for similar contexts and slot-values across languages, which in turn would help the model better utilize the training data from high resource language.

Though our aim in this study is not to beat state of the art, but to study the effectiveness of MixOut, we also list state-of-the-art results from [13] and [10] in SOTA column of Tables 2 and 3 to put our results in perspective.

Using MixOut leads to impressive improvements over the baseline for Thai, as seen in Table 3. Across the three domains considered, we observe an average relative error reduction of over 14%. For Thai, our baseline model lags behind [10] who used a large corpus of parallel data to train a bidirectional NMT based encoder [8].

Table 4 Effect of increase in size of synthetic data. Numbers reported are slot F1 scores averaged over the 7 SNIPS domains

Augmentation ratio	Baseline	1:1	1:2	1:4
English (BERT) Average	95.23	95.50	95.59	95.80

However, improvements from MixOut help push our BERT based model to a new state-of-the-art.

The Turkish results are particularly interesting, since the augmented samples for Turkish are least realistic (English and Turkish have different word orders). In spite of that, MixOut leads to an improvement of 1.5 F1 over the baseline. On the other hand, MixOut does not lead to improvements with Spanish. One possible explanation is that English and Spanish already have a large amount of subword overlap which is already encoded in the BERT embeddings.

Overall, the results indicate that MixOut would be particularly beneficial for linguistically distant language pairs. Finally, we also obtain a new state-of-the-art for all languages in the cross-lingual setting.

4.5 Effect of Amount of Synthetic Data

To study the relationship between performance and size of synthetic data, we use the 7 English (SNIPS) datasets and add synthetic data in 1:1, 1:2 and 1:4 ratios. Here 1:x indicates that for each natural utterance in the dataset, we add x synthetic utterances. Baseline doesn't use augmentation. For each of the 7 domains, 1:4 configuration gives the best performance. The results of this study are reported in Table 4. We notice that, on average, performance increases with the amount of synthetic data.

5 Conclusion and Future Work

We study MixOut, a simple data augmentation technique for slot-filling systems which does not rely on any external resources. It leads to good improvements across a variety of domains, languages and even benefits cross lingual transfer learning. In the future, we would like to extend MixOut to other sequence labelling tasks and examine its usefulness in code-switched utterances. Another interesting line of future work is to leverage knowledge bases like WikiData in the lexicon generation process of MixOut.

References

1. Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information. Trans Assoc Comput Ling 5:135–146
2. Coucke A, Saade A, Ball A, Bluche T, Caulier A, Leroy D, Doumouro C, Gisselbrecht T, Caltagirone F, Lavril T, Primet M, Dureau J (2018) Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. ArXiv **abs/1805.10190** (2018)
3. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT
4. He X, Deng L, Hakkani-Tur D, Tur G (2013) Multi-style adaptive training for robust cross-lingual spoken language understanding. In: International conference on acoustics, speech and signal processing (ICASSP)
5. Hou Y, Liu Y, Che W, Liu T (2018) Sequence-to-sequence data augmentation for dialogue language understanding. In: COLING (2018)
6. Kobayashi S (2018) Contextual augmentation: Data augmentation by words with paradigmatic relations. ArXiv **abs/1805.06201**
7. Liu B, Lane I (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. In: INTERSPEECH
8. McCann B, Bradbury J, Xiong C, Socher R (2017) Learned in translation: contextualized word vectors. In: Advances in neural information processing systems, pp. 6294–6305
9. Mesnil G, Dauphin Y, Yao K, Bengio Y, Deng L, Hakkani-Tur D, He X, Heck L, Tur G, Yu D et al (2014) Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans Audio Speech Lang Process 23(3):530–539
10. Schuster S, Gupta S, Shah R, Lewis M (2018) Cross-lingual transfer learning for multilingual task oriented dialog. In: NAACL-HLT
11. Siddhant A, Goyal A, Metallinou A (2018) Unsupervised transfer learning for spoken language understanding in intelligent agents. In: AAAI
12. Strubell E, McCallum A (2018) Syntax helps elmo understand semantics: Is syntax still relevant in a deep neural architecture for srl? ArXiv **abs/1811.04773**
13. Upadhyay S, Faruqui M, Tür G, Dilek HT, Heck L (2018) (almost) zero-shot cross-lingual spoken language understanding. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6034–6038
14. Wen TH, Gaić M, Mrksic N, Rojas-Barahona LM, hao Su P, Vandyke D, Young SJ (2016) Multi-domain neural network language generation for spoken dialogue systems. In: HLT-NAACL
15. Wu S, Dredze M (2019) Beto, bentz, becas: the surprising cross-lingual effectiveness of bert. In: EMNLP/IJCNLP (2019)
16. Xu P, Sarikaya R (2013) Convolutional neural network based triangular CRF for joint intent detection and slot filling. In: 2013 IEEE workshop on automatic speech recognition and understanding. IEEE, pp 78–83
17. Yoo KM, Shin Y, goo Lee S (2018) Data augmentation for spoken language understanding via joint variational generation. In: AAAI

Towards Similar User Utterance Augmentation for Out-of-Domain Detection

Andoni Azpeitia, Manex Serras, Laura García-Sardiña, Mikel D. Fernández-Bhogal, and Arantza del Pozo

Abstract Data scarcity is a common issue in the development of Dialogue Systems from scratch, where it is difficult to find dialogue data. This scenario is more likely to happen when the system's language differs from English. This paper proposes a first text augmentation approach that selects samples similar to annotated user utterances from existing corpora, even if they differ in style, domain or content, in order to improve the detection of Out-of-Domain (OOD) user inputs. Three different sampling methods based on word-vectors extracted from BERT language representation model are compared. The evaluation is carried out using a Spanish chatbot corpus for OOD utterances detection, which has been artificially reduced to simulate various scenarios with different amounts of data. The presented approach is shown to be capable of enhancing the detection of OOD user utterances, achieving greater improvements when less annotated data is available.

1 Background

Due to the increasing digitisation of our society, the creation of more communication channels and recent advances on Natural Language Processing (NLP), Decision Making and Automatic Speech Recognition, the use of Dialogue Systems (DS)—more

A. Azpeitia (✉) · M. Serras · L. García-Sardiña · M. D. Fernández-Bhogal · A. del Pozo
Vicomtech Parque Científico y Tecnológico de Gipuzkoa, Paseo Mikeletegi 57, Donostia/San, Sebastián, Spain
e-mail: aazpeitia@vicomtech.org

M. Serras
e-mail: mserras@vicomtech.org

L. García-Sardiña
e-mail: lgarcias@vicomtech.org

M. D. Fernández-Bhogal
e-mail: mfernandez@vicomtech.org

A. del Pozo
e-mail: adelpozo@vicomtech.org

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_22

commonly known as Voice Assistants or chatbots—is increasing, as they provide a natural, frictionless, and automatic way to solve tasks of multiple domains and complexities [4].

Any DS that intends to serve its purpose should perform three principal tasks: understand the user, plan dialogue interaction accordingly, and give a comprehensible response. These tasks are usually performed by Natural Language Understanding (NLU), Dialogue Manager (DM), and Natural Language Generation (NLG) components respectively.

Focusing on the first step of a DS, a wide variety of technological approaches have been proposed to try to understand the users' communicative goals in a dialogue. Early methods employed handcrafted rules and grammars to parse the users' utterances. Other data-driven techniques use machine learning algorithms to classify the users' utterances into a set of pre-defined labels, while sequential algorithms such as Conditional Random Fields have been employed to retrieve specific entities that appear in the surface representation of the phrase. Recent advances in deep learning have enabled the application of more sophisticated data-driven neural network approaches, becoming state of the art.

Despite these advances, the principal issue that arises when deploying DS at industrial scale in new scenarios is the lack of available dialogue corpora. Several efforts have attempted to reduce the impact of data scarcity through text augmentation in different NLP applications. [13] proposed cluster-based document modelling to augment existing text samples and improve information retrieval. Another explored text augmentation method is the use of a thesaurus such as WordNet, in order to augment words and phrases with their synonyms [17]. However, the compilation of such thesauri requires a high manual effort. The arise of neural networks and distributed representations of words and language models allowed experimenting with techniques that exploit existing corpora. In [9], word2vec embeddings were used to retrieve nearest neighbours of each token in a query to automatically augment it. [5] replaced the words of a query with words predicted by a Long Short Term Memory (LSTM) based Language Model (LM), that also takes context into account. Similar LM-based text augmentation approaches were applied in Machine Translation by [3]. More recently, the fine-tuning of pre-trained language representation models (e.g., BERT, ERNIE, RoBERTa, ALBERT) has pushed the state-of-the-art in several NLP tasks [2, 7, 12, 14] and proved to be an effective technique even in data-scarce scenarios[1, 8].

This work focuses on exploring the use of pre-trained language representation models to augment a minimal set of user utterances with existing similar data in other communication styles (e.g., emails, web-pages, Wikipedia). The aim is to improve a task oriented DS's ability to detect out-of-domain (OOD) user utterances and reply accordingly. This is important in practical applications, where the acceptance of unsupported OOD input may lead to failure and considerably worsen users' perception of the DS.

The paper is organised as follows: Sect. 2 presents the explored user utterance augmentation approach and methods; Sect. 3 describes the datasets employed and the setup of the experiments carried out; the results obtained are presented in Sect. 4 and

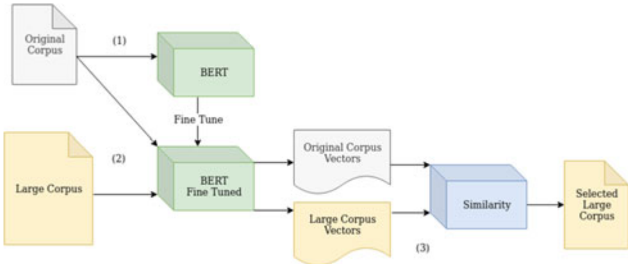


Fig. 1 Similar User Utterance Augmentation Approach. (1) A BERT model is fine-tuned, (2) sentence-vectors are computed using the fine-tuned BERT, and (3) similar utterances are selected

analysed in Sect. 5; finally, Sect. 6 draws some conclusions and presents suggestions for future research.

2 Similar User Utterance Augmentation

Our main goal is to augment a small corpus of annotated user utterances with samples from existing unannotated larger corpora. To this purpose, the proposed approach takes advantage of the pre-trained BERT [2] language representation model. As shown in Fig. 1, the explored similar user utterance augmentation method follows the steps below:

1. First, a sequence classifier is trained fine-tuning a pre-trained BERT model with the small corpus.
2. Second, the BERT model from the previous step is used to compute sentence-vectors from the small and the large corpus.
3. Finally, sentence-vectors are compared to select the most similar utterances in the larger corpus.

In the second step, all sentence-vectors are calculated for both the small and the large corpus. Sentence-vector creation is done as follows. Let S be an utterance of length n, $w_1, ..., w_n$ the words in S, SV the vector representation for S, and wv_i the vector representation for w_i. SV is computed as the sum of word-vectors: $SV = \sum_{i=1}^{n} wv_i$. In order to extract word-vectors, the hidden states of the last four encoding layers are added. Although other combinations are possible, the last four layers achieved almost the best results [2] in the CoNLL- 2003 Named Entity Recognition task and in addition is more efficient than concatenation in terms of memory consumption and processing time.

The final step requires a method to evaluate similarity between sentence-vectors. So far, three methods have been tested: K-Best Similarity, Average Similarity, and Weighted Average Similarity. All these methods compare sentence-vectors in the small and large corpora based on their cosine similarity. Each of them is described in more detail in the following subsections.

2.1 K-Best Similarity (KBS)

For each utterance in the small corpus, this method selects the k most-similar utterances in the large corpus, based on their cosine similarity.

It is very likely that the k-best utterances for a utterance S_i have higher scores than the k-best utterances for another utterance S_j, or even for the same utterance to have some k-best utterances with low scores (especially if the k value is high). To address this issue, a similarity threshold is used—i.e. an utterance of the large corpus is only selected if its similarity score is greater than the threshold.

2.2 Average Similarity (AS)

KBS has two main counterparts: (i) there are two parameters that need to be adjusted (k-best and the similarity threshold); and (ii) k similar utterances are searched for all the utterances in the small corpus, therefore, if an utterance in the small corpus is not very representative of the domain, it is very likely that its most similar utterances in the large corpus will not be representative either.

The idea behind the Average Similarity (AS) method is to sort the utterances in the large corpus by their similarity to the entire small corpus, instead of by their similarity to each single utterance. This way, data augmentation is simply done by selecting the desired amount of sorted utterances in the large corpus, and the impact of dubious utterances in the small corpus is reduced.

Let S_s and S_l be the small corpus of size n and the large corpus respectively, and SV_{si} and SV_{lj} the sentence-vector representation for the ith utterance in S_s and the sentence-vector representation for the jth utterance in S_l respectively. Then, the similarity score of the AS method is computed as shown in Eq. 1.

$$sim\_as(S_{lj}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \cos(SV_{si}, SV_{lj}) \tag{1}$$

2.3 Weighted Average Similarity (WAS)

In an utterance classification scenario, the AS method scores an utterance S_{li} based on its similarity to a label L_{s1} regardless of its similarity to another label L_{s2}. Intuitively, if S_{li} and another utterance S_{lj} are equally similar to L_{s1} but S_{lj} is more similar to L_{s2}

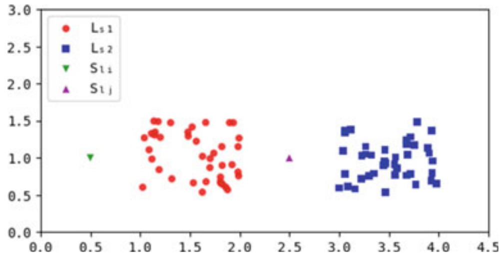


Fig. 2 Weighted Average Similarity (WAS). Utterance S_{li} is scored higher than S_{lj} for label L_{s1}

than S_{li}, it might be preferable to score S_{li} higher than S_{lj} for L_{s1}. Figure 2 shows an example of this type. To address such cases, WAS adds a weight to the AS metric.

Let S_{s1} and S_{s2} be two small corpora with sizes n and m respectively and L_{s1} and L_{s2} the labels which identify an utterance within corpus S_{s1} and S_{s2}, respectively. Now, let S_l be the large corpus to be used to augment the S_{s1} corpus, and S_{lj} an utterance in S_l. Then, the similarity between the utterance S_{lj} and the label L_{s2} is calculated as shown in Eq. 2 and the WAS is computed as shown in Eq. 3.

$$sim\_label(S_{lj}, L_{s2}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \cos(SV_{s2i}, SV_{lj}) \tag{2}$$

$$sim\_was(S_{lj}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\cos(SV_{1i}, SV_{2j}) \cdot (1 - sim\_label(S_{lj}, L_{s2})) \right) \tag{3}$$

3 Experimental Setup

In this section we present the corpora used for experimentation, as well as the setup of the testing environment that was carried out.

3.1 Corpus Description

The small corpus where the proposed similar user utterance augmentation approach is applied is MOVE CHATBOT. This corpus contains user utterances that happen in the context of a public transportation related chatbot, both In-Domain (IND, i.e. utterances in which users ask about something related to the public transportation domain) and Out-of-Domain (OOD, i.e. utterances in which users ask about information related to other domains such as the weather, sports, jokes or insults). In addition, the corpus also contains user utterances that can be considered to be neutral (N, i.e. applicable across domains, such as greetings and farewells). It was designed to train

Table 1 MOVE CHATBOT corpus statistics

Corpus	IND	N	OOD	Total
MOVE CHATBOT 500	205	129	166	500
MOVE CHATBOT 400	163	111	126	400
MOVE CHATBOT 300	128	80	92	300
MOVE CHATBOT 200	81	53	66	200
MOVE CHATBOT 100	41	31	28	100
MOVE CHATBOT 50	18	18	14	50
MOVE CHATBOT DEV	56	37	44	137
MOVE CHATBOT TEST	303	204	227	734

an OOD detector in order to improve the naturalness of the public transportation related chatbot's responses, by discriminating between IND/N/OOD user utterances and responding accordingly.

The corpus was crowd-sourced among people who had knowledge of the application domain and whose mother tongue was Spanish. A total of 1,371 different user utterances were collected and split into train/dev/test sets. To mimic different situations of data scarcity, several training sets were prepared reducing the original training set. Training sets with 500, 400, 300, 200, 100 and 50 samples were created, in such a way that smaller sets were always included in the larger sets. The testing set is significantly larger than the training and development sets. This proportion was set to simulate the real challenges of industrial chatbot deployment, where the variability of the possible user utterances dialogue systems needs to handle is huge. The partition numbers are described in Table 1.

To perform data augmentation, two additional corpora were used: MOVE EMAIL and OPENSUBS. MOVE EMAIL consists of email data used by the public transportation back-office to solve common issues and answer users' questions. It contains a set of 8,314 utterances within the transport card's customer support domain. Despite MOVE CHATBOT and MOVE EMAIL can both be considered to be IND, they differ in style and content, so exploiting the latter corpus for user utterance augmentation is not straightforward.

In order to augment OOD and neutral data, we used the OPENSUBS corpus [6]: a collection of 2.6 billion available utterances gathered from OpenSubtitles.[1] This source was chosen because dialogues that can be considered similar in style to chatbot user utterances constitute a large portion of the subtitle corpus. It was gathered from the OPUS Open Parallel Corpus [15]. To reduce the search space for similar utterances, as well as to improve memory management, 5M utterances were randomly selected. This quantity provided a good balance between size and usefulness. Depending on the size of the user utterance set to be augmented, the volume of the OPENSUBS corpus used in each experiment varied accordingly.

[1] www.opensubtitles.org: a webpage containing subtitles for a vast amount of movies in many different languages.

3.2 Testing Setup

The presented user utterance augmentation approach was tested on a multi-label classification scenario. As explained before, the MOVE CHATBOT corpus was used to train a utterance classifier that tags each user utterance as IND, OOD or neutral. The proposed approach is used to augment a small annotated set of user utterances from MOVE CHATBOT by sampling similar utterances from the MOVE EMAIL and the OPENSUBS corpora. A short summary of the labels and corpora used throughout the experiments is described next:

- IND: questions or requests related from the MOVE CHATBOT corpus related to public transport card management that are augmented with similar utterances from the MOVE EMAIL corpus.
- N: utterances from the MOVE CHATBOT corpus such as greetings or farewells which are applicable across domains and are augmented with similar utterances from OPENSUBS.
- OOD: utterances from the MOVE CHATBOT corpus which are not related to public transport card management that are augmented with similar OPENSUBS utterances.

To test the usefulness of the proposed augmentation approach under varying degrees of data scarcity, different sizes of the MOVE CHATBOT corpus were sampled. Note that all the utterances of the smaller sets are included in the larger ones, e.g. all utterances in MOVE CHATBOT 50 are in MOVE CHATBOT 100, which are also contained in MOVE CHATBOT 200 and so on. Table 1 describes the sizes of the corpus with which we have experimented.

The pre-trained BERT BASE MULTILINGUAL CASED model was fine-tuned for each explored IND/N/OOD classification experiment using the HUGGINGFACE library [16]. All models were trained with a fixed seed to reduce the variability of results and make fair comparisons across the models. Model training was done on 30 epochs with a batch-size of 64. Each epoch was evaluated on the MOVE CHATBOT DEV corpus and the best performing model was then chosen to evaluate on the MOVE CHATBOT TEST corpus.

All experiments were conducted on a server with two CPUs of 32 threads each, 128GB of RAM and four GPUs: two TITAN X (PASCAL) and two GEFORCE GTX 1080. Because training corpora were quite small (the biggest one containing 15,170 utterances), training-time took only between 5 min and 2 h (depending on the parameters of the computed similarity method) in this setting. Thus, it would be feasible to train on fewer GPUs and/or CPUs.

Table 2 Baseline average F1 scores achieved training the multi-label IND/N/OOD classifier on different sample sizes of the MOVE CHATBOT corpus

Samples	IND	N	OOD	Total
500	90.06	80.00	78.43	82.83
400	88.37	78.61	75.72	80.90
300	88.10	78.55	74.19	80.28
200	82.14	76.84	68.36	75.78
100	78.79	74.59	65.32	72.90
50	77.47	73.24	52.71	67.81

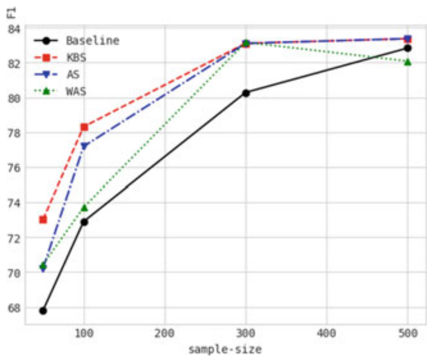


Fig. 3 Average F1 scores comparison with KBS, AS and WAS using different sample sizes on MOVE CHATBOT corpus, and then augmenting each sample size with MOVE EMAIL and OPENSUBS corpora

4 Results

This section presents the preliminary results obtained by our system. Table 2 shows baseline evaluation results, in which the multi-label IND/N/OOD classifier is trained with the different sample sizes of the MOVE CHATBOT corpus. As expected, the performance of the classifier decreases as the training sample size is reduced. This trend is specially marked below 300 samples.

The performance of the proposed user utterance augmentation approach was explored for the three similarity metrics defined (KBS, AS and WAS) and tested on the MOVE CHATBOT corpus of 50, 100, 300, and 500 sample sizes. Figure 3 shows the obtained results. Overall, KBS is the top scoring metric, achieving a highest improvement of +5,41 on a sample size of 100 user utterances. In addition, performance improvements decrease for all metrics as the size of user utterance samples is increased.

Figures 4, 5 and 6 show the impact of parameter setting for KBS, AS and WAS, respectively. In KBS, k-bests bigger than 15 obtain better results and, in general, increasing the similarity threshold up to 0.85 helps. In AS, the baseline is improved in all sample sizes, achieving a greatest improvement of +4,28 on a sample size of 100 user utterances. However, with this similarity metric it is difficult to estimate the amount of candidates that achieve best results across sample sizes. Finally, WAS

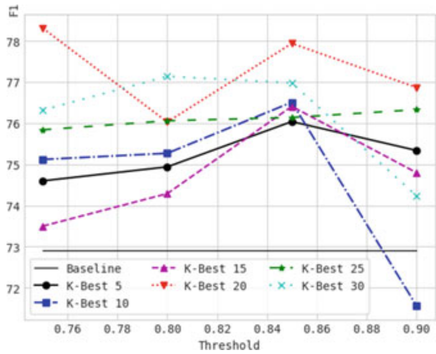


Fig. 4 Average KBS F1 score evolution for a sample size of 100 on MOVE CHATBOT corpus, and varying k-best and similarity thresholds on MOVE EMAIL and OPENSUBS corpora

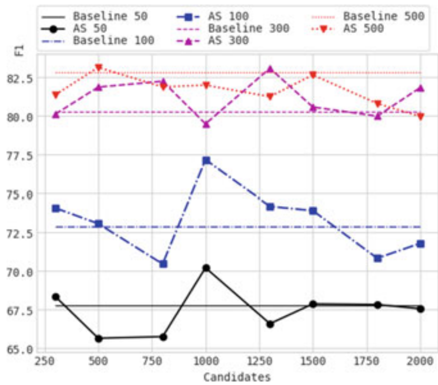


Fig. 5 Average AS F1 score evolution across sample sizes on MOVE CHATBOT corpus for varying amounts of candidates on MOVE EMAIL and OPENSUBS corpora

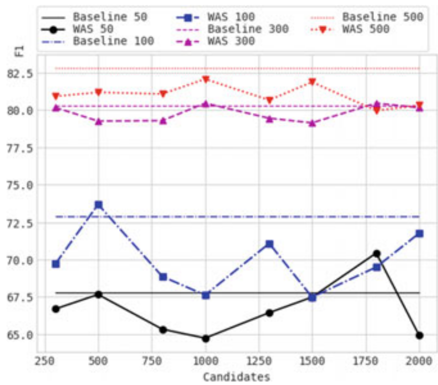


Fig. 6 Average WAS F1 score evolution across sample sizes on MOVE CHATBOT corpus for varying amounts of candidates on MOVE EMAIL and OPENSUBS corpora

results show that this metric is not really useful. In general, it performs worse than AS and does not even beat the baseline with a sample size of 500 user utterances. This can be explained by the nature of the method itself: it was designed to favour utterances that discriminate well across domains, which does not seem to represent the distribution of the IND/N/OOD classes in the targeted application scenario.

5 Discussion

In this section, a brief analysis of the achieved results is presented and a manual evaluation is performed over the utterances selected for augmentation by each metric, in order to achieve a better understanding of the performance of the three different scoring methods.

Overall, the proposed user utterance augmentation approach is capable of improving the performance of the IND/N/ODD classification task over the baseline. When more MOVE CHATBOT data samples are fed to the classifier, the proportional performance gain of the proposed approach is reduced. This behaviour is somehow expected, but the initial goal of the system, to overcome situations of data scarcity, is satisfied. The variation in the scoring curves regarding threshold and sample size is related to the high variability of the utterances collected in the MOVE CHATBOT corpora, as there was almost no constraint associated to the type of utterances users could say.

As described in Sect. 4, the best performing scoring metrics to automatically augment user utterances with similar sentences from existing corpora were KBS and AS. With regard to the WAS metric, which will be further analysed next, the utterances sampled for corpus augmentation did not manage to improve baseline results.

In order to better understand the kind of sample utterances each method selects for corpus augmentation, a short analysis per similarity method and class is provided. For this purpose, the models that achieved best scores on the MOVE CHATBOT 50 user utterance corpus for each metric were analysed: KBS with k-10 and threshold 0.80, AS with 1,000 candidates, and WAS with 1,800 candidates. The choice to focus on the MOVE CHATBOT 50 corpus was made so as to contemplate situations where the available data is most limited and the effect of data augmentation is more pronounced.

5.1 Augmentation with KBS

For the **IND** label, the best performing version of this method augmented the baseline corpus with 161 samples from the MOVE MAIL corpus. The selected samples are correctly tagged with this class and relevant, showing variability in the type of samples chosen.

For the **neutral** label, 180 samples from the OPENSUBS corpus were selected for augmentation using the best performing k and threshold. One characteristic of the selected utterances is that there is a high amount of repetition, since series of very similar samples were selected. For example, we found the following samples: *Has sido de muchas ayuda*, *Has sido de gran ayuda*, *Ha sido de mucha ayuda*, *De hecho has sido de gran ayuda*, all saying that "you have been of great help"; there also were samples like *no lo creo*, *No lo creo*, *no me lo creo*, *No, no lo creo*, *No creo*, *No creo* *eso*, *Yo no lo creo*, all saying "I don't think/believe so", in very similar ways. Apart from this, some languages other than Spanish were captured with this label (e.g., *Buongiorno*, although the neutral label in this case is correct). Among the samples selected as neutral, there were some that would actually correspond to the OOD label (e.g., *Quien eres tú para hablar así de un obispo de la Iglesia?*, "Who are you to talk about a bishop of the Church like that?"). Some of them contain parts that can be considered neutral, like salutations and thanking, but the sample in its full form should be considered OOD: *Hola, puedo hablar con alguien para encargar unas flores, por favor?*, "Hello, can I talk to someone to order some flowers, please?".

As for **OOD** samples, the best performing KBS method adds 127 new cases with this label. As in the case of the neutral label, some selected utterances are repetitive and too similar, but less than in the previous case: e.g., *De vacaciones?*, *Mis vacaciones?*, *Unas vacaciones?*, all asking about holidays. This set al.so includes some onomatopoeic samples that should have been labelled as neutral rather than OOD, for example - *Mm-hmmm*, *Ahhhhhhhh*, or *espera, shhh, shhh*.

5.2 *Augmentation with* AS

For the **IND** label, many of the chosen utterances include the word *tarjeta* ("card") or are about card top-ups and transport fares, so they can be considered relevant to expand our IND corpus.

For **neutral** samples, the lack of variability found with the KBS method does not show with this metric. There is variability in the samples tagged as neutral selected for augmentation, and they tend to be short, as in the original MOVE CHATBOT corpus. Some of them, though, should have been considered as OOD rather than neutral: *Nada de balas.* ("no bullets"), *Que no se moje* ("Do not let it get wet"), *Dame el dinero* ("Give me the money"), *Buscado por robo* ("Wanted for robbery"), *NO compres azúcar* ("Do not buy sugar").

Samples in the **OOD** selected set are varied, touching many topics, and having the correct assigned label for a class that is noisy and sparse by nature.

5.3 *Augmentation with* WAS

For the **IND** label, one striking difference with the other methods is formatting, since many samples were upper-cased. Yet, many of these samples included the word "card" and related topics, as happened in the selection with AS. Although most of the samples were correctly selected, some of them should have been considered as neutral: *GRACIAS POR COLABORAR!!* ("Thanks for your collaboration!!"), *GRACIAS ANTICIPADAS* ("Thank you in advance"), *Un saludo!!!* ("Greetings!!!"), *HOLA.* ("Hello"), *No*, ("No"), *Se lo agradecería* ("I would appreciate that"), *BUENOS DIAS* ("Good mornin;"), *Y Feliz Año* ("And happy new year"). This set of chosen samples also includes addresses, telephone and fax numbers, names, dates, and so on, which could have been considered neutral, even if relevant to the IND scenario in some cases, and channel-related information (these utterances come from emails) that could be rather considered OOD: e.g., *Enviado desde el movil*, “Sent from mobile”.

As for samples augmenting the **neutral** label set, many of the chosen samples should be considered OOD or even directly discarded, like spare names, numbers, letters, punctuation marks and symbols, and words in foreign languages: e.g., *71*, *Pato* (“Duck”), *Bear*, *L*, }, _. Also, as happened when using the KBS method, there are repetitions of very similar samples, but with different casing and punctuation. Other utterances selected as neutral should also have been considered OOD instead: e.g., *No más poker* (“No more poker”), *No está casado* (“He is not married”).

The set of samples selected by this method as **OOD** includes all sorts of data: letter (*E, F, G, H, l, J*) and number (*1, 2, 3, 4, 5*) sequences, texts in foreign languages (*Murasaki no shingou ga hikatte shikou teishi.*), singing (*Na -na- na-na -na- na -na- na , na -na- na -na- na -na- na-na*), a few repetitions of samples different only by punctuation... It also includes a few cases that should have been considered neutral rather than OOD: e.g., *Si, de hecho s.* (“Yes, actually yes”), *Si, ya se* (“Yes, I know”). It is also interesting to note that this method selected some name-only samples as neutral and some others as OOD. Although the samples in this set are varied, many of them might not be very relevant for the target chatbot scenario, since cases similar to them may never happen in such setting.

5.4 Summary

As a wrap-up of the augmented corpora analysis, despite KBS was the metric selecting less samples and a small part of its selections had little variability among themselves, it was the best performing method for IND and neutral labels, and for the overall classification. Although the AS metric provided the most diverse and consistent corpus augmentation, its performance is slightly worse than that achieved by the KBS method. Finally, the WAS similarity metric generally falls behind the other two in all evaluations.

According to the results, the similarity metric settings do not follow a clear pattern. However, some conclusions and recommendations can be drawn. For KBS, selecting 20–25 similar sentences for each sentence with a threshold of 0.80–0.85 usually helps in our datasets. A threshold bigger than 0.85 is too aggressive for IND label because due to the size of the MOVE EMAIL corpus it is difficult to retrieve sentences with near perfect similarity scores. AS performs worse than KBS but it is easier to tune: a number of 1,000–1,500 candidates for each label is the top scoring setting in almost all the cases. WAS performs the worst and the fact that for each sample size the top scoring setting is totally different shows that. Furthermore, for a sample size of 300 WAS hardly improves the baseline, and for 500 samples it does not even beat the baseline.

Overall, the most consistent class for user utterance augmentation across methods was the IND label. This phenomenon is partly explained by the fact that the OOD label is more heterogeneous and varied than the other labels. The results also support the idea that the MOVE EMAIL is better suited for the IND label than the OPENSUBS corpus for the other labels. The statement that the MOVE EMAIL corpus is well suited for the IND label is not surprising, given that it is a collection of emails about questions from users of the public transport service and is directly related to that label.

6 Conclusions and Future Work

We have presented a text augmentation approach capable of improving the detection of Out-of-Domain user utterances in dialogue interaction, by automatically selecting sentences similar to those in a small annotated corpus from other existing corpora. Three different similarity metrics have been defined and their performance has been evaluated.

Overall, the proposed approach achieves greater improvements when less annotated data is available. This result is not surprising due to the fact that similar data is not as accurate as domain data. However, the achieved results demonstrate that augmenting user utterances with similar data helps model training in data-scarce application scenarios and that similar domain data is useful for data augmentation.

Nevertheless, as a first data augmentation approach, there is still room for research and improvement. Despite the presented preliminary results are quite satisfactory, the explored similarity metrics need to be tuned in order to achieve robust results. This fact indicates that the studied metrics do not correlate exactly with utterance and/or class similarity. Trying to learn weights for the soft cosine similarity [11] and/or building other spatial representations could be helpful to better identify similar sentences.

To make the presented methodology closer to its application in industry, human-in-the-loop methodologies such us Active Learning [10] could be tested, with the aim of reducing the cognitive load of building domain and task-specific Natural Language Understanding modules in data-scarce scenarios.

Acknowledgments This work has been partially supported by the HAZITEK program (CONTACT ZL-2020/00237) of the Economic Development and Infrastructure department of the Basque Government.

References

1. Cruz JCB, Cheng C (2019) Evaluating language model finetuning techniques for low-resource languages. arXiv preprint arXiv:1907.00409

2. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
3. Fadaee M, Bisazza A, Monz C (2017) Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), pp 567–573
4. Gorin AL, Riccardi G, Wright JH (1997) How may i help you? Speech Commun 23(1–2):113–127
5. Kobayashi S (2018) Contextual augmentation: Data augmentation by words with paradigmatic relations. In: Proceedings of NAACL-HLT, pp. 452–457
6. Lison P, Tiedemann J (May 2016) OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp 923–929. https://www.aclweb.org/anthology/L16-1147
7. Liu X, He P, Chen W, Gao J (2019) Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 4487–4496
8. Mulcaire P, Kasai J, Smith NA (2019) Low-resource parsing with crosslingual contextualized representations. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL), pp 304–315
9. Roy D, Paul D, Mitra M, Garain U (2016) Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608
10. Settles B (2009) Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, Technical report
11. Sidorov G, Gelbukh A, Gómez-Adorno H, Pinto D (2014) Soft similarity and soft cosine measure: similarity of features in vector space model. Computación y Sistemas 18(3):491–504
12. Sun C, Huang L, Qiu X (2019) Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of NAACL-HLT, pp. 380–385
13. Tao T, Wang X, Mei Q, Zhai C (2016) Language model information retrieval with document expansion. In: Proceedings of the main conference on human language technology conference of the north American chapter of the association of computational linguistics. Association for Computational Linguistics, pp 407–414
14. Tenney I, Das D, Pavlick E (2019) Bert rediscovers the classical NLP pipeline. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 4593–4601
15. Tiedemann J (May 2012) Parallel data, tools and interfaces in opus. In: Chair NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the eight international conference on language resources and evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey
16. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Huggingface's transformers: State-of-the-art natural language processing. ArXiv abs/1910.03771
17. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Advances in neural information processing systems. pp 649–657

Response Generation to Out-of-Database Questions for Example-Based Dialogue Systems

Sota Isonishi, Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara

Abstract Example-based dialogue systems are often used in practice because of their robustness and simple architecture. However, when these systems are given out-of-database questions that are not registered in the question-response database, they have to respond with a fixed backup response, which can make users disengaged in the dialogue. In this study, we address response generation for out-of-database questions to make users perceive that the system understands the question itself. We define question types observed in the speed-dating scenario which is based on open-domain dialogue. Then we define possible response frames for each question type. We propose a sequence-to-sequence model that directly generates an appropriate response frame from an input question sentence in an end-to-end manner. The proposed model also explicitly integrates a question type classification to take into account the question type of the out-of-database question. Experimental results show that integrating the question type classification improved the response generation, and could exactly match 69.2% of response frames provided by human annotators.

1 Introduction

While various kinds of dialogue systems have been proposed and developed, example-based dialogue systems are still widely used as a simple, practical approach.

S. Isonishi · K. Inoue (✉) · D. Lala · K. Takanashi · T. Kawahara
Graduate School of Informatics, Kyoto University, Kyoto, Japan
e-mail: inoue@sap.ist.i.kyoto-u.ac.jp

S. Isonishi
e-mail: isonishi@sap.ist.i.kyoto-u.ac.jp

D. Lala
e-mail: lala@sap.ist.i.kyoto-u.ac.jp

K. Takanashi
e-mail: takanashi@sap.ist.i.kyoto-u.ac.jp

T. Kawahara
e-mail: kawahara@sap.ist.i.kyoto-u.ac.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_23

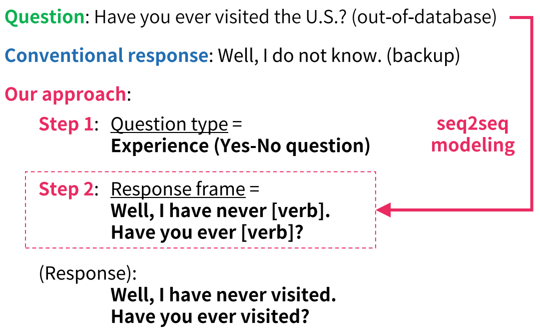


Fig. 1 Overview of response generation to out-of-database questions

Example-based dialogue systems use the database of pairs of an expected user question and the corresponding system response. When a question from the user is provided, an example-based dialogue system retrieves the nearest question (or the most related response) from a database using methods such as keyword matching or vector space modeling, and the corresponding response will be generated as the system response. Recently, the matching part has massively adopted neural networks [7, 15–20]. Those works have made the example-based dialogue systems more robust and also made it possible to increase the scale of the database.

The obvious limitation of example-based dialogue systems is that it is impossible to answer out-of-database questions—questions that are not registered in the database. In this case, the system has to generate the above-mentioned safe responses, which might degrade the naturalness of dialogue and decrease user engagement. Although it is important for example-based dialogue systems to register as many expected questions as possible, it is not practical to cover all possible questions, especially in practical dialogue systems that deal with a wide range of domains. Therefore, we need an effective approach to generate proper responses for out-of-database questions in order to retain the naturalness of dialogue and user engagement.

We propose response generation to out-of-database questions to make users perceive that the system is not able to answer but is able to understand the question itself. Our approach is summarized in Fig. 1. First, we classify the question type of the input question (Step 1). For example, the question type of "Have you been to the U.S.?" is "experience". Question type classification has been studied as a key function in conventional question answering systems [9–12]. In this study, we design a set of question types in the domain of speed-dating which is a dialogue of first encounters. Since it is open-domain dialogue, the proposed question type has the potential to be extended to other kinds of dialogue. Next, we choose from pre-defined response frames according to the classified question type to generate a response sentence (Step 2). Since the question type classification is error-prone and there are multiple possible response frames, we propose an approach that generates a proper response. In this study, we apply a neural network-based sequence-to-sequence (seq2seq) model to directly generate a response frame from an input question sentence in the manner of end-to-end modeling. To take into account question type classification, we

Fig. 2 Snapshot from speed-dating dialogue corpus (left: subject and ERICA, right: operator)

conduct pre-training of question type classification independently and integrate the pre-trained model to the seq2seq model. This study contributes to realizing example-based dialogue systems that are robust against out-of-database questions and continue dialogue without breakdown.

2 Human-Robot Speed-Dating Dialogue Corpus

We use a human-robot dialogue corpus where a human subject talked with an android robot ERICA [2, 3]. From this corpus, we chose one of the dialogue tasks, speed-dating which is an open-domain dialogue. In this task, participants meet each other and exchange their profile information to build the relationship between them. ERICA plays the role of a conversation practice partner for a male subject.

ERICA was operated by another human subject, called an operator. Figure 2 shows a snapshot of the dialogue recording. The operator's voice was directly played by a speaker placed on ERICA. The operator manually controlled non-verbal behaviors of ERICA such as eye gaze, head nodding, and hand gesture. With this setting, 31 dialogue sessions were recorded where each session lasted about 10 min. Whereas the subject was a different person in each session, the operator was one of four trained actresses. In advance, both operator and subject were given a list of dialogue topics which are generally talked about in first encounter dialogues.

For the current study, we extracted question sentences from this corpus. At first, we annotated dialogue act labels based on the standard definition [1] to recognize questions uttered by both ERICA and the subjects. To simplify the problem in this study, we use only questions that do not require the dialogue context to understand the meaning of the question. We call these questions base questions. On the other hand, questions that require additional dialogue context such as "Where was it?" were excluded from the current scope. The number of the base questions in the corpus was 370.

Table 1 Definition and example of question types

Question type	Definition
In-database	Questions about personal information of the system which can be searched in our database and frequently asked in first-encountering dialogue
	e.g. “What is your hometown?”, “What is your hobby?”
Habit	Questions about actions regularly done by the system
	e.g. “What do you do in your day off?”
Preference	Questions about preferences except those directly asking about hobbies
	e.g. “Do you like traveling?”
Experience	Questions about past actions
	e.g. “Have you ever been to the U.S.?”
Desire	Questions about future plans and hopes
	e.g. “How will you spend the next summer vacation?”
Subjective thoughts	Questions about subjective thoughts on something
	e.g. “Do you feel worthwhile to work?”
Personal information	Questions about personal information of the system which is not contained in our database
	e.g. “Do you have any pets?”
General knowledge	Questions about general knowledge, not related to the system
	e.g. “Do you know the Uyuni salt lake?”

3 Question Type

We now define question types and report a manual annotation for the speed-dating corpus including augmented data.

3.1 Definition of Question Type

We analyzed the base questions observed in the corpus and defined question types. The question types consist of *in-database*, *habit*, *preference*, *experience*, *desire*, *subjective thoughts*, *personal information*, and *general knowledge*. The definition and example of the question types are summarized in Table 1. Since the speed-dating corpus covers many topics for exchanging their profiles in open-domain, this set of question types will be applied to other kinds of social dialogue. Currently we have a database of frequently used question-answer pairs for ERICA, based largely on previous user interactions. We assume that *in-database* question types can be searched for within this database. Therefore, *in-database* questions are out of scope in the current study.

Furthermore, we took into account the form of questions because the corresponding response frames are different, which is explained in the next section. Based on

Table 2 Distribution of question types

Question type	Question form		Total
	Yes-No	5W1H	
Habit	115	120	235
Preference	100	122	222
Experience	97	99	196
Desire	58	80	138
Subjective thoughts	67	54	121
Personal info.	60	57	117
General knowledge	81	58	139
Total	578	590	1,168

the dialogue act labels, we classified the base questions into two forms: propositional questions (Yes-No) and set questions (5W1H). In this study, we consider the combination of question form (Yes-No or 5W1H) and the question type. We term this combination a *question type condition*.

3.2 Annotation of Question Type

We manually annotated the question type condition for the base questions observed in the speed-dating corpus. Since the number of question samples in the corpus is not enough for machine learning, we conducted data augmentation. We asked third-party augmenters to create base question sentences, giving the combination of a dialogue topic and a question type condition. Note that *in-database* questions were not The given dialogue topics are based on those observed in the speed-dating corpus. Each augmenter was given one of 14 question type conditions, excepting *in-database*. For example, when an augmenter was given a topic of *travel* and a question form of *Yes-No*, they created a question sentence like *Do you like traveling since you were young?*. We recruited four people as the augmenters and obtained 923 additional base question sentences.

Table 2 reports the distribution of the questions types. The total number of questions was 1,168. Note that *in-database* question are not included here. In the next section, we also annotate appropriate response frames for each question sentence. As a result, some question sentences did not correspond to any response frames in the viewpoint of context validity, so in later experiments we use 1,109 question sentences that were associated with adequate response frames. Although we regard all the 1,109 questions as out-of-database questions in the current study, this scope depends on the database of example-based dialogue systems.

Table 3 Result of question type classification

Question type	Question form					
	Yes-No			5W1H		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Habit	0.938 (106/113)	0.964 (106/110)	0.872	0.763 (87/114)	0.757 (87/115)	0.760
Preference	0.957 (88/92)	0.880 (88/100)	0.917	0.787 (107/136)	0.856 (107/125)	0.820
Experience	0.750 (75/100)	0.833 (75/90)	0.789	0.736 (67/91)	0.744 (67/90)	0.740
Desire	0.660 (33/50)	0.600 (33/55)	0.629	0.652 (45/69)	0.600 (45/75)	0.625
Subjective thoughts	0.754 (52/69)	0.800 (52/65)	0.776	0.789 (30/38)	0.600 (30/50)	0.682
Personal information	0.518 (29/56)	0.580 (29/50)	0.547	0.472 (17/36)	0.378 (17/45)	0.420
General knowledge	0.782 (68/87)	0.850 (68/80)	0.814	0.641 (25/39)	0.417 (25/60)	0.505

3.3 Question Type Classification

We conducted a preliminary experiment on question type classification. The input is a bag-of-words feature of function words, pronouns, adverbial nouns and adjective verbs. Note that we used only words that appeared in more than 1% of the training data. Furthermore, we also used a flag indicating whether the verb in the past tense. The output is a question type condition (14 types). Therefore, the current task is 14-class classification and we implemented this using a logistic regression model.

We evaluated the above model with 5-fold cross validation of the 1,109 question sentences. Table 3 reports the classification result. The overall accuracy of 0.746 was much higher than chance level (0.112) which classifies to the majority class (*preference*, 5W1H). We observed especially high scores on well-observed question types such as *habit*, *preference*, and *experience*. This result suggests the practicality of the question type classification in the current dialogue task.

4 Response Frame

For each question type, we designed several response frames that make users perceive that the system is not able to answer but is able to understand the question itself. Since there are several possible response frames for each question type, we then manually annotated appropriate frames for each question sentence.

4.1 Design of Response Frames

We manually designed several response frames for each question type condition (the combination of a question form and a question type). For example, if a question sentence is “Do you often watch movies?”, the question form is *Yes-No* and the question type is *habit*. One of the response frames is “Well, I do not [verb] [focus]. Do you often [verb] [focus]?”. The full set of response frames is summarized in Tables 4 and 5. Note that the response frames are designed in the Japanese language.

Each response frame consists of two parts: *reaction* and *question*. *Reaction* is the first part of the response frame and is used to return a denial answer to a user question. This denial answer prevents any elaboration on the out-of-database question that cannot be handled by the current example-based dialogue system. *Question* is the second part of the response frame, and is used to ask a question back to the user. Therefore, the system can take the dialogue initiative. In the above example, *reaction* corresponds to “Well, I do not [verb] [focus].” and *question* corresponds to “Do you often [verb] [focus]?”. When the dialogue system generates a response sentence based on this response frame, it is expected to make the user perceive that the system understood the question itself, and then also make it engaged in the dialogue. In this work, the proposed method generates response frames including slots without filling specific words in slots. The method to extract focus and verb words is left in future work, though it is possible to use tools from previous studies on morphological analysis [4] and focus word detection [5].

In each question type condition, we needed to create several response frames to cover the expected variation of the input question sentences. For example, on the question type condition of *preference* and *5W1H*, we created two kinds of response frames. The appropriate response frame depends on the input question sentence. If the input question is “What kind of food do you like?”, the response frame should be “Well, nothing in particular. What do you like?”. On the other hand, if the question is “Which football players do you like?”, the response frame should b “Well, nobody in particular. Who do you like?”. Furthermore, we also created several response frames depending on the presence of slot words. For example, on the question type condition of *preference* and *Yes-No*, we created two kinds of response frames depending on the presence of a focus word inside the question. If there is a focus word inside the question such as “Do you like football?”, the response frame should be “Well, not really. Do you like [focus = football]?” If a focus word is ambiguous, the response frame can be “Well, not really. Do you like that?”. From the above, there is no one-to-one relationship between the question type and the response frame.

4.2 Annotation of Corresponding Response Frames

We annotated appropriate frames for each question sentence. As each question sentence was already associated with a question type condition, we checked if each

Table 4 Response frame for Yes-No question form (✓ represents that the corresponding item, e.g. a focus word, must be found in an input question.)

Question type	Focus	Verb	Response frame
Habit	✓	✓	Well, I do not [verb]. Do you often [verb] [focus]?
	✓	–	Well, I do not do that. Do you often do [focus]?
	–	✓	Well, I do not [verb]. Do you often [verb]?
	–	–	Well, I do not do that. Do you often do that?
Preference	✓	–	Well, not really. Do you like [focus]?
	–	–	Well, not really. Do you like that?
Experience	–	✓	Well, I have never [verb]. Have you ever [verb]?
	–	✓	Well, I did not [verb]. Did you [verb]?
	–	–	Well, I have never done that. Have you ever done that?
	–	–	Well, I did not do that. Did you do that?
Desire	–	✓	Well, I think nothing in particular. Will you [verb]?
	–	–	Well, I think nothing in particular. How about you?
Subjective thoughts	–	–	Well, not really. How about you?
Personal information	–	–	Well, I am not sure about that. How about you?
	–	–	Well, nothing special. How about you?
	–	–	Well, not really. How about you?
General knowledge	–	–	Well, I do not know that. Do you know that well?

response frame candidate was appropriate in the viewpoint of the dialogue context and the presence of slot words inside the question sentence. Note that the number of appropriate frames was not restricted to one, which means that in some cases there are several frames for an input question sentence. For this question type condition, there are four response frames, but all frames contain a word "often" inside the response frame. This word is inconsistent with the meaning of the question sentence, so this question was annotated as there was no appropriate response frame. As a result, 1,109 of 1,168 question sentences were associated with more than one response frame. In this study, we do not use the 59 samples that were not associated with any frames. We use the pairs of the input question sentences and associated appropriate response frames in order to train a sequence-to-sequence model in the next section.

5 Response Frame Generation

To generate the appropriate response frame, we need to classify the question type condition and then find the appropriate response frame from several candidates. Since the question type classification is error-prone and there is ambiguity in the mapping from the question type condition to the appropriate frame, we propose sequence-to-

Table 5 Response frame for 5W1H question form (✓ represents that the corresponding item, e.g. a verb, must be found in an input question. WH represents an interrogative.)

Question type	Verb	WH	Response
Habit	✓	✓	Well, there it nothing [WH]. [WH] do you [verb]?
	✓	–	Well, I do not [verb] in particular. What do you [verb]?
	–	✓	Well, there is nothing [WH]. [WH] do you do?
	–	–	Well, I do not do in particular. What do you do?
Preference	–	–	Well, nothing in particular. What do you like?
	–	–	Well, nobody in particular. Who do you like?
Experience	✓	–	Well, I did not [verb] in particular. How about you?
	✓	–	Well, I did not [verb]. Did you [verb] anything?
	–	–	Well, I did nothing in particular. How about you?
	–	–	Well, I did not do that. Did you do anything?
Desire	✓	–	Well, I think nothing in particular.
			Is there any places you want to [verb]?
	✓	–	Well, I think nothing in particular.
			Is there anything you want to [verb]?
	–	–	Well, I think nothing in particular. Do you have any plan?
	–	–	Well, I think nothing in particular. Is there anything?
Subjective thoughts	–	–	Well, I do not think anything in particular. How about you?
Personal information	–	–	Well, I am not sure about that. How about you?
	–	–	Well, there are various. How about you?
	–	–	Well, nothing special. How about you?
General knowledge	–	–	Well, I do not know. Do you know that well?

sequence (seq2seq) modeling to generate the response frame directly from the input out-of-database question sentence. First, we introduce a simple seq2seq model. Then, we integrate the question type classification model in order to take into account which question type condition the input question sentence is.

5.1 *Simple Seq2seq Model*

The first model is a simple seq2seq model implemented by recurrent neural networks [14]. We use gated recurrent units (GRU) in the current study. The input is a word sequence of a question sentence and the output is a word sequence of a response frame annotated in the previous section. The input word is embedded in a distributed representation (word2vec) by using a continuous bag-of-word (CBOW) model [6]. This word embedding is trained independently from the training of the seq2seq model by using the same training data. Note that we gave a random vector of zero-mean and unit-variance for unknown words that did not appear in the training data. The output word sequence is searched using a greedy method. An input sequence sometimes corresponds to several output sequences when several response frames were annotated as adequate for the same input question sentence. In this case, we regard that there are several data pairs with the same input sequence and different output sequences, and we use them individually.

The seq2seq model has the potential to be able to handle two tasks, the question type classification and the response frame selection, with a simple architecture in an end-to-end manner. Another advantage of using the seq2seq model for the current task is the extendability of the model. Even if we use more variations of output response frames, the seq2seq model can easily handle the increase of the variation of output sequences. On the other hand, if we use another model that selects the kind of output response frame instead of generating a response frame itself, the output dimension would increase with the number of response frames, which is inefficient. In future work, we also plan to directly generate a response sentence, automatically filling in the slot words. In this case, one possible and sophisticated way is to extend the seq2seq model to something like a pointer network which directly generates a response sentence by referring to some words in the input sequence [8, 13].

5.2 Integration with Question Type Classification

We extend the simple seq2seq model. In our preliminary experiment in Sect. 3.3, we confirmed that we can classify the question type with a reasonable accuracy. The information about question type is useful for the seq2seq model to generate an adequate response frame. Accordingly, we propose to integrate the question type classification into the seq2seq model. Figure 3 illustrates the architecture of the proposed model. We concatenate the output of the logistic regression of the question type classification (14 dimensions) and the output of the encoder of the seq2seq and then feed the concatenated vector to the decoder. In training, we pre-train the logistic regression of question type classification and fine-tune the entire network. This integrated network is expected to generate a response frame which considers the question type.

6 Evaluation

We evaluated response generation to out-of-database questions with the seq2seq models.

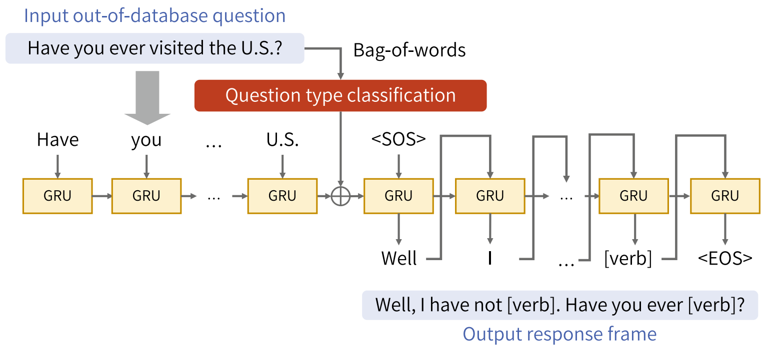


Fig. 3 Seq2seq model generating a response frame by taking into account question type classification

6.1 Setup

We conducted 5-fold cross validation of the 1,109 question sentences. Since each input question sentence can correspond to more than one output response frame, the number of training data pairs was 1,016. The evaluation metric was an perfect matching between a generated response frame and a reference response frame. We regarded that it was correct if the generated response frame matched one of the corresponding reference response frames.

We also implemented a baseline method that is based on the question type classification method described in Sect. 3.3. This model first classifies the question type condition and consequently select a response frame that is the most frequent one in the training dataset within the classified question type condition. We compared the two seq2seq models with this baseline model.

The parameters of the seq2seq models are as follows. We use one layer of GRUs for the encoder and decoder of the seq2seq models. The number of hidden units of the GRUs was 128. Note that when we integrated the question type classification, the number of hidden units of the decoder increased to 142 ($128 + 14$). The dropout ratio was 20%. We regarded one sequence as one minibatch. The number of training epochs was 100.

6.2 Result

Table 6 reports the ratios of perfect matching between the generate response frame and the reference response frame. Compared with the baseline model, the seq2seq models showed higher scores, which demonstrates the effect of this end-to-end modeling. The integrated seq2seq model increased the matching ratio to 0.692 from 0.672.

Table 6 Evaluation of response generation

Method	Perfect matching ratio
Baseline	0.635
Simple seq2seq	0.672
Integrated seq2seq	0.692

This improvement shows that it is effective to explicitly take into account the kind of question type condition in the current response generation task.

6.3 Generated Examples

We report several generated examples by the seq2seq models. Examples have been translated from Japanese, so the grammar of these may not be necessarily correct in English. Correct examples of both seq2seq models are shown below. Note that **Q** and **R** are an input question sentence and a generated response frame, respectively.

Q1 Do you go to music concerts? (Habit, Yes-No)
R1 Well, I do not [verb]. Do you often [verb] [focus]?
Q2 What do you do on your days off? (Habit, 5W1H)
R2 Well, there is nothing [interrogative] I do in particular. [interrogative] do you [verb]?
Q3 Have you ever kept any pets? (Experience, 5W1H)
R3 Well, I did not [verb] in particular. How about you?

Some examples where the simple seq2seq model failed but the integrated one succeeded are follows. Note that **Si** is the generated response frames by the simple seq2seq and **In** is one generated by the integrated seq2seq model, respectively.

Q4 Do you know a music instrument called a Cajon? (General knowledge, Yes-No)
Si4 Well, there it nothing [interrogative]. [interrogative] do you [verb]? (Habit, 5W1H)
In4 Well, I do not know. Do you know that well?
Q5 What do you do when you meet your family? (Habit, 5W1H)
Si5 Well, I think nothing in particular. How about you? (Desire, 5W1H)
In5 Well, there is nothing [interrogative]. [interrogative] do you [verb]?

In the above examples, the generated response frame by the simple seq2seq frame came from an incorrect question type condition. Since the integrated seq2seq model explicitly takes into account the kind of question type condition, these cases were enhanced.

Table 7 Result of human evaluation

		Human evaluation		Total
		Accept	No	
Generation	Correct	57	12	69
	Wrong	16	15	31
Total		73	27	100

6.4 Human Evaluation

We conducted a human evaluation to confirm the appropriateness of the generated response frames. We recruited three human evaluators. From the response frames generated by the integrated seq2seq model, we randomly selected 100 samples keeping the balance of correct and wrong samples. We showed each pair of the input question sentence and the generated response frame and then asked each evaluator to judge if he/she feels that the system understands the question itself. Note that we also asked the evaluators to interpolate adequate slot words by themselves as much as possible. The evaluator had to select one of three choices: *accept* (the system understands the question), *not accept* (not understand), or *neither*. We used the majority voting that regarded the sample as *accepted by human* if more than two persons selected *agree* for the sample.

Table 7 reports the result of the human evaluation. In total, 73 of 100 samples were accepted by the evaluators, slightly higher than the perfect matching ratio. In the correct samples, 57 of 69 samples were accepted by human. The remaining 12 samples were not accepted even though the response frame was correctly generated. This suggests that the set of response frames needs to be revised in future work. Within the wrong samples, surprisingly, 16 of 31 samples were accepted. When we analyzed these samples they seemed to be grammatically incorrect but semantically acceptable and meaningful. These types of response can be accepted and used in conversations such as speed-dating dialogue where engagement is more important than the grammatical correctness.

7 Conclusion

We have addressed response generation to out-of-database question for example-based dialogue systems to make users perceive that the system understands the question itself. We defined question types such as *habit* and *preference* based on observation of speed-dating first-encountering dialogue which is open domain. We then designed response frames for each combination (called question type condition) of the question type and the question form (Yes-No or 5W1H). The sequence-to-sequence neural network model was used to directly generate the adequate response frame from the input question sentence. Meanwhile, we confirmed that we could classify the question type condition with an accuracy of 0.746. Therefore, we integrated the question type classification (logistic regression) into the seq2seq model to explicitly consider the question type condition. This model achieved an exact sequence matching ratio of 0.692, improving the basic seq2seq model. Finally, we confirmed that human evaluators accepted 73 of 100 generated samples in the viewpoint of whether the system understands the question itself.

Our future work is as follows. In the current study, our model generates response frames that contains slots for entries such as focus and verb words. To fill in the slots, we will first use conventional tools such as morphological analysis to extract the slot values. As a more sophisticated approach, we will investigate how to extract the slot values by using an integrated neural network such as pointer networks [8, 13]. We will also conduct a subjective evaluation for the generated response sentences including slot values. Finally, we will integrate the response generate to out-of-database questions into current example-based dialogue systems to apply them in practice.

Acknowledgments This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant number JPMJER1401) and Grant-in-Aid for Scientific Research on Innovative Areas "Communicative intelligent systems towards a human-machine symbiotic society" (Grant number JP19H05691).

References

1. Bunt H, Alexandersson J, Carletta J, Choe JW, Fang AC, Hasida K, Lee K, Petukhova V, Popescu-Belis A, Romary L, Soria C, Traum D (2010) Towards an ISO standard for dialogue act annotation. In: LREC, pp 2548–2555
2. Inoue K, Milhorat P, Lala D, Zhao T, Kawahara T (2016) Talking with ERICA, an autonomous android. In: SIGDIAL, pp 212–215
3. Kawahara T (2018) Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS
4. Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: EMNLP, pp 230–237
5. Lala D, Milhorat P, Inoue K, Ishida M, Takanashi K, Kawahara T (2017) Attentive listening system with backchanneling, response generation and flexible turn-taking. In: SIGDIAL, pp 127–136
6. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: ICML, pp 1188–1196
7. Lowe R, Pow N, Serban I, Pineau J (2015) The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: SIGDIAL, pp 285–294
8. Merity S, Xiong C, Bradbury J, Socher R (2017) Pointer sentinel mixture models. In: ICLR
9. Mitamura T, Lin F, Shima H, Wang M, Ko J, Betteridge J, Bilotti MW, Schlaikjer AH, Nyberg E (2007) JAVELIN III: cross-lingual question answering from Japanese and Chinese documents. In: NTCIR
10. Mizuno J, Akiba T, Fujii A, Itou K (2007) Non-factoid question answering experiments at NTCIR-6: towards answer type detection for realworld questions. In: NTCIR
11. Shima H, Mitamura T (2007) JAVELIN III: answering non-factoid questions in Japanese. In: NTCIR

12. Tamura A, Takamura H, Okumura M (2005) Classification of multiple-sentence questions. In: IJCNLP, pp 426–437
13. Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. In: NIPS, pp 2692–2700
14. Vinyals O, Le Q (2015) A neural conversational model. In: ICML deep learning workshop
15. Wang H, Lu Z, Li H, Chen E (2013) A dataset for research on short-text conversations. In: EMNLP, pp 935–945
16. Wang M, Lu Z, Li H, Liu Q (2015) Syntax-based deep matching of short texts. In: IJCAI, pp 1354–1451
17. Wu Y, Wu W, Xing C, Zhou M, Li Z (2017) Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots. In: ACL, pp 496–505
18. Yan R, Song Y, Wu H (2016) Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: SIGIR, pp 55–64
19. Zhou X, Dong D, Wu H, Zhao S, Yu D, Tian H, Liu X, Yan R (2016) Multi-view response selection for human-computer conversation. In: EMNLP, pp 372–381
20. Zhou X, Li L, Dong D, Liu Y, Chen Y, Zhao WX, Yu D, Wu H (2018) Multi-turn response selection for chatbots with deep attention matching network. In: ACL, pp 1118–1127

Packing, Stacking, and Tracking: An Empirical Study of Online User Adaptation

Jean-Sébastien Laperrière, Darryl Lam, and Kotaro Funakoshi

Abstract This paper explores the application of expert tracking to online user adaptation based on a set of basic predictors in order to classify input in multimodal interaction settings. We compare the performances of this approach to other common approaches that aggregate multiple predictors, like stacking and voting. To realistically assess the performances of algorithms that require feedback, we added noise to feedback to simulate an imperfect system. Using two datasets, we obtained inconsistent results. With one dataset, expert tracking was the best option for short interactions, but with the other dataset, it was outperformed by other algorithms. In contrast, voting worked surprisingly well. On the basis of these results, we discuss implications and future directions.

1 Introduction

To interact with humans, a system must be equipped with machine learning-based modules that accurately recognize various social signals [17] from different users. In a training dataset composed of interactions with many users, only a subset of the dataset might be similar enough to approximate a new user's behavior. Finding and using that subset of similar users in an online adaptation manner could substantially increase the accuracy of a module.

K. Funakoshi - (His current affiliation is Tokyo Institute of Technology, Tokyo, Japan).

J.-S. Laperrière
Polytechnique Montréal, Montréal, Canada
e-mail: jean-sebastien.laperriere@polymtl.ca

D. Lam
University of British Columbia, Vancouver, Canada
e-mail: darryl.lam@alumni.ubc.ca

K. Funakoshi (✉)
Kyoto University, Kyoto, Japan
e-mail: funakoshi@lr.pi.titech.ac.jp

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture
Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_24

In this paper, we explore *expert tracking* [2, 10, 13, 23] in terms of online user adaptation in multimodal interactions. An expert tracking algorithm finds the most reliable predictors (experts) among all experts and bases its predictions on the best experts. By regarding as experts the prediction models that are individually built for different users or groups, expert tracking would work as an online adaptation method that finds relatively better models to work with a new user in accordance with feedback obtained through interaction with the user. We define and test a few feedback settings to empirically determine the feasibility and limitations of expert tracking in a real-world interaction. We compare the performance of expert tracking to other common approaches, that is, *stacking* [7, 14, 24], *majority voting* [5], and *packing*. Here, we use the term *packing* to refer to a single bulk model trained with all available data before an interaction.

The primary contribution of this paper is that, to the best of our knowledge, this is the first effort to apply expert tracking (hereafter, ET) to online user adaptation in multimodal human-machine interaction. Our contribution consists of multiple experimental results and lessons learned from the results:

- an empirical analysis of an ET algorithm under different feedback settings
- a detailed comparison of the performances of ET and other approaches and
- the identification of the best practice and future directions.

In the rest of the paper, Sect. 2 first describes ET and how its use could benefit user adaptation. Next, stacking, a statistical ensemble learning method that can also be applied to online user adaptation, is described in Sect. 3. Then, Sect. 4 overviews the datasets used in our experiments. The experimental settings and results are presented in Sects. 5 and 6, respectively. Section 7 discusses the implications of those results, and Sect. 8 concludes this paper.

2 Expert Tracking and Online User Adaptation

In this section, we first describe the general concept of expert tracking (ET). Then, we discuss the application of ET to online user adaptation and consider a particular ET algorithm that is examined in this paper.

2.1 Expert Tracking

The simple idea behind ET is to find the most reliable predictor (expert) among all experts available for a particular prediction task and base predictions on the best expert [13, 23].

Initial input:

- A set of N predictors (experts).
- A weight vector $\boldsymbol{w} \in \mathbb{R}^N$ with initial values (*e.g.*, $w_i = \frac{1}{N}$ where $i \in [1, N]$).

For each trial t:

1. Receive the vector of the N experts' predictions $\boldsymbol{x}_t \in \mathbb{R}^N$.
2. Predict $\hat{y}_t$ (the output at t) with the predictions $\boldsymbol{x}_t$ and the weights $\boldsymbol{w}$, *i.e.*, $\hat{y}_t = \mathbf{pred}(\boldsymbol{x}_t, \boldsymbol{w})$.
3. Compute the loss of each expert after receiving outcome y_t.
4. Update the weights $\boldsymbol{w}$ using the losses.

Fig. 1 Generic procedure of expert tracking

Figure 1 shows the generic procedure of ET. Concrete details for steps 2 to 4 are given for each particular ET algorithm. Each expert can be given as a prediction module that uses any method including ad hoc rules based on heuristics and statistical machine learning, independently of the ET algorithm.

In general, as the subject (*i.e.*, the target and nature of the prediction task concerned as in the after-mentioned example situations) can change (subject shift) over time, so can the reliability of experts (expert shift). Therefore, when an ET algorithm receives an actual outcome as feedback after each trial, it must dynamically adapt the weight given to each expert to follow those shifts as closely as possible. To do so, a loss function quantifies the performance of the experts, and the expert tracker uses these values to rebalance how it will use the predictions that it receives from the experts in the next trial.

One significant qualitative difference between ET and common statistical machine learning is that it does not presuppose stable statistical distributions behind sequential observations along time line. Therefore, it may work out better than statistical approaches against possible but problematic situations in multimodal human-machine interactions such as:

- a restless user changing posture suddenly (*e.g.*, standing → sitting)
- a user changing behavior suddenly in response to a change in the environment (*e.g.*, their boss coming into their office) *and*
- a user gradually changing their disposition or interest and at some moment, causing their feedback to almost contradict their previous feedback.

2.2 Tracking as Online User Adaptation

As mentioned in the introduction, this paper presupposes the use of a training dataset composed of interactions with many users (and the dataset can be partitioned per user). Because users often behave differently, only a subset of the dataset might be similar enough to approximate the behavior of a new user who does not reside in the training dataset.

Given this perspective, for any specific prediction task, by regarding as experts the prediction models that are individually built for different users or groups, ET can work as an online adaptation method that changes the weights of given experts (*i.e.*, the subsets of the dataset) in accordance with the feedback obtained during an interaction with a new user.

Formally speaking, given a training dataset D of N users, *i.e.*, $D = \{D_1, \ldots, D_N\}$, a predictor for a particular task in a multimodal interaction system is created for each user data $D_j \in D$. Those predictors are input as experts to an expert tracker in advance as described in Fig. 1. Then, the tracker makes prediction $\hat{y}_t$ for each input from a new user with whom the system is currently interacting. The tracker updates weights $\boldsymbol{w}$, namely, adapts to the new user in an online manner while receiving outcomes through the interaction.

This paper focuses on exploring the feasibility of this data-subset-wise multi-expert approach. Note that, in a theoretical sense, this approach (the proposed use of ET as an online adaptation method) and any other online adaptation methods which are applicable to a single prediction model do not compete with each other (or are not mutually exclusive). That is, one can apply any other online adaptation to each single expert in conjunction with this ET-based multi-expert approach if it is feasible and useful.

In a realistic interaction context, apart from basic inferences that can be reached by observing visible attributes (*e.g.*, gender and physique [22]), an ET algorithm typically starts from scratch with every new user (*i.e.*, all of the elements of $\boldsymbol{w}$ in Fig. 1 are initialized with a flat weight of $1/N$). The algorithm must then rely on feedback obtained in real time. The pace of human interactions being quite slow, around an utterance every few seconds, an expert tracker must rapidly converge on the best expert that can reliably predict the current user's behavior. However, the feedback can contain noise when feedback signals are recognized by other automatic modules.

ET cannot simply be locked into a state once a good expert is found in the initial adaptation. It might need to readjust the weights of the experts later in the interaction due to a slight subject shift. Indeed, the user might change his behavior depending on the flow of the conversation. For example, the subject might adopt a more robotic posture and voice if the model has difficulty understanding requests or answers or relax and speak more casually if the model responds well.

In short, user adaptation is a real challenge for ET as it requires a rapid focus on the best expert and a tolerance to noise in feedback.

2.3 Bousquet's Algorithm

For this study, we chose an ET algorithm developed by Bousquet and Warmuth [2].[1] Bousquet's algorithm was one of the earliest algorithms to model expert shifts.

Bousquet's tracking algorithm keeps track of all of the past weights for all of its experts and uses a mixing scheme to decide the new weights for the next trial. After each trial, when it receives feedback, the current weights are adjusted in accordance with the experts' losses. The reactivity of the weights is determined by the importance of the latest trial in accordance with a given mixing scheme, which is chosen in accordance with each problem domain. The reader is referred to the original paper [2] for the detailed algorithm and the definitions of various mixing schemes.

3 Stacking

Stacking is an ensemble learning technique and trains a blending model to combine the predictions of multiple basic models [25]. Stacking is a popular approach in many fields because a blending model often outperforms the basic models that compose it [7, 14, 24].

3.1 Stacking as Online User Adaptation

In this work, the blending model aggregates the predictions from the basic models (*i.e.*, experts) in the same way as expert tracking. Stacking is fundamentally not an online scheme but a batch one. To use stacking in an online adaptation manner, we simply retrain the blending model with all of the available data once it receives a new feedback label for the last prediction. The pair of a received feedback label and the corresponding predictions of the basic models as a feature vector is stored as a new instance in the training data for the blending model. The initial training data contains only two instances that are meant to represent trivial cases; if every expert's prediction is positive, predict positive; if they are all negative, predict negative.

[1] We tested another recent tracking algorithm, CBCE [10]. In a simple simulated situation, we confirmed the superiority of CBCE to Bousquet's, as was claimed. However, in none of the settings examined in this paper did CBCE outperform Bousquet's. Therefore, we omit the results with CBCE due to space limitations.

3.2 Data Augmentation for Stacking

Using only past trials as the training data at the very early stage of adaptation is problematic in the above framework. Mainly, the size of the training data is very small, and the balance of the classes (positives and negatives) is left to chance. To lessen the former problem and eliminate the latter, we augment the training data by systematically adding the inverse of each example. Let us say that the first-level predictions of four experts are [1, 0, 1, 1] and the feedback label is 1 (positive). Then, we also create an artificial data instance with the inverted feature values [0, 1, 0, 0] and with the inverted label 0 (negative) to add to the training data. This balances the training data for the blending model while highlighting the reliability of each expert without linking experts to a specific label. We found that augmenting the training data in this way improved the performance.

4 Datasets

Here, the two datasets used for our experiments are described. One dataset is based on a corpus of multimodal human-robot interactions. The other dataset is based on another corpus of multimodal human-agent interactions.

4.1 MPR Corpus and ROE Task

4.1.1 MPR Corpus

The multiparty robot (MPR) dialogue corpus consists of two editions of multimodal human-robot dialogues in Japanese [6]. The first edition was collected in 2012 by using Microsoft Kinect V1 and the second edition was collected in 2016, by using Kinect V2. In each of the collected dialogues, a small humanoid robot (Aldebaran NAO) and a group of three people who were familiar with each other (friends, family members, or colleagues) participated in one-to-many style quiz games.

4.1.2 ROE Task

Sugiyama *et al.* [21] defined the task of response obligation estimation (ROE) and evaluated their proposed ROE method with a part of the 2012 edition of the MPR corpus. ROE is a binary classification task of estimating whether an input sound should be responded to by a robot or not. Input sounds are all that occur while the robot is interacting with people. It is an extended integration of noise rejection [12] and addressee identification [15]. Each sound segment is classified into "ought-to-respond" (positives) or "ought-not-to-respond" (negatives). While the robot will respond to input sounds classified as the former, it will not respond to those classified as the latter as those are monologues, user utterances to other users, surrounding noises, or user utterances to which the robot need not respond to even if they are directed to the robot (*e.g.*, expressions of impressions). For the ROE task, Sugiyama *et al.* [21] proposed 1 dialogue act feature and 49 non-verbal features that exploit prosody, face, and motion information. We use the 49 features for the sake of simplicity.[2]

4.1.3 MPR Dataset

Hereafter, MPR refers to a dataset based on the 2012 edition of the MPR corpus. While the evaluation of the original ROE method was done on the basis of the manual annotation of participants' speech segments [21], this dataset was created by processing 14 dialogues with an automatic voice activity detector and was provided with the labels of one annotator.

The dataset was the set of labeled feature vectors for the ROE task using the MPR corpus. Each dataset was separated into exclusive subsets. Each subset was a list of labeled feature vectors (instances) from the three members of a group. As those members are friends or family, they often exhibited similar behaviors in the group. Therefore, we regarded a set of instances from the three members as that from one virtual individual subject. We refer to this subset (group) as "subject" hereafter.

4.2 JMD Corpus and MIE Task

We also examined the performances with a dataset from a different corpus and a different task. The dataset shares the same format and structure as the aforementioned ROE dataset.

4.2.1 JMD Corpus

We refer to the multimodal corpus of human-computer dialogues in Japanese [1, 11] as the JMD corpus,[3] in which the participants talked about various news topics with a Wizard-of-Oz dialogue system, represented as a virtual agent on the screen, in a one-to-one situation.

[2]While the dialogue act feature requires costly annotation work, the contribution of this feature to the overall performance is limited. It is the second least contributing one among the seven feature sets investigated in the ablation study in [21].

[3]The corpus is now officially named as *Hazumi* corpus.

Table 1 MIE dataset's feature sets

Feature set	# features	# after PCA
Face	35	35
Dialogue	12	12
Voice	384	30
Text	951	50

4.2.2 MIE Task

The JMD corpus is designed especially for the multimodal interest estimation (MIE) task, which is another task of classifying a user's spoken response to a system's suggestion on the basis of whether the user is interested in the topic (positive) or not (negative).

4.2.3 MIE Dataset

The MIE dataset consists of 39 subjects. With regard to this dataset, each "subject" (see Sect. 4.1.3) corresponds to a real individual person.

The features in this dataset are grouped into four feature sets, *i.e.*, Face, Dialogue, Voice, and Text in accordance with [16].[4] Face features are the landmarks extracted by using Dlib.[5] Dialogue features consists of the response time of an user utterance, the number of the words in the previous system utterance, the number of the words in the user utterance, the difference of the two word numbers, and dialogue act type of the user utterance (one-hot vector of eight types: information-offer, positive-answer, negative-answer, other-answer, opening-new-topic, wh-question, yes-no-question, and suggestion). Voice features are the acoustic features used in [19]. Text features includes the frequencies of the word stems that appeared in the corpus more than one time in addition to frequencies of four parts of speech: noun, adverb, adjective, and interjection.

Table 1 shows the number of original features per feature set and the number used in the experiments. We reduced the numbers of Voice and Text features by using primary component analysis (PCA) as PCA performed better in a preliminary experiment.

This task is more subjective than the ROE task. Therefore, the correct labels were assigned by majority voting of six or three external judges.

[4][8] uses an extended version of the features used in this work. The description of the feature extraction process in [8] is mostly applicable to the features used in this work.

[5]http://dlib.net.

Table 2 Label and subject counts for datasets

Dataset	Negatives	Positives	# subjects
MPR	1382	1267	14
MIE	1416	1162	39

Table 3 Major qualitative differences between datasets

Dataset	Task subjectivity	Labeling manner	Speech segmentation
MPR	Low	Roughly done by 1	Automatic
MIE	High	Voting of 6 or 3 (moderate agreement)	Manual

4.3 Characteristics of Datasets

4.3.1 Label Distributions

Table 2 shows the label distribution and the number of subjects for each dataset. Though the imbalance of the total data is not so significant, the distribution of the labels varied greatly from subject to subject in each dataset. Therefore, we augmented the training data for each subject to reach parity between positive and negative labels by using random oversampling [9].

4.3.2 Qualitative Differences

Table 3 shows the qualitative differences between the datasets that impacted the difficulty of the tasks and the qualities of the datasets. There were three major factors that varied among the datasets.

Task Subjectivity

The subjectivity of the ROE task was not high because it is not difficult for a human observer to decide the correct label for most inputs in the ROE task on the basis of the content of spoken utterances. In comparison, the subjectivity of the MIE task was higher because the content of spoken utterances is not that reliable of an indicator of interest.

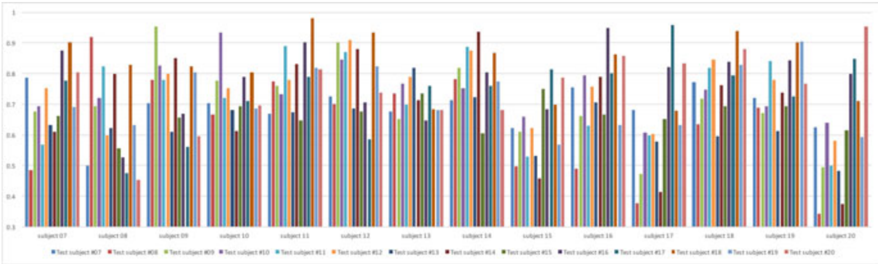


Fig. 2 MPR ROE task performances in all-play-all setting

Labeling Manner

Given the task subjectivity, the ROE datasets were labeled by only one person. The MIE dataset was labeled by majority voting of six or three people.

Speech Segmentation

Data samples of all of the datasets used in this paper were based on speech segments. During the dataset preparation, sound segments were extracted by an automatic voice activity detector for MPR. For MIE, sound segments were extracted on the basis of manual sound annotation.

4.3.3 Differences Among Subjects

Figure 2 shows the ROE task performance results from an all-play-all manner experiment[6] with subjects 07 to 20 in the MPR dataset. The graph demonstrates the significant differences among subjects. For example, the model trained with subject 20 (the right-most) performed with subject 20 the best, and it worked fine with subjects 16 and 17. However, its performance with subjects 08 and 14 was terrible. The model of subject 13 worked very evenly but moderately with all of the other subjects. This suggests that user adaptation combining multiple different models may perform better than a single general model built with all of the available data.

[6] A model is trained with the data of one subject and is tested with all of the subjects including itself individually.

5 Experimental Settings

This section describes baseline methods without adaptation, evaluation metrics, and other experimental settings. Hereafter, we refer to a prediction model trained with the data from a single subject in a dataset as an "expert."

5.1 Baselines (non-Adaptation Methods)

5.1.1 Voting

This is the simplest algorithm, *i.e.*, majority voting by all available experts. Each expert outputs a prediction as a real number in [0, 1]. The average of the predictions is then rounded to obtain the final prediction. Such majority voting is also used in ROVER [5], which is a popular method for integrating multiple speech recognition results.

5.1.2 Packing

Packing refers to a single prediction model built in the same way as with experts, except for one point; it is trained with all data from all available subjects. The original ROE evaluation experiment in [21] corresponds to this method.

We see the comparison between voting and packing as a debate over a single question: genericity versus specificity. Packing represents genericity, and voting represents specificity.

5.2 Experimentation Setups

5.2.1 Experimentation Method

Because we wanted to observe the performances of the algorithms in adapting to unseen subjects, we performed our experiments in a leave-one-subject-out (LOSO) manner.

At the beginning of each round of LOSO, the ET algorithms and the blending model of stacking were refreshed to the initial states. This means we assumed that the deployed systems could detect when subjects were switched by any means, *e.g.*, face recognition.

We also assumed that a system could recognize feedback signals from users by any means.[7] Therefore, for the algorithms that rely on feedback (tracking and stacking), we simulated feedback signals by using the correct labels in the test data. That is, an adaptation algorithm receives the feedback (hit or miss) for each prediction immediately. To simulate errors in real feedback, we added noise to the simulated feedback. In this binary feedback context, we simulated noise by randomly flipping the specified proportion of the labels.

5.2.2 Implementations and Parameters

All experts were built as a random forest model with 50 decision trees with a maximum depth of 11 following our preliminary investigation. To ensure a fair comparison, all of the algorithms that base their final predictions on those of experts use the exact same instances of random forest models because the randomness in the random forest algorithm creates fluctuations in the resulting models even with the exact same training data.

We used the Python scikit-learn implementation of random forest.[8] We implemented Bousquet's algorithm by ourselves. For stacking, we used our modified version of an open source framework,[9] which is adapted for models trained with different data. The blending model used for stacking was a logistic regression model also from scikit-learn.

The learning rate of Bousquet's algorithm was left to its default of $\eta = 1$. In our preliminary experiment, the default rate seemed to be the best compromise between adaptability and resilience to errors in feedback. As for the mixing scheme, we used "Fixed Share to Decaying Past." Like in the original paper, this mixing scheme led to the best results when its parameters (α, γ) were set to (0.3, 2).

6 Results

Now we show the performances of all of the algorithms mentioned in the previous sections, *i.e.*, tracking (Sect. 2), stacking (Sect. 3), and voting and packing (Sect. 5.1), under multiple settings to observe their behavior during simulated interactions of variable length by using the datasets described in Sect. 4. All of the results are averages of three trials with different random seeds.

Our results are separated into two large settings. General observations concerning the performances of the algorithms were gathered in long term simulations (Sect. 6.1).

---

[7] One may deploy automatic recognition modules for natural reactions from users as discussed in [6] or may adopt any specially designed interaction devices so that the users can provide feedback precisely but easily.

[8] https://scikit-learn.org/stable/modules/ensemble.html#random-forests.

[9] https://github.com/dustinstansbury/stacked_generalization.

We recognized the major merit of ET in the early stages of interactions, where the accumulated data for each subject was not enough to build an individualized model yet. Therefore, we observed the performances in shorter simulations, too (Sect. 6.2).

6.1 Run-Thorough Experiment

To observe the general long-term performance of the algorithms, we first tested the models on all of the data instances. Table 4 shows the performance for each dataset separately. In the tables, we shorten precision to P, recall to R, and F1-score to F. In each table, the better approach between voting or packing in terms of macro average F1 is indicated in italic. The adaptation performances better than the emphasized baselines are indicated in bold.

The results can best be described as inconsistent. ET was almost the best option for MPR and was not too far behind the best options for MIE. Since ET performed worse than voting, its default state, user adaptation seemed to be unsuccessful on MIE.

Stacking performed well only for MPR and only with low noise. Tracking algorithm on two of the three MPR datasets. The noise generally affected the stacking algorithm more than tracking.

Voting was unexpectedly competitive. It outperformed packing in the two datasets and was the best algorithm for MIE.

6.2 Initial Adaption Experiment

Next, we observed how well tracking and stacking performed in the early stages of interactions. For this experiment, we used only the first few t trials of each test set. By repeating the test with different t, we could observe the adaptation of the expert trackers and the stacking algorithm. Figure 3 shows the performances of the algorithms in terms of macro average F1 score for $t = 10, 20, 30$, respectively.

For MPR, the expert trackers rapidly surpassed voting (their initial state) and almost reached their performance in the long-term simulations. For MIE, however, the opposite happened; the tracking algorithms' performances stabilized in a state consistently worse than voting.

Generally, stacking seems to adapt more slowly than tracking. While stacking outperformed both tracking algorithms during long simulations on MPR, stacking was consistently outperformed by tracking in short simulations, from 10 to 30 trials and from 0 to 30% noise.

Stacking was much more vulnerable to noise in short simulations. This was expected because shorter simulations imply a smaller training set.

Since they do not require feedback, the performances of voting and packing obviously nearly matched their performances in longer simulations.

Table 4 Full length run-thorough experiments

MPR	Negative class			Positive class			Average F1
	P	R	F	P	R	F	
Tracking: 0% noise	0.818	0.877	0.846	0.855	0.787	0.819	**0.833**
Tracking: 10% noise	0.822	0.872	0.846	0.850	0.793	0.821	**0.833**
Tracking: 20% noise	0.813	0.843	0.828	0.822	0.789	0.805	**0.816**
Tracking: 30% noise	0.790	0.792	0.791	0.773	0.770	0.772	**0.781**
Stacking: 0% noise	0.829	0.875	0.852	0.856	0.803	0.829	**0.840**
Stacking: 10% noise	0.818	0.874	0.845	0.851	0.788	0.818	**0.832**
Stacking: 20% noise	0.798	0.853	0.824	0.827	0.764	0.794	**0.809**
Stacking: 30% noise	0.766	0.809	0.787	0.779	0.731	0.754	0.770
Voting	0.777	0.794	0.785	0.770	0.751	0.760	*0.773*
Packing	0.733	0.746	0.739	0.717	0.703	0.710	0.725
MIE	**Negative class**			**Positive class**			**Average F1**
	P	R	F	P	R	F	
Tracking: 0% noise	0.723	0.751	0.737	0.682	0.649	0.665	0.701
Tracking: 10% noise	0.722	0.754	0.737	0.683	0.646	0.664	0.701
Tracking: 20% noise	0.719	0.745	0.732	0.675	0.644	0.659	0.695
Tracking: 30% noise	0.715	0.726	0.720	0.660	0.647	0.653	0.686
Stacking: 0% noise	0.732	0.653	0.690	0.626	0.709	0.665	0.678
Stacking: 10% noise	0.724	0.641	0.680	0.616	0.702	0.657	0.668
Stacking: 20% noise	0.710	0.630	0.668	0.604	0.687	0.643	0.655
Stacking: 30% noise	0.686	0.601	0.640	0.577	0.664	0.618	0.629
Voting	0.732	0.785	0.758	0.712	0.648	0.678	*0.718*
Packing	0.740	0.743	0.741	0.683	0.679	0.681	0.711

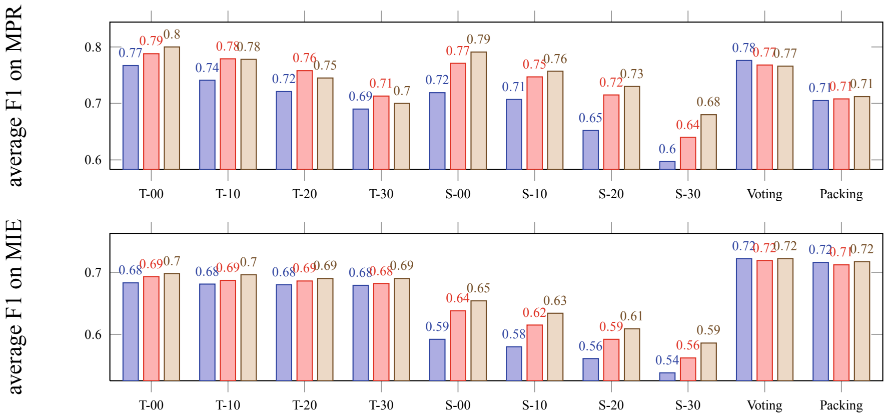


Fig. 3 Macro average F1 scores of algorithms when using first 10/20/30 trials (left to right) on MPR and MIE (top and bottom). T and S refer to tracking and stacking with noise ratio, respectively. *E.g.*, T-10 means tracking with 10% simulated noise in feedback

7 Discussion

On the basis of the observations on the results presented in Sect. 6, we discuss possible implications and recommendations. First of all, simple voting seems to work better than packing on subject-wise datasets like those examined in this paper. It is worth considering voting over packing if the computational cost is affordable. The reason that voting works better than packing has not been identified yet. Our hypothesis, however, is that by mixing all of the data from different subjects (*i.e.*, by packing), a trained model could suffer from over-generalization or under-generalization due to inconsistencies between differently behaving subjects. In comparison, voting would have more reliable sub-models that are individually trained with consistent data. This hypothesis implies that voting will not have an advantage over packing if inconsistencies among subjects are negligible. Another possibly related aspect is the fact that a simple ensemble of weak predictors (*i.e.*, voting) can mitigate the over-confidence of predictors [3].

Unfortunately, ET is not a magic bullet. However, in some cases, it would work as expected. Although so far we do not know how to identify those cases in advance, we can pose a hypothesis, *i.e.*, divergence or inconsistencies among subjects as discussed above regarding voting. In our experiments, tracking performed considerably better only with the MPR dataset, and voting was also much better than packing in comparison to the MIE dataset result. This suggests that the divergence in subjects of MPR may be larger than in MIE. With regard to MIE, the utterance (text) features are rather dominant for the task [16]. Such textual features would be more consistent and less divergent than other non-verbal features between subjects. This could be a reason that voting and packing show almost the same performances for MIE.

Finally, stacking may work better than tracking in longer interactions. However, stacking performs generally worse in the early stages than tracking and is vulnerable to noise in feedback. Therefore, it would be impractical to choose stacking for online user adaptation if one is focusing on improving performance in early stages.

8 Conclusion

When a machine interacts with a new user, it must correctly recognize the user's behavior to facilitate interaction. This paper looked into the feasibility of expert tracking for quickly adapting to a new user by using multiple models trained on different known subjects. To our knowledge, this is the first work dedicated to that direction.

Our results suggest that, especially for a noisy or divergent situation, separating the data by individuals or groups to form experts and using expert tracking or voting can be better than using a single bulk model (packing). A lesson we learned from this study is that one must be careful in choosing an approach to adapting to different users on the basis of the nature of incoming data. If enough information about a domain is not available, it would be desirable to reserve voting as the default choice because, unlike tracking or stacking, it does not rely on constant feedback, and it would give performances that are acceptable or better than packing.

There are several opportunities to elaborate on this issue. First, an expert tracker's adaptation could be improved by providing better initial weights based on visible similarities between users like gender and age. This approximation could jump-start the algorithm in its search for the best expert and thus decrease the amount of mistakes committed in the search. Second, the weights given to experts could be used to refine the training set of a bulk model. Concretely, these weights could be used to remove the worst experts rather than to select a small number of best experts. This would possibly improve the performance while maintaining the generality of the bulk model. Finally, we must investigate another user adaptation scheme for transfer learning [4, 20], which could be combined with ET effectively. Last but not least, consideration of other online learning approaches such as [18, 26] is also necessary.

Acknowledgements This work was done in Kyoto University Cooperative Intelligence Laboratory funded by Honda Research Institute Japan Co., Ltd. We gratefully thank to Haruto Nishimoto and Kazunori Komatani for the MIE data.

References

1. Araki M, Tomimasu S, Nakano M, Komatani K, Okada S, Fujie S, Sugiyama H (2018) Collection of multimodal dialog data and analysis of the result of annotation of users' interests. In: Proceedings of language resources and evaluation conference (LREC), pp 1584–1588

2. Bousquet O, Warmuth MK (2002) Tracking a small set of experts by mixing past posteriors. J Mach Learn Res 3(Nov):363–396
3. Criminisi A, Shotton J, Konukoglu E (2011) Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Tech. rep, Microsoft Research
4. Daumé III H (2007) Frustratingly easy domain adaptation. In: Proceedings of the 45th annual meeting of the association of computational linguistics, pp 256–263
5. Fiscus JG (1999) A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). In: Proceedings of IEEE workshop on automatic speech recognition and understanding
6. Funakoshi K (2018) A multimodal multiparty human-robot dialogue corpus for real world interaction. In: LREC 2018 special speech sessions, pp 35–39
7. Henderson M, Thomson B, Williams JD (2014) The second dialog state tracking challenge. In: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), pp 263–272
8. Hirano Y, Okada S, Nishimoto H, Komatani K (2019) Multitask prediction of exchange-level annotations for multimodal dialogue systems. In: Proceedings of 2019 international conference on multimodal interaction (ICMI), pp 85–94
9. Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6(5):429–449
10. Jun KS, Orabona F, Wright S, Willett R (2017) Improved strongly adaptive online learning using coin betting. In: Proceedings of the 20th international conference on artificial intelligence and statistics, pp 943–951
11. Komatani K, Okada S, Nishimoto H, Araki M, Nakano M (2019) Multimodal dialogue data collection and analysis of annotation disagreement. In: Proceedings of international workshop on spoken dialogue systems (IWSDS)
12. Lee A, Nakamura K, Nisimura R, Saruwatari H, Shikano K (2004) Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In: Proceedings of Interspeech, pp 173–176
13. Littlestone N, Warmuth MK (1994) The weighted majority algorithm. Inf Comput 108(2):212–261
14. Malmasi S, Dras M (2018) Native language identification with classifier stacking and ensembles. Comput Linguist 44:403–446
15. Nakano Y, Baba N, Huang HH, Hayashi Y (2013) Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In: Proceedings of ACM international conference on multimodal interaction (ICMI), pp 35–42
16. Nishimoto H, Takeda R, Komatani K (2018) Predicting user's interest level in dialogues with multimodal features. In: Proceedings of the 32nd annual conference of the japanese society for artificial intelligence, vol 3C2-OS-14b-04. (in Japanese)
17. Pentland A (2007) Social signal processing. IEEE Signal Process Mag 24(4):108–111
18. Saffari A, Leistner C, Santner J, Godec M, Bischof H (2009) On-line random forests. In: Proceedings of 3rd IEEE ICCV workshop on on-line computer vision (2009)
19. Schuller B, Steid S, Batline A (2009) The interspeech 2009 emotion challenge. In: Proceedings of 10th annual conference of the international speech communication association (INTERSPEECH), pp 312–315
20. Segev N, Harel M, Mannor S, Crammer K, El-Yaniv R (2017) Learn on source, refine on target: a model transfer learning framework with random forests. IEEE Trans Pattern Anal Mach Intell 39(9):1811–1823
21. Sugiyama T, Funakoshi K, Nakano M, Komatani K (2015) Estimating response obligation in multi-party human-robot dialogues. In: Proceedings of 2015 IEEE-RAS 15th international conference on humanoid robots (Humanoids), pp 166–172
22. Sztyler T, Stuckenschmidt H (2017) Online personalization of cross-subjects based activity recognition models on wearable devices. In: Proceedings of IEEE international conference on pervasive computing and communications (PerCom), pp 180–189

23. Vovk V (1990) Aggregating strategies. In: Proceedings of the 3rd annual workshop on computational learning theory, pp 371—383
24. Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. J Theor Biol 242(4):941–946
25. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259
26. Zhao P, Hoi SC, Wang J, Li B (2014) Online transfer learning. Artif Intell 216:76–102

Language Identification, Grammar and Syntax

On the Use of Phonotactic Vector Representations with FastText for Language Identification

David Romero and Christian Salamea

Abstract This paper explores a better way to learn word vector representations for language identification (LID). We have focused on a phonotactic approach using phoneme sequences in order to make phonotactic units (phone-grams) to incorporate context information. In order to take into consideration the morphology of phone-grams, we have considered the use of sub-word information (lower-order n-grams) to learn phone-grams embeddings using FastText. These embeddings are used as input to an i-Vector framework to train a multiclass logistic classifier. Our approach has been compared with a LID system that uses phone-gram embeddings learned through Skipgram that do not implement sub-word information, using Cavg as a metric for our experiments. Our approach to LID to incorporate sub-word information in phone-grams embeddings significantly improves the results obtained by using embeddings that are learned ignoring the structure of phone-grams. Furthermore, we have shown that our system provides complementary information to an acoustic system, improving it through the fusion of both systems.

1 Introduction

Learning word embeddings plays an important role in Natural Language Processing (NLP) tasks, being a core component for several applications, such as Speech Recognition [14], Machine Translation [15], Image description [7] and Machine reading [20]. While word representations have been successfully used in NLP, this approach

D. Romero (✉) · C. Salamea
Interaction, Robotics and Automation Research Group, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador
e-mail: davidromerom.19@outlook.es

C. Salamea
e-mail: csalamea@ups.edu.ec

C. Salamea
Speech Technology Group, Information and Telecommunication Center, Universidad Politécnica de Madrid, Av-Complutense 30, 28040 Madrid, Spain

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_25

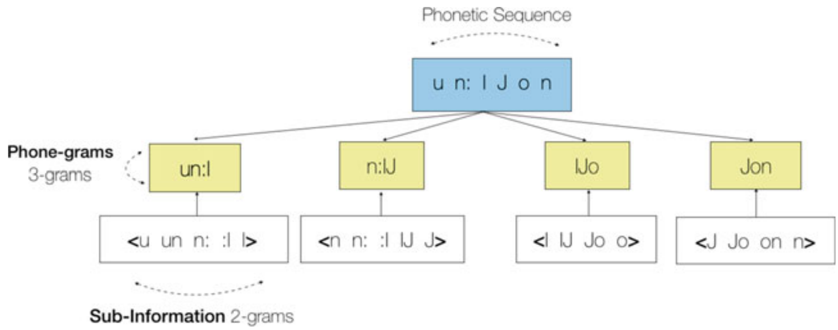


Fig. 1 Learning phone-grams representations (3-grams) using sub-information (2-grams) with FastText

has also been applied in other fields such as Automatic Spoken Language Identification (LID) [3, 6, 10, 17], unlike many NLP tasks where these embeddings are usually learned at a word level using semantic and syntactic features, in LID, these embeddings are learned at a phonetic/phonotactic level, where these features are unusual, making this task more challenging.

LID systems play an important role in dialogue systems applications such as online translation, call center routing and data labelling. In these systems the best results have been achieved using acoustic-based systems, however, their fusion with phonetic/phonotactic-based systems give a boost of performance [17].

In this work, we strive to make better representations at a phonotactic level for the LID task. The primary challenge towards this goal is to overcome the limitations that phonemes do not have semantic and syntactic relationships, due to this, the use of Phone-Grams has been considered. These new units concatenate two or more phonemes to incorporate context information (see Fig. 1), being this the same approach used in [17]. The second challenge is that, at a word level, learning a distinct representation for each word, ignoring its internal structure is a limitation for morphologically rich languages, which leads to having very bad representations of words that occur rarely in the training corpus [15]. While at the phonotactic level we will have more combinations than at the word level, we could have a lot of phone-grams that occur rarely in some languages, which would result to have units that have poor representations or does not have representation at all. Our core insight is that we could represent these phonotactic-units (phone-grams) in a more efficient way using Fast-Text. Due to many phone-grams formations could follow rules, learning its representations using sub-information (lower-order n-grams) could lead to better representations, and therefore to better results in the LID task. As shown in Fig. 1, we use 2-grams as sub-information to learn the representation of 3-grams. Our primary contribution is incorporating sub-word information in phone-gram embeddings in order to take into account the morphology of these units in the LID task. This paper is organized as follows. In Sect. 2 we review related works in LID using sub-word information. Section 3 presents our approach and explains how we improve LID performance. Finally, Sects. 4 and 5 presents our experiments and results.

2 Related Word

Modern speech systems handle large words vocabularies making it infeasible to collect enough repetitions of each word and train good models, this has motivated the use of sub-word units to build vector representations of words [9]. It has been shown that subword ngrams models have been used to improve word representations for many tasks [9, 13, 18], being useful in challenging applications such as native language identification [8] and language discrimination [2], where the approach of bag-of-n-grams has been used to discriminate languages in a collection of documents. Also, tasks such as code-switching [12, 19] where the authors extended the LID task to the sub-word level to make language identification in conversations where people alternate between two or more languages. Subword information also can help to improve generalization in low resourced languages [5], transferring word embeddings from high resourced languages using sub-word level information.

In this work, we want to take advantage of the usefulness of sub-word information in LID tasks, using its advantages in generalization and flexibility to represent phone-grams taking into account the morphology of these units. All these works have used subword units at a word level, however, to our knowledge, the use of sub-word units (lower-order n-grams) to make phone-grams embeddings has not been done yet.

3 Model

In this section, we will describe our approach to perform language identification. We will begin by presenting the database and the general model, then the followed approach to learn Phone-Grams embeddings using sub-word information with n-grams and how we managed these set of embeddings, finally a brief description of the acoustic system used to perform the fusion with the phonotactic system will be exposed.

3.1 Database

We evaluate our LID system performance on the Kalaka-3 database [16]. This database contains clean and noisy audio recordings in 6 different languages in the closed set condition (Basque, Catalan, English, Galician, Portuguese and Spanish), including 108 h of speech in total, all of them stored as a single channel 16 kHz.

This database has 4656, 458 and 941 audio recordings distributed for training, development, and evaluation respectively.

3.2 General Model

Our system (see Fig. 2) consists of two parts. The first one is called “Front End”. In this part, the acoustic signal is processed, and for each language in the corpus the phonemes sequences are generated by a phoneme recognizer. For this part, we have used a phoneme recognizer designed by Brno University [1] that has 3 sets of HMMs, for Hungarian, Russian and Czech. In the second phase that is called “Back End”, we create Phone-Grams through the concatenation of phonemes, then, we take the Phone-Grams in the training corpus to train their embeddings using FastText, (explained in Sect. 3.3). After that, all the Phone-Grams in the corpus have been replaced by their corresponding embeddings that were learned previously. Because we will have multiple FastText models one for each language, we have used the “Multiple Vector Embedding” approach (explained in Sect. 3.4). These Phone-Grams embeddings are used as input to an i-Vector System; thus, we have used the Phone-Grams embeddings to train the T-Matrix and the UBM model needed to obtain the i-Vectors. Finally, we have used these i-Vectors as features to train a multiclass logistic classifier (MLR).

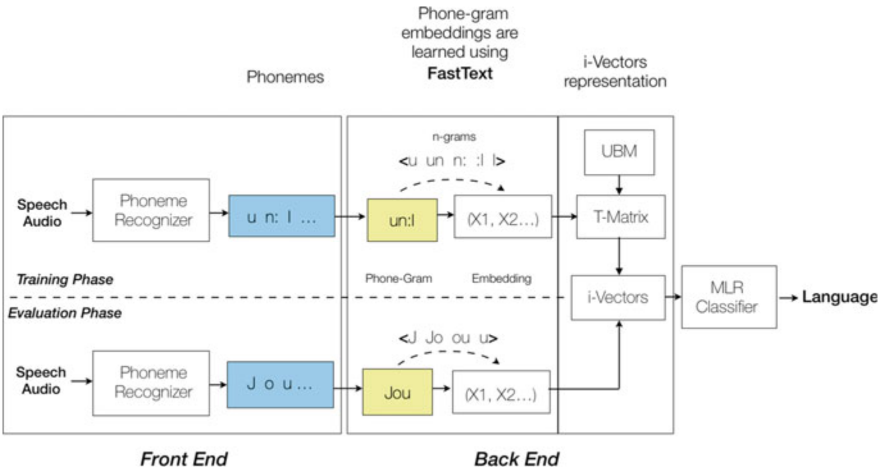


Fig. 2 Global system architecture

3.3 Phone-Grams Embeddings

In our approach we learn phone-grams embeddings using character n-grams with FastText [4]. In this way, to take into consideration the internal structure of each unit each Phone-Gram will be represented as a bag of character *n-grams*. FastText include special symbols < and > to distinguish prefixes and suffixes in each unit, also, include the word itself in the set of its n-grams, that in our case would be the Phone-Gram (see Fig. 2). So, given a Phone-Gram p the set of n-grams that appear in p would be $G_p = \{1 \ldots G\}$, being G_p a dictionary of n-grams of size G. Then, a vector is associated to each n-gram g, with V_g and the vector of the context Phone-Gram c, with V_c. FastText uses the following equation to represent the scoring function of the actual word and its context, that in our case using phone-grams would be defined as follows

$$s(p, c) = \sum_{g \in G_p} V_g^T V_c \tag{1}$$

So, we represent a Phone-Gram by the sum of the vector representations of its n-grams, which will allow us to share representations across Phone-Grams to learn representations of rare units in the corpus.

3.4 Multiple Vector Embedding (MVE)

The approach that we have used to manage the set of embeddings is the MVE, which has been proposed in [17]. Because we have multiple FastText Models L_m each one trained for each language, we have used all of them to obtain i-Vectors for each model and then we have fused the scores S_L provided by each individual language model in the MLR in order to complement the information of each system and get a better performance in the final LID system (final phonotactic score fps). Specifically:

$$S_{Lx} = \text{MLR}\ (L_{mx}) \tag{2}$$

$$fps\ (finalphonotacticscore) = Fusion(S_{L1}, S_{L2} \ldots\ldots, S_{L6}) \tag{3}$$

3.5 Acoustic System

Once we have fused the scores of each language and obtained the final score of the phonotactic system, we have fused it with the score obtained by an acoustic system, this was done to complement the information of both systems. The acoustic system has been generated as follows: from each speech utterance, 12 MFCC coefficients are extracted for each frame. The silence and noisy segments of the acoustic signal have been removed using a Voice Activity Detector. Also, a RASTA filter has been used along with cepstral mean and variance normalization to reduce noise perturbations. We have a feature vector of dimension 56, generated from the concatenation of SDC parameters using a 7-1-3-7 configuration. These feature vectors are used to train the total variability matrix, and the UBM with 512 Gaussians in order to extract the i-Vectors with dimension of 400 [17].

4 Experiments

4.1 Baseline

In our experiments, we compare our approach to LID using FastText to obtain Phone-Grams embeddings using sub-word information to the use of Skipgram that learns embeddings ignoring the internal structure of Phone-Grams. We make this comparison using C_{avg} as the metric for our experiments, which consider the number of false acceptances and false alarms generated by the recognizer [11].

4.2 Model Setup

For both our approach using FastText and the baseline experiments with Skipgram, we use the following parameters: The phoneme sequences used to make the phone-grams are the sequences generated by the phoneme recognizer that has 3 sets of HMMs, in this work we use the Hung phoneme recognizer [1], because this HMM has given us a better performance in previous experimentations [17]. For our experiments, we considered two Phone-Gram sizes, 3-grams (3-phone-gram) and 2-grams (2-phone-gram). We choose these two sizes because although large order Phone-Grams could have more context information, these units produce a large dispersion of information due to the increase in phonemes combinations. The Phone-Grams embeddings have dimension 10, a context window size of 5, and a negative sampling of 25. For our approach better results were obtained using small embeddings instead of using larger sizes. We train these embeddings with 3 interactions (epochs) and a learning rate of 0.009. Also, we have used a min-count of appearances of 5, which is a default value that works well for our approach. In the i-Vector system, the best results were obtained training the UBM with 512 Gaussians using 5 interactions, the same amount of interactions was used to train the Matrix-T. All these parameters have the optimal values for our approach, which have been established during the experimentation phase.

5 Results

In this section, we show our experiments and the comparison with the baseline system. After that, we fusion our model with an acoustic LID system in order to complement the information of both systems.

5.1 Phone-Gram Embeddings with Sub-word Information-FastText

As we described in Sect. 4.2, we have used 2 sizes of phone-grams for our experiments (3-phone-gram and 2-phone-gram). To use sub-word information in these phone-grams we have performed tests with different lengths of lower-order n-grams, finding the best performance using 2-grams as sub-word information for 3-phone-grams and 1-gram for 2-phone-grams.

5.2 Phone-Gram Embeddings Without Sub-word Information-Skipgram

The phone-grams embeddings obtained with Skipgram does not have sub-word information. So, with this approach, the phone-grams that were not found in the training phase have been replaced by a vector with low values of 0.0001, which is different than our approach using FastText where these phone-grams embeddings are formed using their sub-word information.

5.3 LID—Results: Comparing Embeddings

In this section, we summarize the results of our experiments. Table 1 contains the results for all languages using phone-gram embeddings obtained using sub-word information with FastText, compared to the LID system that uses Skipgram embeddings. With our approach using sub-word information in phone-grams we have improved the results in each language and in the fusion of all the scores. We have obtained a relative improvement of 36.2% in 3-phone-grams and 15.3% in 2-phone-grams. The use of sub-word information improved the performance of the LID task, although we have used very small n-grams as sub-word information; these units were capable of giving useful information to the phone-grams embeddings. Also, with FastText we were capable to have embeddings for rare phone-grams that are not in the training corpus, removing in this way the approach of using low vector values for these units.

Table 1 Summary of results

	C_{avg}			
	3-Phone-gram		2-Phone-gram	
Languages	FastText	Skipgram	FastText	Skipgram
Basque	25.72	35.52	27.61	35.98
Catalan	27.92	34.22	27.37	35.86
English	27.48	40.98	26.29	36.81
Galician	25.68	33.95	28.51	35.74
Portuguese	26.19	37.62	24.92	37.72
Spanish	27.75	32.34	28.70	33.49
Fusion	**20.06**	31.46	**22.59**	26.68

Table 2 Fusion with the acoustic system

	System	C_{avg}	Improvement %
	Acoustic system	7.60	–
3-phone-gram	Fusion with FT-Emb	**5.71**	**24.8**
	Fusion with SK-Emb	6.42	15.5
2-phone-gram	Fusion with FT-Emb	**6.0**	**21.1**
	Fusion with SK-Emb	6.27	17.1

5.4 Fusion with the Acoustic Model

In order to complement the information of the phonotactic system that uses FastText, we have fused it with an acoustic system to improve the LID performance. Table 2 contains the results of the fusion of both systems. The results show that the fusion of both systems improves the acoustic system performance. The performance is improved in both cases, using 3-phone-grams and 2-phone-grams, obtaining a bigger improvement in the first case using higher-order phone-grams, getting our best performance with a Cavg of 5.71.

6 Conclusions

In this paper, we focus on learning phone-gram representations by using sub-word information with FastText. The proposed approach has been evaluated on 6 languages for the LID task and we showed that despite the sub-word information use low order n-grams that carry little context information these units were capable of improving the phone-grams representations and the performance of the LID system. Our approach outperforms the baseline LID system which uses Skipgram to learn embeddings without sub-word information, our best performance has been obtained with 3-phone-grams, having a relative improvement of 36.2% over the baseline. The phonotactic system performance was improved by the fusion with the acoustic system with an improvement of 24.8% over the acoustic system alone. In this way, we have shown that the use of sub-word information in phonotactic units give a boost of performance in LID systems, improving the representations of the common and rare units in the corpus.

As future work we aspire to incorporate probabilities to each phone-gram with the aim of having an initial notion of the possible language in order to decide which vectors to use to represent each unit, this could lead to a better and faster performance which is an important factor in dialogue systems applications.

References

1. Ace P, Schwarz P, Ace V (2009) Phoneme recognition based on long temporal context
2. Barbaresi A (2017) Discriminating between similar languages using weighted subword features. In: Fourth workshop on NLP for similar languages, pp 184–189
3. Berkling K, Arai T, Barnard E (1994) Analysis of phoneme-based features for language identification. In: Proceedings of the international conference on acoustics, speech and signal processing. IEEE, pp 289–292
4. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. arXiv:1607.04606v2
5. Chaudhary A, Zhou C, Levin L, Neubig G, Mortensen D, Carbonell J (2018) Adapting word embeddings to new languages with morphological and phonological subword representations. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 3285–3295
6. D'Haro L, Glembek O, Plchot O, Matejka P, Soufifar M, Córdoba R, Cernocky J (2012) Phonotactic language recognition using i-vectors and phoneme posteriogram counts. In: ISCA 13th annual conference, Proceedings of the INTERSPEECH, pp 42–45
7. Karpathy A, Fei-Fei L (2017) Deep visual-semantic alignments for generating image description. IEEE Trans Pattern Anal Mach Intell 39(4):664–676
8. Kulmizev A, Blankers B, Bjerva J, Nissim M, Noord G, Plank B, Wieling M (2017) The power of character n-grams in native language identification. In: Proceedings of the 12th workshop on innovative use of NLP for building educational applications, pp 382–389
9. Livescu K, Fosler-Lussier E, Metze F (2012) Sub-word modeling for automatic speech recognition. IEEE Signal Process Mag 29:44–57
10. Matejka P, Schwarz P, Cernock J, Chytil P (2005) Phonotactic language identification using high quality phonome recognition. In: Proceedings of the IberSPEECH, pp 2237–2240
11. Martin A, Greenberg C (2010) The 2009 NIST language recognition evaluation. In: Odyssey, p 30
12. Mager M, Cetinoglu O, Kann K (2019) Subword-level language identification for intra-word code-switching. arXiv:1904.01989v1
13. Mikolov T, Sutskever I, Deoras A, Le H, Kombrink S, Cernocky J (2011) Subword language modeling with neural networks
14. Palaskar S, Raunak V, Metze F (2019) Learned in speech recognition: contextual acoustic word embeddings. arXiv:1902.06833v1

15. Qi Y, Sachan D, Felix M, Padmanabhan S, Neubig G (2018) When and why are pre-trained word embeddings useful for neural machine translation? arXiv:1804.06323v2
16. Rodriguez L, Penagarikano M, Varona A, Diez M, Bordel G (2016) KALAKA-3: a database for the assessment of spoken language recognition technology on YouTube audios. Lang Resour Eval 50(2):221–243
17. Salamea C, Córdoba R, D'Haro L, Segundo R, Ferreiros J (2018) On the use of phone-based embeddings for language recognition. In: Proceedings of the IberSPEECH, pp 55–59
18. Singh R, Raj B, Stern R (2002) Automatic generation of subword units for speech recognition systems. IEEE Trans Speech Audio Process 10(2):89–99
19. Xia M (2016) Codeswitching language identification using subword information enriched word vectors. In: Proceedings of the second workshop on computational approaches to code switching, pp 132–136
20. Zhang Z, Huang Y, Zhu P, Zhao H (2018) Effective character-augmented word embedding for machine reading comprehension. arXiv:1808.02772v1

The Influence of Syntax on the Perception of In-Vehicle Prompts and Driving Performance

Daniela Stier, Ulrich Heid, Patricia Kittel, Maria Schmidt, and Wolfgang Minker

Abstract Advances in Natural Language Generation technically enable dialog systems to output utterances of an arbitrary length. However, in order to provide the most efficient form of interaction, the complexity of voice output needs to be adapted to individual user needs and contexts. This paper investigates the influence of syntactic complexity on user experience and primary task performance with spoken interaction representing a secondary task, such as in the automotive context. For this purpose, we validate the approach of assessing user preferences concerning voice output. On this basis, we report the results of a user study, where participants interact with a simulated dialog system producing utterances of differing syntactic complexity. We conclude that the choice of a particular syntactic structure affects primary task performance. Equally, our analyses of user preferences suggest an influence on the perception of syntactic forms dependent on individual context and user characteristics.

1 Introduction

From a technical perspective, voice output of an arbitrary length and complexity can be produced. However, it is generally agreed that more intelligent software is required for Spoken Dialog Systems (SDSs) to enable complex Human-Machine-

D. Stier (✉) · P. Kittel · M. Schmidt
Mercedes-Benz AG, Sindelfingen, Germany
e-mail: daniela.stier@daimler.com

P. Kittel
e-mail: patricia.kittel@daimler.com

M. Schmidt
e-mail: maria.m.schmidt@daimler.com

U. Heid
Hildesheim University, Hildesheim, Germany
e-mail: heidul@uni-hildesheim.de

D. Stier · W. Minker
Ulm University, Ulm, Germany
e-mail: wolfgang.minker@uni-ulm.de

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_26

Interaction (HMI; [11]). The aim to provide the most efficient form of interaction becomes of particular interest in situations, where the SDS interaction is deprioritized to the secondary task, such as in the automotive context. Here, the requirement to support and not distract an individual driver from the primary task of driving results in the need of efficient and intuitive user interfaces. In this context, Natural Language Generation (NLG) capable of flexibly adapting voice output to a particular user or situation has gained increased attention.

In this paper, we focus on natural and intuitively comprehensible SDS voice output with regards to its syntactic form. We assume that structural complexity directly relates to an increased workload while driving a car [6]. We further hypothesize that in a driving environment the appropriateness of a particular syntactic structure and its inherent complexity depends to a large extent on both the individual characteristics of the driver and the respective driving situation. As such, a linguistically affine user might be used to process complex syntactic structures and as a consequence will be less distracted than a user with low linguistic knowledge. An adaptive in-vehicle SDS should include such individual factors in its NLG strategy when choosing a syntactic form. Focusing on user experience, we therefore consider it necessary to investigate the influence of syntactic forms in in-vehicle voice output on user- and application-side parameters.

In our investigations, we adopt a user-centered approach and rely on user preferences. For this purpose, we (1) investigate whether the subjective assessment of user preferences concerning SDS output (*i.e.* via audio) allows valid conclusions about syntax. In a pilot user study, we analyze whether participants are capable of explicitly identifying differences between varying syntactic realizations. We hypothesize that user preferences are influenced by an awareness of syntax due to established opinions concerning its appropriateness. We prove whether this (lack of) awareness permits intuitive, unbiased assessment results. On this basis, we (2) investigate the relevance of syntactic forms in relation with different user characteristics and application contexts. We report the results of a user study in a driving simulator, where participants were asked to rate syntactic voice output paraphrases with regard to their perceived *naturalness* and *comprehensibility*. We furthermore (3) analyze the influence of syntactic structures on driving performance under the assumption that an increasing syntactic complexity is directly related with an increase in cognitive load derivable by means of objective performance measures.

2 Related Research

Research is done on all levels of the NLG pipeline [22, 24], such as adaptive information presentation (*e.g.* [3, 25]; as *content planning*) or lexical and structural adaptation (*e.g.* [4]; as *sentence planning*). Focusing on a linguistic level, the terms ‘alignment’ and ‘priming’ in human-human communication [20] were transferred to HMI by Branigan et al. [2], who indicated that human-computer alignment is even stronger than between human interlocutors to avoid communicative failure. As a consequence, alignment by computers with human users is assumed a promising approach and employed as a general strategy in literature. As an example, Hu et al. [9], Mairesse and Walker [16, 17] demonstrate how linguistic styles between a fine-grained distinction of personality traits can be learned and generated. However, from a user perspective the question of applying an NLG strategy following either a *similarity* [18, 19, 32] or *complementarity* [10, 14] principle remains unsolved. Furthermore, the exclusiveness of this binary distinction is questionable, since both concepts may be found in human-human discourse [7]. This confusion might be resolved in relation to the *context* of interaction [1].

Another outstanding point in literature is how an alignment strategy should be adjusted in SDS interaction as a secondary task, such as in the automotive context. Here the terms ‘safety’ and ‘cognitive load’ represent key concepts, and the primary driving task is prioritized over any system interaction or communicative success. A rich literature exists discussing the effect of (human-human) language use while driving, *e.g.* speaking on a telephone [12] or with an in-vehicle passenger [30, 33]. With regards to voice output, linguistic complexity *e.g.* induced by means of ambiguous relative clauses was found to influence cognitive load in a dual-task environment [6, 8]. However, concrete implications for adaptive in-car SDS voice output considering individual user characteristics and application contexts are still to be defined. Although Stier and Sigloch [28] investigated the influence of syntactic structures on personal- and system-side parameters in voice output while driving, the authors indicate several weaknesses in their data elicitation approach. Their results thus should be interpreted with reservations.

In comparison to the mainly theory-oriented approaches in previous works to implement and prove alignment theories, we focus on a user perspective. We revise the work by Stier and Sigloch [28] in order to identify the relevance of syntax and particular parameters, which should be taken into account in the adaptation of SDS voice output in dual-task interaction contexts. In future work, this research will contribute to an NLG strategy by matching the influence of syntax in voice output with the use of syntax in natural human, in-vehicle linguistic behaviour. In this way, the confusion between the *similarity* and *complementary* principles will be resolved with a focus on the user and the interaction situation.

3 Controlling Paraphrases

In order to draw valid conclusions about syntax, we generated textual paraphrases which are comparable in terms of content and information density. For our purpose, we limited the scale of dialog turns to controllable *one-shot Question-Answer* sequences (QAS; *cf.* [28]). The length of answers was determined as two main clauses. Possible QAS were further limited to three types of explanations (*cf.* [26, 27]; Table 1). We chose two domains and a subset of their functions from a Mercedes Benz S-Class driver’s manual as contents from the vehicle and driving context, *i.e.* driving assistants (*DAS*) and comfort functions (*COF*). After matching the contents

Table 1 Explanation types (*What, How, When*), Domains (*DAS, COF*) and functions (F)

What	The answer to the question "What is function F?" (ger. *Was ist Funktion F?*) supplies a general definition of a particular function F
How	"How does F work?" (ger. *Wie funktioniert F?*) requires an explanation of F's functionality
When	The additional question "When can I use F?" (engl. *Wann ist F einsetzbar?*) in our work is a special case of *How* asking for particular limitations of function F
Driving Assistants (DAS)	Blind Spot Assist (*Totwinkel-Assistent*), Space Assist (*Abstands-Assistent*), Lane Keeping Assist (*Spurhalte-Assistent*), Emergency Stop Assist (*Nothalt-Assistent*)
Comfort Functions (COF)	Well being (*Behaglichkeit*), Joy (*Vergügen*), Vitality (*Vitalität*), Warmth (*Wärme*)

to one of the explanation types, we computed several surface measures[1] on the resulting texts to ensure a comparable lexical and structural complexity. In a last step, we subordinated the produced main clause variants (*MCV*) to paraphrases containing subject-oriented relative clauses (*RCV*). Further syntactic realizations were omitted (*cf.* [28]); so only textual bases for *MCV* and *RCV* were constructed as two extremes on a continuum of syntactic complexity. While *MCV* is represented by simple linear structures, the nested syntactic structure of *RCV* is assumed to be more difficult to process [34]. This difference in syntactic complexity is exemplarily demonstrated for the Lane Keeping Assist by means of dependency parses in Fig. 1, where the longest path [21] of *MCV* consists of five nodes in contrast to seven for *RCV*.

This difference in complexity between our syntactic paraphrases represents the basis for our dual-task analyses. We assume that it has a significant impact on human cognitive processing and perception, which can be measured by means of linguistic preferences and objective performance measures.

4 Validating the Data Elicitation Method

User preferences concerning syntactic structures are commonly assessed via textual samples (*e.g.* [17]). However, we consider the audio channel as the decisive factor in the evaluation of voice output. For this reason, we conducted a pilot study to validate whether audio can be used to assess valid user preferences regarding syntax. In this context we investigated whether our participants were able to identify and distinguish different syntactic forms in voice output and whether their (lack of) awareness of syntax allowed intuitive user ratings.

[1] Number of words, av. sentence length, prop. of words > 6 characters, LIX index, Idea Density [5].

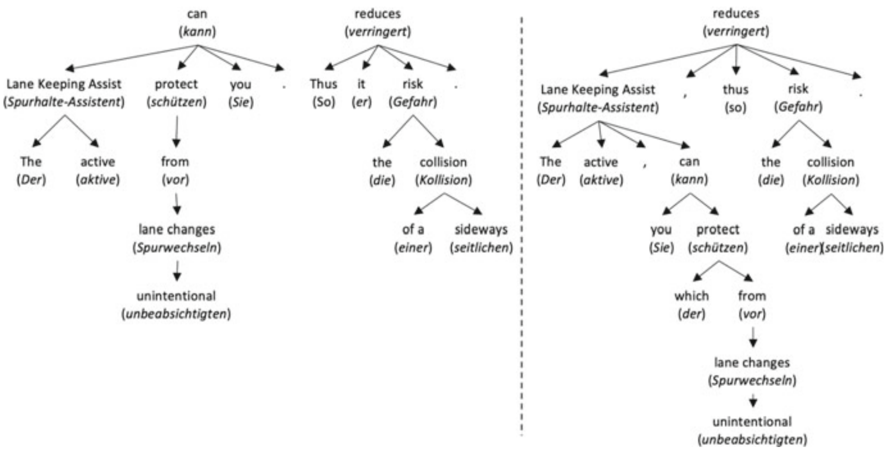


Fig. 1 Visualization of the dependency tree depths for the main clause variant (*MVC*; left) and relative clause variant (*RCV*; right) of *What*, demonstrated for the Lane Keeping Assist

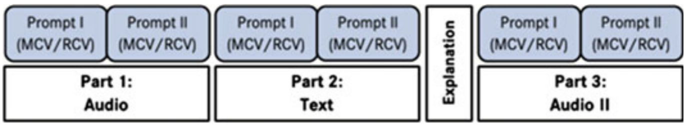


Fig. 2 Schematic procedure of pilot user study

This pilot study took place without a parallel secondary task in order to focus the participants' attention on syntactic differences.

Pilot Study Design and Participants. A total of 77 German native speakers (male 60%, female 40%) between 18 and 69 years (*mean* 43.60, *sd* 14.65) were invited and asked to follow the instructions by the investigator without any further preparation. As visualized in Fig. 2, participants were first asked to listen to two synthesized voice prompts (one *MCV* and one *RCV* in random order; Part 1). In Part 2, the instructor presented the identical prompts in text form to the participants and afterwards revealed the differences between the two syntactic variants (Explanation). The participants were then asked to listen to the same synthesized voice prompts again (Part 3). In each step, the participants were asked whether they noticed any peculiarity in the voice/text prompts and which of the two variants they preferred.[2]

Pilot Study Results. Firstly, only 12 subjects (15.6%) explicitly identified the syntactic differences between our voice prompts via audio. In contrast, a clear majority perceived the syntactic differences through text in Part 2 (40 participants [51.9%]). On a conscious level, the syntactic differences were therefore perceived significantly

[2]The chosen procedure was not randomized in order to investigate the perception of syntactic forms and how user preferences change depending on the awareness of syntactic structures.

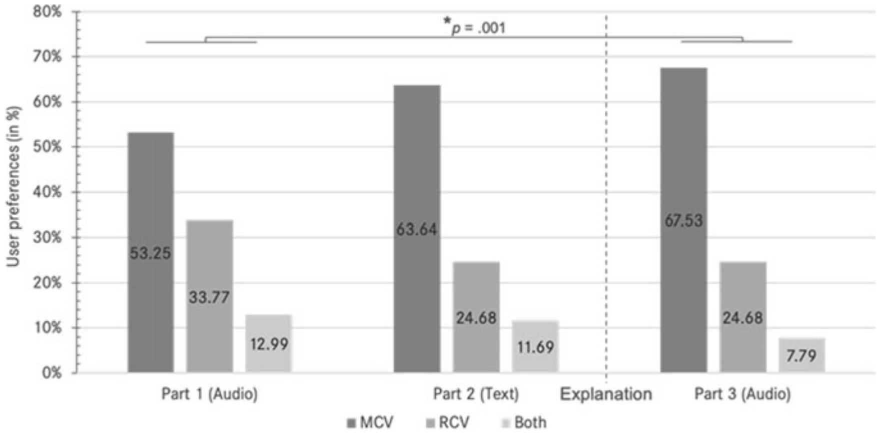


Fig. 3 Indicated user preferences (in %)

less as a distinguishing feature via the audio channel than in the text (Wilcoxon signed-ranks test: $Z = -5.29$, $p < .001$, $r = .60$).

Secondly, we observed that the assessed user preferences (Fig. 3) differed significantly between Parts 1 and 3, *i.e. before* and *after* the explanation phase ($Z = -3.19$, $p < .001$, $r = .26$).[3] Subjects, who were aware of syntactic differences *after* the explanation phase, were more likely to prefer *MCV*. In a post-interview, they stated that relative clauses were too complex and therefore rather unsuitable in voice output.

These observations generally confirm our hypothesis that the awareness of syntax does influence user preferences. In this pilot study, we have seen that the perception of syntactic differences over audio was lacking. But precisely this lack of awareness seems to allow for intuitive user ratings (without a fixed opinion on syntactic complexity and applicability) concerning the preferred syntactic variant. We therefore conclude that the assessment of voice output via audio concerning individual syntactic preferences represents a valid methodology.

5 The Role of Syntax in a Dual-Task Environment

Based on this pilot study, we conducted a Wizard-of-Oz (WoZ) experiment in a driving simulator and revised the weaknesses outlined by Stier and Sigloch [28].

Experimental Setup. The study took place in a fixed-base simulator (Mercedes C-Class) with a 180-degree screen. Real-time driving data (RTDD) was assessed via a Controller Area Network (CAN bus). The vehicle bus was additionally connected to the WoZ-Tool to synchronize assessed user ratings with user driving performance.

[3] No significant difference between Parts 1 and 2 was revealed ($Z = -1.95$, $p = .051$, $r = .16$).

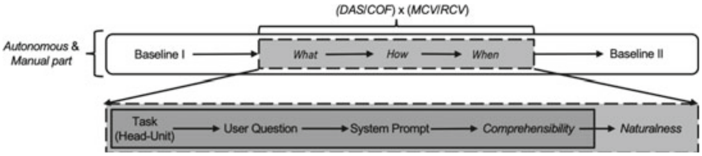


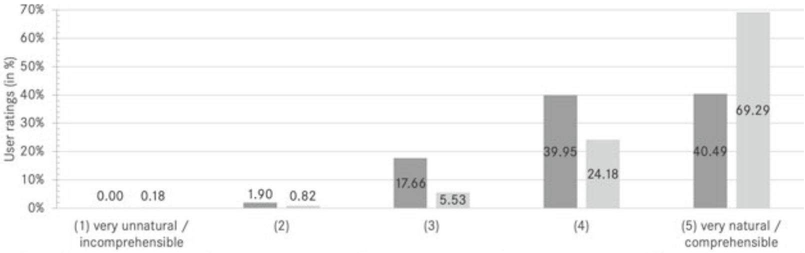
Fig. 4 Schematic procedure of user study

A speaker was installed in the front passenger footwell, not visible to the driver, to output dialog turns directed by the WoZ. We chose a freeway with moderate traffic as driving environment. The driving task was split into one manual (*MAN*; SAE level 0, *i.e.* without automated system) and one autonomous part using a prerecorded video (*AUT*, SAE level 5; each part 15 min). The order of driving parts was randomized. There was a short break of approx. 3 min for the participants waiting outside the car to reload the simulation between the driving parts.

Study Design. Participants were instructed to drive at a speed of 100 km/h and to follow a lead vehicle at a distance of approximately 100 m (*i.e.* two delineator posts). Baselines were included at the beginning and end of the drive to gather performance data without SDS interaction as a secondary task (Fig. 4). Spoken interaction between a participant and our WoZ was based on consecutive QAS for *What*, *How* and *When*. In each step user-initiative was triggered by displaying a task on the head-unit screen, indicating the respective type of question to be formulated and a *COF* or *DAS* function the participant should enquire. A question by the user was followed by a system explanation and the request to assess the voice prompt on a five-point Likert scale concerning its perceived *comprehensibility*. After completing all three QAS, the participant was asked to evaluate the *naturalness* of system responses on a five-point scale. The outlined question-answer dialog sequences were repeated for *AUT* and *MAN*, so all participants experienced each QAS eight times (2 driving complexities x 2 domains x 2 syntactic realizations).

In order to keep up the illusion of real SDS interaction, each participant was instructed to activate the 'voice assistant' (*i.e.* our WoZ) with the phrase "Hello Mercedes" (ger. *Hallo Mercedes*) before stating their question. Furthermore, the subjects were asked to focus on the quality and formulation of voice output in their evaluations and to leave out aspects such as text-to-speech quality.

Comparison with Stier and Sigloch [28]**.** The difference between our approaches mainly consists in the reduction of cognitive load of the originally highly demanding/overloading task (*cf.* [29, 35]). We used a lead vehicle as an orientation. Our subjects were thus relieved of the stress factor to maintain speed on their own responsibility. In addition, we replaced the short QAS, which merely consisted of *What* (or *How*), with consecutive QAS in order to prime subjects over a longer time period with a particular syntactic structure. We also split the question of *comprehensibility* and *naturalness* into two separate steps to provide a clearer understanding of the assessment task. Most importantly, we omitted randomly placed phone calls by the WoZ (subjects should reproduce the last voice prompt in own words) to decrease



		(1)	(2)	(3)	(4)	(5)
Naturalness	*MCV*	0.00	1.09	15.76	39.13	44.02
	RCV	0.00	2.72	19.57	40.76	36.96
Understandability	*MCV*	0.18	0.36	3.80	21.20	74.46
	RCV	0.18	1.27	7.25	27.17	64.13

■ Naturalness (n=368) ■ Understandability (n=1104)

Fig. 5 Indicated user ratings (in %)

cognitive load of participants. The phone calls intended to simulate sincere interest of subjects in the voice output. Although we acknowledge this motivation, it has been shown by Stier and Sigloch [28] that this approach increased the stress level of participants to an extent that their results should be interpreted with reservations.

Participants. A total of 48 German native speakers[4] from four age groups (19–29, 30–44, 45–59, 60–70 years) participated in the experiment with an average of 42 years (*sd* 15.07) and a gender distribution of 27 (59%) male and 19 (41%) female subjects. Further personal information was assessed via questionnaires on five-point Likert scales. As such, experience with *DAS* (*mean* 2.22, *sd* 0.72) and *COF* (*mean* 1.54, *sd* 0.62) was self-assessed as rather low, while an average linguistic knowledge (*mean* 2.70, *sd* 0.51) was indicated. On average, participants considered themselves as poorly open to technology in general (*mean* 1.26, *sd* 0.44; [13]). The self-assessed Big Five traits [23] displayed an even distribution with Openness (*mean* 2.59, *sd* 0.49), Conscientiousness (*mean* 2.93, *sd* 0.25), Extraversion (*mean* 2.76, *sd* 0.43), Agreeableness (*mean* 2.93, *sd* 0.25) and Neuroticism (*mean* 1.91, *sd* 0.46).

5.1 Evaluation of Subjective User Ratings

Our syntactic paraphrases were assessed 368 and 1104 times with regard to their perceived *naturalness* and *comprehensibility*, respectively. Ratings are found within the higher ranges of the five-point scales (Fig. 5). In general, *MCV* was assessed as more *natural* ($Z = -3.14$, $p = .002$, $r = .23$) and better *comprehensible* ($Z = -7.53$, $p < .001$, $r = .32$) than *RCV*. Evaluation was conducted fitting two 2-level generalized linear mixed models (*GENLINMIXED* in SPSS) with subjects introduced as random intercepts to account for the repeated measures character of our data. A

[4]Two participants were excluded from analyses due to technical problems in the driving simulator.

Table 2 User and situation parameters (left) and significant results for the interaction effect with the parameter *Sentence type* (right)

	Parameter	Levels	*Naturalness*	*Comprehensibility*
	Complexity	*AUT*, *MAN*	*n.s.*	*n.s.*
	Domain	*COF*, *DAS*	*n.s.*	F(1,318) = 5.032*
	Question type	*What*, *How*, *When*	–	F(2,318) = 3.466*
	Sentence type	*MCV*, *RCV*	–	–
	Age groups	18-29, 30-44, 45-59, 60-70	*n.s.*	*n.s.*
	Gender	male, female	*n.s.*	*n.s.*
Exp[a]	*Linguistics*	low < mid < high	F(2,318) = 93.287**	F(2,318) = 327.406**
Exp[a]	*COF, DAS*	low < mid < high	*n.s.*	*n.s.*
BFT[b]	*Openness, Extraversion*	mid < high	*n.s.*	*n.s.*
BFT[b]	*Conscientiousness*	mid < high	F(1,318) = 5.093*	*n.s.*
BFT[b]	*Agreeableness*	mid < high	F(1,318) = 6.300*	F(1,318) = 28.130**
BFT[b]	*Neuroticism*	low < mid < high	F(2,318) = 3.781*	*n.s.*
TA[c]	*Competence*	mid < high	*n.s.*	*n.s.*
TA[c]	*Enthusiasm*	low < mid < high	*n.s.*	*n.s.*
TA[c]	*Pos. attitude, Neg. att.*	mid < high	*n.s.*	*n.s.*

[a] Experience, [b] Big Five Traits, [c] Technical Affinity
Note: *p <.05, **p <.001; *n.s.* not significant

cumulative logit link function was chosen given the ordinal scale of the dependent variables *naturalness* and *comprehensibility*. The parameters listed in Table 2 were entered as fixed effects. Additionally, their interactions with the parameter *Sentence type* were computed. Our models revealed a number of statistically significant differences in the perception of *naturalness* and *comprehensibility* (Table 2). In these cases, *post hoc* analyses were performed (Table 3): In accordance with our pilot study, the assessment of varying syntactic complexities appeared to depend on the linguistic knowledge of a user. Subjects, who indicated low or average experience in *Linguistics*, were more likely to assess *MCV* as more *natural* and *comprehensible* than highly experienced participants. Subjects with low experience revealed a higher odds ratio in preference for *MCV* than averagely experienced ones. Interestingly, a different user acceptance was also observed in relation with the self-assessed personality traits. The stronger a participant was associated to one of the personality traits *Conscientiousness, Agreeableness* and *Neuroticism*, a higher preference for *RCV* was found. The odds ratio of rating *MCV* as more *natural* or *comprehensible* was found higher for participants with an average and low manifestation of these traits. Significant effects for application-dependent contexts were mainly identified with regards to *comprehensibility*. We observed higher assessment ratings for *MCV* in the case of *DAS*, whereas *RCV* was preferred for contents of *COF*. Equally, we observed an effect between our syntactic variants and *Question type*s. *What* revealed a higher odds ratio in preference for *MCV* compared to *How* and *When*. A higher preference for *MCV* was found for *How* compared to *When*.

Table 3 *Post hoc* analyses for the interaction with *Sentence type* and syntactic preferences

	Parameter levels	***Naturalness*** Odds ratio [95% CI]	***Comprehensibility*** Odds ratio [95% CI]	**Interpretation:** *MCV* $\xleftrightarrow{\text{preference}}$ *RCV*
	DAS vs. *COF*	*n.s.*	0.512 [0.285, 0.920]	*DAS* - - - *COF*
	What vs. *When*	–	1.311 [1.186, 4.500]	*What* - *How* - *When*
	How vs. *When*		1.666 [0.907, 3.058]	
	What vs. *How*		1.387 [0.656, 2.932]	
Linguist.	low vs. high	4.80e9 [1.93e8, 1.19e11]	14.43e5 [4.71e5, 44.19e5]	low - mid - high
	mid vs. high	1.501 [0.687, 3.281]	2.470 [0.891, 6.846]	
	low vs. mid	3.19e9 [1.26e8, 8.08e10]	58.42e4 [11.38e4, 29.99e5]	
Consc.	mid vs. high	9.631 [1.337, 69.387]	*n.s.*	mid - - - high
Agree.	mid vs. high	15.635 [1.812, 134.929]	7.468 [3.549, 15.712]	mid - - - high
Neurot.	low vs. high	28.642 [2.283, 359.338]	*n.s.*	low - mid - high
	mid vs. high	19.797 [2.168, 180.752]		
	low vs. mid	1.447 [0.358, 5.843]		

Note: Redundant lines omitted. Results are based on *MCV* as referent.

As the automation level constantly increases, it should be mentioned that our data did not show any clear relation between the driving task's *Complexity* and our *Sentence type*s. There was no significant difference in the perceived *naturalness* nor *comprehensibility* between *AUT* and *MAN*. Considering user preferences, the long-term goal to design SDS interaction according to interpersonal models seems independent of the SAE level.

5.2 Evaluation of Objective Performance Measures

Driving performance measures were only assessed for the manual driving part *MAN*, since no data reflecting user performance was generated during *AUT*. The analyses of objective performance measures thus only refer to *MAN*.

A set of driving performance measures was logged at intervals of 20 ms. To ensure valid conclusions, we limited our data to sequences during which a voice prompt was played (12/participant) and computed standard deviations for *Speed* [km/h, *SP*], *Distance to lead vehicle* [m, *DL*] and *Lateral position* on the lane [m, *LP*] (Table 4).

We first compared all three performance measures for the combined baseline drives and sequences with voice output (including *MCV and RCV*; Table 4). A Wilcoxon signed-ranks test revealed significant results for *SP*, *DL* and *LP*. All three measures indicated higher values during voice output sequences than during the com-

Table 4 Standard deviations of performance measures assessed during baselines, *MCV* and *RCV* and results of Wilcoxon signed-ranks tests

	Speed (SP)	*Distance to Lead (DL)*	*Lat. Position (LP)*
Baselines	0.81	5.63	0.17
VO sequences (*MCV*+*RCV*)	1.03	13.24	0.24
MCV	1.06	11.95	0.190
RCV	1.07	13.22	0.187
Baselines vs. VO sequences	$Z = -8.34^*$, $r = .36$	$Z = -18.66^*$, $r = .79$	$Z = -15.69^*$, $r = .67$
MCV vs. *RCV*	$Z = -0.73$, $r = .03$	$Z = -6.12^*$, $r = .26$	$Z = -0.54$, $r = .02$

Note: $^*p < .001$.

bined baseline drives. These results coincide with the general consent that driving performance is influenced by the secondary task (limited to voice output sequences, in our case) given an increased cognitive load [31].

We further evaluated whether a difference exists in our data between driving performance during voice output sequences with *MCV* or *RCV* (Table 4). No significant results were found for *SP* and *LP*, but for *DL*. While *SP* and *DL* indicated slightly higher values for *RCV* compared to *MCV*, the opposite was observed for *LP*.

6 Discussion

In this paper we established the foundation for an adaptive NLG strategy concerning the syntactic design of voice output in SDS interaction as a secondary task. In comparison to previous theory-oriented approaches, we focused on experience and usability from a user perspective. We conducted a pilot study and argued that the approach of assessing linguistic preferences via audio represents a valid methodology in our context. The lacking awareness of syntactic variation results in intuitive user ratings. On this basis, we evaluated a WoZ experiment in a driving simulator and investigated the influence of syntax in voice output on the perceived *naturalness*, *comprehensibility* and driving performance. Our results indicate significant differences in the appropriateness of syntactic forms and their inherent complexity.

We have demonstrated a clear preference for SDS voice output in the form of *MCV* as opposed to *RCV*. This observation was first made in our pilot study and then repeated in the driving simulator. Apparently, SDS users subconsciously differentiate and prioritize different syntactic realizations in voice output. The concomitant structural complexity thus provably represents a distinguishing factor in the perceived *naturalness* and *comprehensibility* of voice output.

The investigation of the relationship between varying *Sentence type*s and several user and context parameters allowed a precise identification of the relevant factors to be considered in the design of voice output. As such, the *Domain* and *Question type* were indicated as dependent factors in the perceived *comprehensibility* of syntactic paraphrases. While for *RCV* a higher probability was found being rated as *comprehensible* in the context of *COF* and *When*, this holds for *MCV* in the context of *DAS, What* and *How*. A reason is found in the respective contents. While *COF* exclusively provides explanations concerning the recurring ensemble playing of various in-vehicle programs (*e.g.* music, fragrance, lighting, massage, *etc.*) to support the well-being of a driver, the contents for *DAS* are broader and more varied, ranging from an explanation for the cause, action and purpose of a driving assistant. Similarly, contents for *What* and *How* supply varied explanations as opposed to *When* (special case of *How*, Table 1) and refers to system limitations. The comparably complex and varied contents of *DAS, What* and *How* were thus reflected by a preference for simpler, linear structures. As soon as contents get less varied and simpler, *RCV* was the preferred sentence structure. Our results furthermore indicated individual user characteristics as dependent factors in the perception of syntax. We have observed an increased (decreased) probability for *RCV* (*MCV*) being rated as more *natural* and *comprehensible* the stronger a participant was associated to a personality trait (*Conscientiousness, Agreeableness, Neuroticism*) or experienced in *Linguistics*. This observation contradicts the results of our pilot study, where linguistically experienced users (in Part 3, at the latest) were aware of linguistic cues, had a clear opinion regarding the appropriateness of syntactic structures and preferred *MCV*. Since syntax was not revealed as an object of investigation in our driving simulation study, we emphasize here the intuitiveness of user evaluations. Accordingly, we relate the contradictory results to the additional secondary task and conclude that linguistically affine participants are used to process syntactically complex structures and thus evaluate them as more *natural* and *comprehensible*.

Finally, we have seen a direct influence of syntactic forms and their inherent complexity on driving performance. Besides the initial baseline comparison, our measures *SP* and *DL* confirmed the assumption that syntactic complexity is reflected in an increased cognitive load, measurable by driving performance, as they indicated higher values for *RCV* compared to *MCV*. Surprisingly, this effect was reversed for *LP*. Although research exists on the fact that at increased cognitive load a microsteering behavior occurs and improves lane keeping [15], we would then have expected the same relation for the initial baseline comparison. Overall and due to the performance measures available to us, no clear result was achieved here. The influence of different syntactic forms on driving performance should therefore be further investigated in future work.

7 Conclusions and Future Work

From our investigations, we conclude that there is no rigid default solution for *natural* and *comprehensible* SDS voice output in dual-task environments. Rather, the appropriateness of and preference for a syntactic structure (such as *MCV* vs. *RCV*) and their associated complexity depend on individual characteristics of the driver and the application context. At the same time, the choice of a particular syntactic structure has a direct influence on the performance of the primary task. It is therefore obvious that the design of vehicle-related voice output should be carefully chosen to minimize safety-critical aspects, such as cognitive load and driver distraction, while still satisfying user preferences to enable the most efficient form of interaction.

The present work provides the basis towards an adaptive NLG strategy. Based on the identified user and application parameters, future work will compare the role of syntactic structures with the use of syntax in natural linguistic behavior. In this way, it will be possible to elaborate a user-focused alignment strategy while simultaneously taking the interaction context as a parallel secondary task into account.

References

1. Aly A, Tapus A (2016) Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction. Auton Robots 40(2):193–209
2. Branigan HP, Pickering MJ, Pearson J, McLean JF (2010) Linguistic alignment between people and computers. J Pragmat 42(9):2355–2368
3. Buschmeier H, Baumann T, Dosch B, Kopp S, Schlangen D (2012) Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In: Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, pp 295–303
4. Buschmeier H, Bergmann K, Kopp S (2009) An alignment-capable microplanner for natural language generation. In: Proceedings of the 12th European workshop on natural language generation. Association for Computational Linguistics, pp 82–89
5. Chand V, Baynes K, Bonnici L, Farias ST (2010) Analysis of idea density (AID): a manual. University of California at Davis
6. Demberg V, Sayeed A, Mahr A, Müller C (2013) Measuring linguistically-induced cognitive load during driving using the ConTRe task. In: Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications. ACM, pp 176–183
7. Dijkstra P, Barelds DPH (2008) Do people know what they want: a similar or complementary partner? Evol Psychol **6**(4), 147470490800600,406
8. Engonopulos N, Sayeed A, Demberg V (2013) Language and cognitive load in a dual task environment. In: Proceedings of the annual meeting of the cognitive science society, vol 35
9. Hu Z, Tree JF, Walker MA (2018) Modeling linguistic and personality adaptation for natural language generation. In: Proceedings of the 19th annual SIGdial meeting on discourse and dialogue, pp 20–31
10. Isbister K, Nass C (2000) Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. Int J Hum Comput Stud 53(2):251–267
11. Jokinen K (2003) Natural interaction in spoken dialogue systems. In: Proceedings of the workshop ontologies and multilinguality in user interfaces. Crete, pp 730–734
12. Just MA, Keller TA, Cynkar J (2008) A decrease in brain activation associated with driving when listening to someone speak. Brain Res 1205:70–80
13. Karrer K, Glaser C, Clemens C, Bruder C (2009) Technikaffinität erfassen- der Fragebogen TA-EG. Der Mensch im Mittelpunkt technischer Systeme 8:196–201
14. Lee KM, Peng W, Jin SA, Yan C (2006) Can robots manifest personality?: an empirical test of personality recognition, social responses, and social presence in human-robot interaction. J Commun 56(4):754–772

15. Li P, Merat N, Zheng Z, Markkula G, Li Y, Wang Y (2018) Does cognitive distraction improve or degrade lane keeping performance? Analysis of time-to-line crossing safety margins. Transp Res Part F Traffic Psychol Behav 57:48–58
16. Mairesse F, Walker MA (2010) Towards personality-based user adaptation: psychologically informed stylistic language generation. User Model User-Adap Inter 20(3):227–278
17. Mairesse F, Walker MA (2011) Controlling user perceptions of linguistic style: trainable generation of personality traits. Comput Linguist 37(3):455–488
18. Moon Y, Nass C (1996) How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. Commun Res 23(6):651–674
19. Nass C, Moon Y, Fogg BJ, Reeves B, Dryer DC (1995) Can computer personalities be human personalities? Int J Hum Comput Stud 43(2):223–239
20. Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. Behav Brain Sci 27(2):169–190
21. Pinter Y, Reichart R, Szpektor I (2016) Syntactic parsing of web queries with question intent. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 670–680
22. Rambow O, Bangalore S, Walker MA (2001) Natural language generation in dialog systems. In: Proceedings of the first international conference on human language technology research, HLT 2001. Association for Computational Linguistics, Stroudsburg, pp 1–4
23. Rammstedt B, Danner D (2016) Die Facettenstruktur des Big Five Inventory (BFI). Diagnostica
24. Reiter E (1994) Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In: Proceedings of the seventh international workshop on natural language generation. Association for Computational Linguistics, pp 163–170
25. Rieser V, Lemon O, Keizer S (2014) Natural language generation as incremental planning under uncertainty: adaptive information presentation for statistical dialogue systems. IEEE/ACM Trans Audio Speech Lang Process 22(5):979–994
26. Roth-Berghofer T, Cassens J, Sørmo F (2005) Goals and kinds of explanations in case-based reasoning. In: Wissensmanagement, pp 264–268
27. Spieker P (1991) Natürlichsprachliche Erklärungen in technischen Expertensystemen. Ph.D. thesis, University of Kaiserslautern
28. Stier D, Sigloch E (2019) Linguistic design of in-vehicle prompts in adaptive dialog systems: an analysis of potential factors involved in the perception of naturalness. In: Proceedings of the 27th ACM conference on user modeling, adaptation and personalization. ACM, pp 191–195
29. Strayer DL, Cooper JM, Turrill J, Coleman JR, Hopman RJ (2016) Talking to your car can drive you to distraction. Cogn Res Princ Implic 1(1):16
30. Strayer DL, Drews FA, Johnston WA (2003) Cell phone-induced failures of visual attention during simulated driving. J Exp Psychol Appl 9(1):23
31. Strayer DL, Turrill J, Cooper JM, Coleman JR, Medeiros-Ward N, Biondi F (2015) Assessing cognitive distraction in the automobile. Hum Factors 57(8):1300–1324
32. Thomas P, Czerwinski M, McDuff D, Craswell N, Mark G (2018) Style and alignment in information-seeking conversation. In: Proceedings of the 2018 conference on human information interaction & retrieval. ACM, pp 42–51
33. Villing J (2009) In-vehicle dialogue management – towards distinguishing between different types of workload. In: Proceedings of SiMPE, fourth workshop on speech in mobile and pervasive environments (to appear)
34. Warren T, Gibson E (2002) The influence of referential processing on sentence complexity. Cognition 85(1):79–112
35. Wickens CD (2002) Multiple resources and performance prediction. Theor Issues Ergon Sci 3(2):159–177

Learning Grammar in Confined Worlds

Graham Spinks, Ruben Cartuyvels, and Marie-Francine Moens

Abstract In this position paper we argue that modern machine learning approaches fail to adequately address how grammar and common sense should be learned. State of the art language models achieve impressive results in a range of specialized tasks but lack underlying world understanding. We advocate for experiments with the use of abstract, confined world environments where agents interact with the emphasis on learning world models. Agents are induced to learn the grammar needed to navigate the environment, hence their grammar will be grounded in this abstracted world. We believe that this grounded grammar will therefore facilitate a more realistic, interpretable and human-like form of common sense.

1 Introduction

It is generally well understood that humans create abstract models of the world from which they draw inferences [2, 3]. A recent study in mice shows that visual inputs are encoded by predicting the impact of actions on the surroundings [7]. It makes sense that such world models are a byproduct of the necessity to deal with everyday tasks. The complex details of all possible interactions are abstracted to a level of detail that best suits the ability to navigate the world.

Most machine learning applications lack such interactions and reduce training to one or two objectives, e.g. achieve the best caption, achieve the most fluidity, etc. Some models excel at complex natural language benchmarks such as GLUE and SuperGLUE [12, 13] yet they fail to grasp basic, common sense. Studies have shown that these models exploit several fallible syntactic heuristics rather than reason about the underlying meaning [8]. A better goal would be to achieve a type of world

G. Spinks (✉) · R. Cartuyvels · M.-F. Moens
KU Leuven, Leuven, Belgium
e-mail: graham.spinks@cs.kuleuven.be

R. Cartuyvels
e-mail: ruben.cartuyvels@cs.kuleuven.be

M.-F. Moens
e-mail: sien.moens@cs.kuleuven.be

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_27

understanding or common sense, which we describe as the ability to understand things as they really are. This implies that data is processed and understood with respect to a broader, physical context.

We posit that 'making oneself understood' is a task that humans learn as part of the objectives in everyday life. We learn our language from interacting with teachers (humans, books, or even television) and with the world, because it inherently lets us navigate the world better and achieve our goals as humans.

Whereas animals possess a form of common sense, human world knowledge and understanding far exceeds that of animals. While common knowledge is often implicit in communication, and most likely results from world interaction, humans do use language to transmit complex information that relates to the world. It is therefore sensible to study how grammar and common sense relate to each other. It is interesting to see that language has developed across the world in many different ways, but that there are common factors. These universals include basic notions such as 'verbs' or complex ideas as 'Universal Grammar' [1]. The latter states that all humans are born with the capacity to learn and understand grammar.

Taken together, the ideas in this paragraph imply that human grammar is grounded in world interaction too, and we postulate that the resulting structured language facilitates the navigation of world models and common sense. Grammar in our view provides a set of tools that can be used to relate concepts and objects in our internal world models. This increases the capacity of internal models to express the necessary complexity to navigate the world. These models could therefore facilitate a more elaborate common sense understanding. To extrapolate this idea to machine learning, one could deploy models that are capable of learning grammar in an environment that induces them to learn it.

2 From Complexity to Controlled Experiments

Current approaches to NLP tend to rely on more and more complex datasets that are *solved* with larger and larger models. As noted by [6], meaning representations are mostly constructed following the distributional hypothesis [4], grounding symbols in other symbols instead of in reality. As a consequence, models learn to exploit word co-occurrences and unintended features in the dataset. This leads to nontransparent calculations and outputs that conflict with human understanding of the world.

An interesting avenue is to place models in virtual or physical environments that are affected by a temporal component. This idea has been expressed by other researchers, yet the push towards higher complexity has perhaps complicated this line of investigation. Many controlled experiments are possible which could advance our understanding of language learning. By setting up a 'confined world' that follows the virtual embodiment principles proposed by [6], different priors and models can be investigated. Rather than setting up 3D physical or photo-realistic simulation environments with complex interactions as in [5, 11], we suggest abstracting the physical world to a set of parameters of increasing complexity.

Some work in this direction has been done, where e.g. the emergence of language was studied in a cooperative multi-agent game in an abstracted environment [9]. Many more abstractions, tasks, situations, agents and interactions (e.g. human-computer vs. computer-computer) remain to be explored. [10] proposes agent-based interactions to test and sniff out the cognitive mechanisms that lead to the development of languages. Whereas this approach advocates for a semantic representation formalism based on second-order intentional logic, we believe that such agents should form their own implicit models of interactions and meaning. Our suggestion is therefore to focus on architectures that mirror the innate capability of brains to learn syntactical and morphological complexities of language [1] rather than relying on pre-specified symbolic formalisms. Artificial agents could, for example, exploit model architectures that explicitly detect (dis)agreement between language morphology and the state of the world. As a concrete example, for the sentence "The apples are falling", a model might be designed to detect that the sentence contains a plurality of apples as expressed by the subject and verb, and align it with observations of multiple falling apples from the evolution of the world state.

Concretely we propose abstracted environments that contain the following: spatial and temporal axes (coordinates, time), perception (vision and/or touch), grammar (a set of simplified rules to convey messages based on language universals), interaction and objectives. Objectives should prompt agents to implicitly learn a basic model of the world and the language rules to navigate it successfully. In these confined worlds, the importance of teachers, including whether their language supervision is needed at all, as well as the elements needed to express language in relation to the world can be investigated. Methods could focus on different model structures and learning algorithms that are better suited to induce grammar given different types of interactions. Comparing various types of objectives could lead to a better understanding of the implicit goals that align with language learning. We believe such experimentation will lead to a better understanding and new research directions, and enable valuable applications.

Acknowledgements This work is part of the CALCULUS project at the KU Leuven, funded with the European Research Council Advanced Grant H2020-ERC-2017-ADG 788506.

References

1. Chomsky N (2000) New horizons in the study of language and mind. Cambridge University Press, Cambridge
2. Forrester JW (1971) Counterintuitive behavior of social systems. Technol Forecast Soc Chang 3:1–22
3. Ha D, Schmidhuber J (2018) Recurrent world models facilitate policy evolution. In: Advances in neural information processing systems, vol 31, pp 2450–2462
4. Harris ZS (1954) Distributional structure. Word 10(2–3):146–162
5. Hermann KM, Hill F, Green S, Wang F, Faulkner R, Soyer H, Szepesvari D, Czarnecki WM, Jaderberg M, Teplyashin D, et al (2017) Grounded language learning in a simulated 3D world. arXiv preprint arXiv:1706.06551

6. Kiela D, Bulat L, Vero AL, Clark S (2016) Virtual embodiment: a scalable long-term strategy for artificial intelligence research. arXiv preprint arXiv:1610.07432
7. Leinweber M, Ward DR, Sobczak JM, Attinger A, Keller G (2017) A sensorimotor circuit in mouse cortex for visual flow predictions. Neuron 95(6):1420–1432
8. McCoy RT, Pavlick E, Linzen T (2019) Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007
9. Mordatch I, Abbeel P (2018) Emergence of grounded compositional language in multi-agent populations. In: Thirty-second AAAI conference on artificial intelligence
10. Steels L (2016) Agent-based models for the emergence and evolution of grammar. Phil Trans Royal Soc B Biol Sci 371(1701):20150,447 (2016)
11. de Vries H, Shuster K, Batra D, Parikh D, Weston J, Kiela D (2018) Talk the walk: navigating New York City through grounded dialogue. arXiv preprint arXiv:1807.03367
12. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) Superglue: a stickier benchmark for general-purpose language understanding systems. In: Advances in neural information processing systems, vol 32, pp 3261–3275
13. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: EMNLP 2018, pp 353–355

Corpora and Knowledge Management

A Content and Knowledge Management System Supporting Emotion Detection from Speech

Binh Vu, Mikel deVelasco, Paul Mc Kevitt, Raymond Bond, Robin Turkington, Frederick Booth, Maurice Mulvenna, Michael Fuchs, and Matthias Hemmje

Abstract Emotion recognition has recently attracted much attention in both industrial and academic research as it can be applied in many areas from education to national security. In healthcare, emotion detection has a key role as emotional state is an indicator of depression and mental disease. Much research in this area focuses on extracting emotion related features from images of the human face. Nevertheless, there are many other sources that can identify a person's emotion. In the context of MENHIR, an EU-funded R&D project that applies Affective Computing to support people in their mental health, a new emotion-recognition system based on speech is being developed. However, this system requires comprehensive data-management

B. Vu (✉)
FernUniversität in Hagen, Hagen, Germany
e-mail: binh.vu@fernuni-hagen.de

M. deVelasco
Universidad del Pais Vasco UPV/EHU, Leioa, Spain
e-mail: mikel.develasco@ehu.eus

P. Mc Kevitt
Ulster University, Derry/Londonderry, Northern Ireland
e-mail: p.mckevitt@ulster.ac.uk

R. Bond · R. Turkington · F. Booth · M. Mulvenna
Ulster University, Newtownabbey, Northern Ireland
e-mail: rb.bond@ulster.ac.uk

R. Turkington
e-mail: turkington-r@ulster.ac.uk

F. Booth
e-mail: booth-f@ulster.ac.uk

M. Mulvenna
e-mail: md.mulvenna@ulster.ac.uk

M. Fuchs · M. Hemmje
GLOBIT GmbH, Barsbüttel, Germany
e-mail: m.fuchs@globit.com

M. Hemmje
e-mail: matthias.hemmje@globit.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_28

support in order to manage its input data and analysis results. As a result, a cloud-based, high-performance, scalable, and accessible ecosystem for supporting speech-based emotion detection is currently developed and discussed here.

1 Introduction and Motivation

Affective Computing is an emerging inter-disciplinary field developing technology that attempts to detect, analyse, process, and respond to important psychological traits such as emotions, feelings, or behaviours with the goal of improving human-computer interaction [1]. Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is a 4-year project funded by the European Union (EU), that applies Affective Computing to enhance and advance future healthcare processes and systems, especially in providing assistance to people with dementia, medical professionals, and caregivers [2]. By gathering activity and related sensor data to infer the emotional state of the patient as a knowledge stream of emotional signals, SenseCare can provide a basis for enhanced care and can alert medics, professional carer, and family members to situations where intervention is required [3, 4].

One of the systems developed in SenseCare is a machine-learning-based emotion detection platform [5] which provides an early insight into the emotional state of an observed person. SenseCare can process a live video stream or a pre-recorded video which enables analysis to be completed on the fly or at a later stage. Similar to SenseCare, the MENtal Health monitoring through InteRactive conversations (MENHIR) is a EU-funded project that aims to support and improve the mental wellbeing of people by applying Affective Computing, especially conversational technologies, such as emotion recognition in speech, automatic conversation management (chatbots), and other multidisciplinary topics [6]. According to the World Health Organization (WHO), mental, neurological, and substance use disorders make up 10% of the global, and 30% of non-fatal, disease burden. The global economy loses about US$ 1 trillion per year in productivity due to depression and anxiety [7].

In MENHIR, new research assists people with improving their current state of emotion and provides a long-term overview of their state over time. A machine-learning-based emotion detection platform has been developed. Unlike SenseCare, where human emotions are extracted from a live video stream or a pre-recorded video, the MENHIR emotion detection platform identifies emotions from speech. The system relies on short-term features such as pitch, vocal tract features such as formants, prosodic features such as pitch loudness, as well as speaking rate to perform effectively. Furthermore, recurrent neural networks are applied to predict emotion in real-time using a 3D emotional model. This paper discusses the challenges of emotion detection based on speech and its corresponding transcription in the MENHIR project. Furthermore, it provides a solution to overcome these challenges. The architecture of the proposed system and its constituent components are described. Finally, we conclude and discuss future work.

2 Problem Statement

One of the goals of the MENHIR project is to further extend the results of earlier research work, expanding the set of identified depressive speech acoustic features and automating their detection so that depressed and anxious speech can be accurately distinguished from healthy speech [8]. To enable this, challenging scenarios need to be considered and overcome as discussed here.

After a series of human-to-human counselling conversations are recorded in a laboratory setting, a corpus of audio data of conversations is formed. Along with the audio files, their metadata, which consists of documents describing the conversations and spreadsheets describing the conversation results, are also provided for advanced annotation and analysis. All these data need to be stored in a high-performance repository where other analysis systems can connect to and download them when needed. Furthermore, multimedia objects usually take up a lot of storage space. This means the data repository also needs to be scalable to fulfil users' demands in the future.

In MENHIR, not only multimedia objects but also other kinds of scientific content, knowledge, and their metadata need to be imported, stored, and managed. Sharing and exchanging research results powers collaborative and co-creative networking among project participants. Therefore, a solution is needed to support the ingestion of scientific publications from different sources. Here, the imported content can be managed and transformed into learning materials. Similar to multimedia objects, scientific data content also needs high-performance, scalable, and fault-tolerant storage. Furthermore, a content management system will enable users to edit, share, and publish their content.

There are a number of collaborative services producing analysis results and generating observed subject and patient conversational behaviour, such as, authentication, authorization, speech analysis data services, big data speech analysis, collaboration and coordination services, psychological/affective analytics, reporting/result sharing and reproducibility services [8]. It is crucial to have an integration architecture for all the mental health services and applications employed in MENHIR. This architecture will provide a common platform for these systems to communicate in a predefined flow, where input data is received and results are stored.

For research results to make an impact, they need to be easily found and used. Meanwhile, related publications, datasets, and analysis results are distributed in different locations. Therefore, one needs to find a means to automatically gather and combine all these resources into scientific asset packages. Otherwise, users can only find fragments of related information. It will prevent them from having a complete overview of the research topic and discovering important relationships between factors. Organizing related information and data into scientific asset packages is a powerful method of systemizing results produced by conversational technologies.

Finally, classification helps to narrow the choices among content, information, and knowledge resources. By dividing the material into reduced subsets, classification can make content, information, and knowledge resource retrieval and access faster and more accurate [9]. In MENHIR, a considerable volume of subject data will be analysed by an emotion detection server and will be made available for use by, for example, chatbots. Furthermore, the analysis results and related scientific publications will also be generated and managed in MENHIR. Without organizing the content created into suitable categories, researchers will not have capacity for insight on the key data produced in MENHIR, discover connections between data or whether something is missing. Therefore, a system that allows the content, knowledge, analysis results, and datasets to be classified is critical for the success of MENHIR.

3 System Design

Based on these challenges, we have developed a system design to support MENHIR in the task of conversational technologies research and development. In this section, a cloud-based Content and Knowledge Management Ecosystem (KM-EP) for audio files and metadata persistence, human emotion detection, as well as asset packaging, classification, and management is introduced and described. Figure 1 illustrates the architecture of the system, which comprises the MENHIR KM-EP and supporting systems.

The MENHIR Content and Knowledge Management Ecosystem (KM-EP) provides a platform for managing scientific as well as educational content and knowledge resources. Furthermore, the KM-EP will act as a framework for researchers to deploy their work without spending time reimplementing basic functionalities, such as, e.g. user management and task scheduling. In Fig. 1, four components of the MENHIR KM-EP, which are related and crucial for the tasks of audio data persistence, emotion detection, as well as asset packaging, classification, and management, are shown. The components are Media Archive (MA), Digital Library (DL), Taxonomy Manager (TM), and Asset Manager (AM).

The first component that is needed for the MENHIR KM-EP is the Media Archive (MA). The MA manages all multimedia objects in the KM-EP. The MA enables users to create, persist, manage, and classify different types of multimedia objects, such as, e.g. video, audio, images, presentation slides, along with their metadata. In MENHIR, audio files and their metadata need to be imported and stored together in the system. The initial audio files are, e.g., recordings of interviews, which are conducted in order to form a corpus of conversational audio data. This corpus will be used to validate the operation of the Emotion Detection Server. Their metadata consists, e.g., of documents describing the interviews and spreadsheets describing the interview results. Furthermore, audio files of conversations and interviews can also be uploaded and linked to user accounts automatically or manually. The KM-EP provides an interface where users can upload these files into the system and populate basic metadata information related to them, such as, e.g. title, description, authors, and creation date. The uploaded files are stored in a cloud storage service, which is fault-tolerant, flexible, scalable, and has high performance. This will enable users to have fast and stable access to the files worldwide. Furthermore, with the support of

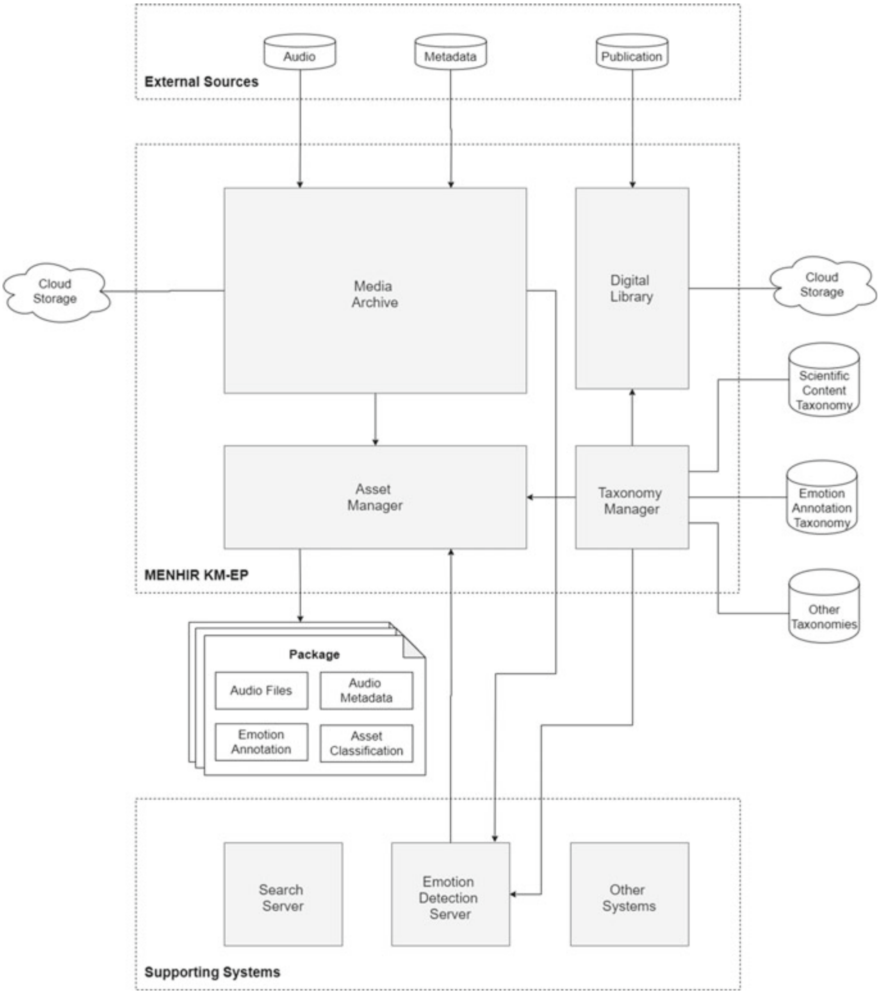


Fig. 1 Architecture of MENHIR content and knowledge management ecosystem (KM-EP)

the TM component, multimedia objects can be classified into different categories. Classification enables objects to be searched and accessed easily and quickly by users.

The next component of the KM-EP is the Digital Library (DL). The DL enables users to import publications into the KM-EP, persist, and manage them. Using a Mediator-Wrapper Architecture, publications from different sources, such as, e.g. Mendeley [10], SlideShare [11], and in different formats, such as BibTex [12] and OAI-PMH [13], can be queried, uploaded, and integrated into the DL [14]. Similar to the MA, after importing or creating a new publication using the DL, users have the option to fill in its metadata, such as, e.g. title, abstract, conference, publisher, and upload the document. The uploaded files will also be stored in cloud storage to maintain their availability and scalability. By indexing file metadata and classifying publications into existing categories, they can be searched by users based on these criteria.

The Taxonomy Manager (TM) component supports the construction, collaboration, management, and evolution of taxonomies. The TM enables users to develop and manage their own taxonomies. With the support of its version control system, users can manage the changes of their taxonomies. Every modification is tracked and can be reverted. A complete history of changes helps users to compare different versions of a taxonomy. Furthermore, the branching feature enables users to create multiple versions of a taxonomy.

Multimedia objects, publications, and assets of the MENHIR KM-EP can be classified with support of the TM. As a result, users can search and browse contents quickly and easily. Classification also enables navigation inside the KM-EP. A persistent identifier introduced for each term in a taxonomy enables taxonomy evolution without affecting existing classifications. A rating system is implemented based on crowd voting to support the evaluation of taxonomies in the KM-EP. With the rating system, authors can improve the accessibility of their taxonomies, and users can also choose quickly more relevant taxonomies. A caching system enables thousands of taxonomies and terms to be retrieved and constructed in just a few milliseconds.

In MENHIR, the TM can not only be used to collect, classify, and provide access to audio materials from initial emotion analysis and results but can also support the emotion detection platform by providing an emotion annotation taxonomy. The machine learning platform can use this taxonomy to label its training and validation set. This creates a standard emotion classification that can be used for classifying results produced later by the platform. This process would be more costly without the classification, annotation, and access support of the TM in the MENHIR KM-EP supporting scientific research in the domain of Affective Computing.

The Asset Manager (AM) component is where related data, metadata, analysis results, and classification are gathered and combined into packages. In order to do this, a cronjob is developed and scheduled to run regularly after a given period of time. This cronjob has 3 tasks, which are: (1) searching for new audio files and their metadata and adding them into a new asset, (2) sending the new audio files and their metadata to the emotion detection server for analysis, and (3) receiving and adding analysis results into its package. This guarantees that new data will always be processed after it is uploaded to the MENHIR KM-EP.

After the cronjob has been started, the daemon searches for audio files along with their metadata in the MA. For each audio record found, the daemon will check if it belongs to a package in the AM or not. If it exists in a package, the emotion detection process, along with other processes, has been already completed for this audio record and the daemon continues to work with other audio records. Otherwise, the daemon needs to gather necessary data, such as uploaded files, documents describing the counseling interview where the audio file was recorded, and the spreadsheets describing the interview results. These files will be downloaded from the current cloud storage service to a temporary location in the local server. Next, the daemon

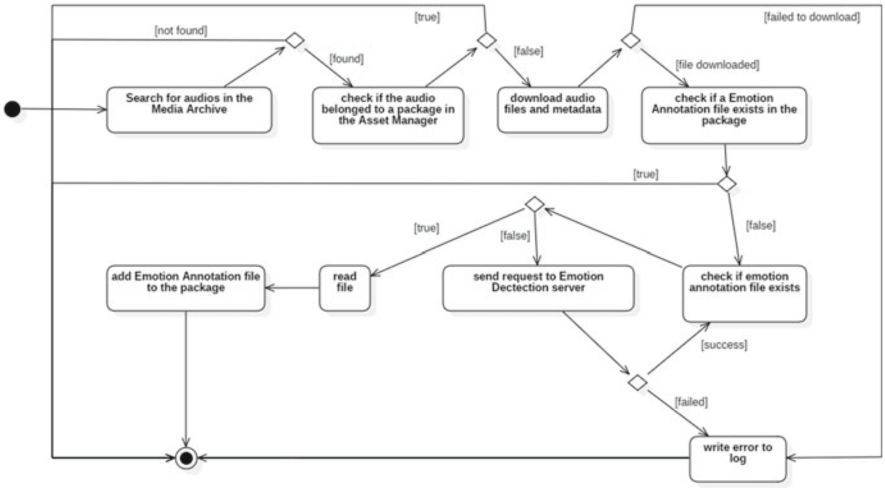


Fig. 2 Activity diagram of the asset manager (AM) cronjob

checks if an emotion annotation was produced for the audio record. If it has been produced, the daemon will go to the next step. Otherwise, it will search for the annotation file produced by the Emotion Detection server. If this file does not exist, the daemon will send a request to the server and let it process the downloaded audio file along with its metadata. An annotation file will then be produced by the server. The daemon reads the file and adds it to the new package. Errors that occur whilst the daemon is running will be written to the log file in order to enable the system administrator to debug them later. Figure 2 describes the activity flow of the cronjob.

After an emotion annotation file is added, users can use the Emotion Audio Player (EAP), which is an important feature of the AM, to play the audio files and discover the current emotional state of the subject in the audio. The emotion in the annotation file will be indexed based on its timestamp. When the audio playback reaches a timestamp, the player will display the emotion associated with it. Furthermore, with this annotation file, emotions of the subject in the audio file can be visualized with various visualization techniques. This enables users to have an overview of the current emotional state of the subject and provides an opportunity to explore hidden information behind human emotion.

Finally, the AM enables users to edit, delete and classify their packages through interacting with a user interface. Not only audio records and respective analysis results inside a package can be classified using the scientific content, emotion annotation, and other types of taxonomies, but the package itself can be classified into different categories using the TM. This classification will be stored in the package as well as indexed in a search server. With the AM, scientific content and respective analysis results can be managed in a central repository. This will reduce dramatically the effort to deploy, maintain, search, and reuse scientific data.

Supporting systems such as Emotion Detection Server and Search Server provide standalone, high-performance services that the MENHIR KM-EP can take advantage of. They provide interfaces, so the KM-EP can send requests and later receive results. In the context of this paper, we focus on the Emotion Detection Server, which is being developed in MENHIR.

The Emotion Detection Server detects human emotion from speech signals extracted from the audio files and their transcriptions. The files will be downloaded from the MA and sent to the processing server by the introduced cronjob from the AM component. The audio samples are processed by the server and their results are exported to annotation files and stored in the local server for the KM-EP to access and use. Automatic recognition of spontaneous emotions from speech is complex [15]. To overcome its challenges, two procedures have been conducted. The first one is the annotation task, that involves the segmentation of the audio samples in order to label them with emotions, and the second one is building a model that is able to distinguish between different emotional states.

In relation to the annotation, transcriptions are used to identify the spoken turns, and those turns have been split automatically into segments of between 2 and 5 5 s, because it is known that there is no emotion change in this time window. Subsequently, each segment is labelled by both professional and crowd annotators following the same questionnaire. The questionnaire includes both categorical and dimensional annotation (valence, arousal, and dominance). Using these annotations, we have experimented with the creation of a model capable of identifying the mood of the speaker through application of neural network algorithms. This model infers the subject's emotional state using both audio features (such as e.g. pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs)) or the spectrogram. With this model, an emotion detection server can be developed to provide the MENHIR KM-EP with emotion annotations from both acoustic signals and their corresponding transcription in real-time.

A high-performance search server is needed to index content objects in the MA and the DL, so they can be searched quickly by users. Furthermore, indexing classifications enables faceted search, which is a way to add specific, relevant options to the results pages, so that when users search for content, they can see where in the catalogue they've ended up [16]. With the faceted feature, users can have an overview of the classification of contents in real-time and quickly find results by selecting only relevant categories. Furthermore, faceted search enables navigation using taxonomy hierarchies, which are created and managed using the TM [17].

Besides a search server, other systems, such as, e.g. cache server, queuing system, are also important for the MENHIR KM-EP. Caching improves performance of the system by pre-processing and storing frequently used data in memory, each time it is required, it can be retrieved from there without requiring reconstruction resources. A queuing system enables the KM-EP to process data in an organized manner. Processing all data at once requires considerable computing power and resources. Therefore, organizing data into a queue and processing them accordingly would allow the resource to be distributed evenly and reduce stress on system components.

4 Conclusion and Future Work

The MENHIR project provides rapid intervention, appropriate feedback and overview on the state of development of subject mood and anxiety levels over time, by monitoring moods, behaviour, and symptoms of subjects in real time. The objective of the work reported in this paper is to develop an integration platform to support the ingestion and management of audio files and their metadata, results on human emotion detection from speech, and scientific asset packaging, classification, and management.

Here, we have described the challenges involved in the development and integration of such a platform. The content and knowledge management ecosystem (KM-EP) proposed here is a cloud-based, high-performance, scalable, and easy to use solution. By relying on its Media Archive and Digital Library, the KM-EP is able to ingest, modify, share, and preserve scientific publications and multimedia objects, such as audio files and their metadata. The Taxonomy Manager enables users to classify content and knowledge, which leads to better quality and faster exploration. Finally, the Asset Manager combines related scientific publications, multimedia objects, datasets, and analysis results into packages. With the Asset Manager, all the related data, information, and knowledge can be gathered and managed in one central repository, which is easier to maintain and reuse. The MENHIR KM-EP will provide a useful foundation for the development of conversational systems in mental health promotion and assistance.

The current emotion detection server uses a model, which needs to be trained offline by AI experts. This model also needs to be re-trained frequently with updated corpora to enhanced its accuracy. The MENHIR KM-EP can be extended in the future to use the uploaded audio records in the MA to form a new corpus. Then, the new model can be trained based on the new data corpus and replace the former model automatically. By doing this, the cost of developing an advanced emotion detection model can be reduced.

Acknowledgements This publication has been produced in the context of the MENHIR project. This project has received funding from the European Union's H2020 Programme under grant agreement No. 823907. However, this paper reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

1. Healy M, Donovan R, Walsh P, Zheng H (2018) A machine learning emotion detection platform to support affective well being. In: IEEE international conference on bioinformatics and biomedicine
2. Sensor enabled affective computing for enhancing medical care, 19 April 2017. https://cordis.europa.eu/project/rcn/199563/factsheet/en. Accessed 27 Aug 2019

3. Engel F, Bond R, Keary A, Mulvenna M, Walsh P, Zheng H, Wang H, Kowohl U, Hemmje M (2016) SenseCare: towards an experimental platform for home-based, visualisation of emotional states of people with dementia. In: Advanced visual interfaces. Supporting big data applications
4. Healy M, Walsh P (2017) Detecting demeanor for healthcare with machine learning. In: IEEE international conference on bioinformatics and biomedicine
5. Donovan R, Healy M, Zheng H, Engel F, Vu B, Fuchs M, Walsh P, Hemmje M, Kevitt PM (2018) SenseCare: using automatic emotional analysis to provide effective tools for supporting wellbeing. In: IEEE international conference on bioinformatics and biomedicine
6. MENHIR. https://menhir-project.eu/. Accessed 20 Jan 2020
7. Mental health. WHO, 2 Oct 2019. https://www.who.int/news-room/facts-in-pictures/detail/mental-health. Accessed 20 Jan 2020
8. Consortium M (2018) MENHIR proposal. European Commission
9. Vu B, Mertens J, Gaisbachgrabner K, Fuchs M, Hemmje M (2018) Supporting taxonomy management and evolution in a web-based knowledge management system. In: HCI 2018, Belfast, UK
10. Mendeley. https://www.mendeley.com/. Accessed 27 Jan 2020
11. SlideShare. https://de.slideshare.net. Accessed 27 Jan 2020
12. Your BibTeX resource (2016). BibTeX. http://www.bibtex.org/. Accessed 28 Oct 2019
13. Protocol for Metadata Harvesting. Open Archives Initiative. https://www.openarchives.org/pmh/. Accessed 28 Oct 2019
14. Vu B, Wu Y, Afli H, Kevitt PM, Walsh P, Engel F, Fuchs M, Hemmje M (2019) A metagenomic content and knowledge management ecosystem platform. In: IEEE international conference on bioinformatics and biomedicine, San Diego, USA
15. de Vázquez M, Justo R, López Zorrilla A, Inés Torres M (2019) Can spontaneous emotions be detected from speech on TV political debates? In: IEEE international conference on cognitive infocommunications
16. What is faceted search and navigation? (2016) Loop54. https://www.loop54.com/knowledge-base/what-is-faceted-search-navigation. Accessed Apr 2019
17. Vu B, Donovan R, Healy M, Kevitt PM, Walsh P, Engel F, Fuchs M, Hemmje M (2019) Using an affective computing taxonomy management system to support data management in personality traits. In: IEEE international conference on bioinformatics and biomedicine, San Diego, USA

A Script Knowledge Based Dialogue System for Indoor Navigation

Juliana Miehle, Isabel Feustel, Wolfgang Minker, and Stefan Ultes

Abstract We present an indoor navigation system that is based on natural spoken interaction. The system navigates the user through the University of Ulm based on scripts, supporting three different routes and varying communication styles for the system descriptions. Furthermore, the system is able to cope with incomplete scripts and inconclusive situations by passing the dialogue initiative to the user. In order to create the scripts, a data collection has been carried out in which 97 native speakers described the routes with the help of videos. In the end, the system has been evaluated in a user study with 30 participants. The work is part of the research project "MUNGO—Multi-User Indoor Navigation Using Natural Spoken Dialog and Self-learning Ontologies", co-funded by the 2018 Google Faculty Research Award.

1 Introduction

Compared to classical GPS based navigation, indoor navigation poses some additional challenges, one of which has long been the task of accurate localisation. Recent breakthroughs in indoor localisation systems (e.g. [1] and [6]) encourage research on improved user experience. This is why we address the task of indoor navigation using a dialogue system based on natural spoken interaction. In order to do so, we have implemented a script knowledge based dialogue system which supports different routes within the University of Ulm and varying communication styles for the system descriptions.

J. Miehle (✉) · I. Feustel · W. Minker
Institute of Communications Engineering, Ulm University, Ulm, Germany
e-mail: juliana.miehle@uni-ulm.de

I. Feustel
e-mail: isabel.feustel@uni-ulm.de

W. Minker
e-mail: wolfgang.minker@uni-ulm.de

S. Ultes
Mercedes-Benz Research & Development, Sindelfingen, Germany
e-mail: stefan.ultes@daimler.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_29

2 Functionalities

With the help of our system, the user is able to navigate through the University of Ulm on three different ways. The system can cope with incomplete scripts and inconclusive situations by passing the dialogue initiative to the user. The communication is based on natural spoken interaction and the user can perform the following actions:

- Request Navigation: The user can ask the system for navigation (e.g. "Where do I have to go?", "What's next?").
- Offer Navigation: The user can give route descriptions in case the system has an incomplete script (e.g. "I turn left at the stairs.").
- End Navigation: During the dialogue, the user can end the navigation at any point if he is not satisfied with the interaction (e.g. "I want to end the navigation.").

Besides these navigation actions, the user can accept or acknowledge the system output or ask the system to repeat the previous description, e.g. in case he did not understand the last output. Our system supports four different communication styles for the route descriptions, following Pragst et al. [5] who have shown that the communication styles *elaborateness* and *directness* influence the user's perception of a dialogue and are therefore valuable possibilities for adaptive dialogue management:

- Elaborate, direct (e.g. "Go straight and turn left near the stand containing magazines in order to reach the stairs.")
- Concise, direct (e.g. "Go straight to the stairs.")
- Elaborate, indirect (e.g. "Find the stairs to the left of the stand containing magazines.")
- Concise, indirect (e.g. "Find the stairs.")

The *elaborateness* refers to the amount of additional information provided to the user and the *directness* describes how concretely the information that is to be conveyed is addressed by the speaker. The selection of the system's communication style is based on three different strategies:

- A fixed strategy, which does not change the communication style during an interaction.
- A random strategy, which randomly selects the communication style at each turn.
- An adaptive strategy, which adapts the system's communication style to the user's style.

In addition to the navigation dialogues, the system supports a small talk scenario. This can be used to estimate the user's communication style before the navigation takes place. The communication style estimation is based on the classification approach presented in [3].

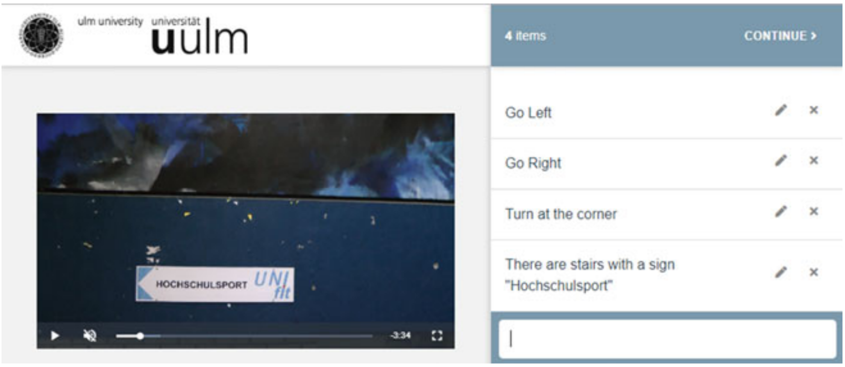


Fig. 1 The video description tool used for the data acquisition

3 Data Acquisition

In order to create the scripts for our indoor navigation dialogue system, we have collected route descriptions from native speakers based on videos. The aim was to acquire data for creating the scripts and to investigate the different description styles of human users in order to integrate them in our dialogue system. We have chosen three routes within the University of Ulm and made videos while walking along the way. The routes contain different points of interest and path elements like the cafeteria, an elevator or stairs. The start and end points of the paths are linked which means that route one ends where route two starts and the end point of route two is the starting point of route three. In order to collect the descriptions, we implemented a web page containing a short demographic questionnaire (age, gender and country of residence) and the video description tool shown in Fig. 1. We have used the platform clickworker[1] to find English native speakers for our route descriptions. Overall, we were able to collect 74 complete scripts for each route. These descriptions were then used to create the XML scripts for our dialogue system. An excerpt of a script is shown in Fig. 2, including examples for the different communication styles.

4 Implementation

Our indoor navigation dialogue system, which has been implemented in Python 3.5, consists of two main components: the Dialogue Handler and the Dialogue Manager. The Dialogue Handler initiates the dialogue with the request for the first system action to the Dialogue Manager. The Dialogue Manager has a stack of system actions which was set within the initialisation. The stack is filled with the actions extracted from the XML script. The Dialogue Manager returns the next system action to the Dialogue

[1] https://www.clickworker.com

```
<path>
  <pathelement id='1' nextElement='2'>
    <concise_direct>
      Go straight to the stairs.
    </concise_direct>
    <elaborate_direct>
      Go straight and turn left near the stand containing magazines
      in order to reach the stairs.
    </elaborate_direct>
    <concise_indirect>
      Find the stairs.
    </concise_indirect>
    <elaborate_indirect>
      Find the stairs to the left of the stand containing magazines.
    </elaborate_indirect>
  </pathelement>
  <pathelement id='2' nextElement='3'>
    <concise_direct>
      Go down the stairs.
    </concise_direct>
    <elaborate_direct>
      Go down the stairs and pass the sign saying "Hochschulsport".
    </elaborate_direct>
    <concise_indirect>
      The path continues down the stairs.
    </concise_indirect>
    <elaborate_indirect>
      The path continues down the stairs, passing a sign saying
      "Hochschulsport".
    </elaborate_indirect>
  </pathelement>
  ...
</path>
```

Fig. 2 Excerpt of a script, showing the first and the second path element

Handler and the Dialogue Handler activates the Text-to-Speech Synthesis. In parallel, a background job is running that listens to the next user input and enables barge-ins. As soon as a speech signal is detected, the voice is recorded and the audio file is sent to the Google Cloud Speech-to-Text API[2]. It returns the text which is passed to the machine-learning based RASA Natural Language Understanding component[3]. The user's intention is classified and assigned to a user dialogue action. If the classification fails, the Dialogue Handler immediately triggers the Dialogue Manager to return a dialogue action asking the user to repeat the previous input. Otherwise, the user action is enriched by the result of the user communication style classifier. The user action is then passed to the Dialogue Manager with the request for the next system action. The Dialogue Manager then decides whether the next description is taken from the system actions stack, a request is created (e.g. ask the user to repeat the input or to go on with the navigation as the script is incomplete) or the dialogue is terminated.

[2]https://cloud.google.com/speech-to-text

[3]https://rasa.com/docs/nlu

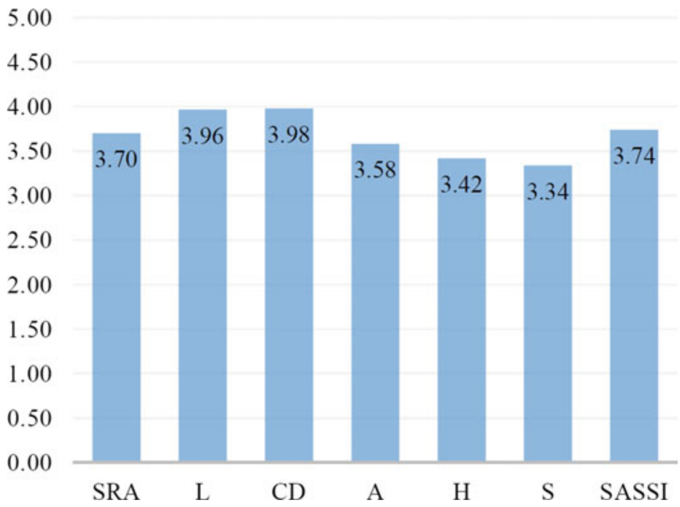


Fig. 3 The results of the participants' ratings grouped by the different categories of the SASSI questionnaire, whereby the rating scale has been inverted for the negatively formulated statements so that the optimal rating is five for every statement

5 Evaluation

In order to evaluate our indoor navigation dialogue system, we have conducted a user study where 30 participants used the system to navigate around the University of Ulm. The course of the study was as follows:

1. Introduction: The participants were introduced to the experimental setting and signed a consent form declaring voluntary participation and consent to the use of their data.
2. Small talk dialogue: The participants conducted the small talk scenario where the system asked questions like "How are you?" or "What are your hobbies?"
3. Navigation dialogues: The participants completed three navigation tasks.
4. Questionnaire: After each navigation task, the participants had to rate statements about the dialogue on a five-point Likert scale in an on-line questionnaire. The statements were taken from the questionnaire developed by Hone and Graham [2] measuring the subjective assessment of speech system interfaces (SASSI). It comprises the categories *System Response Accuracy (SRA)*, *Likeability (L)*, *Cognitive Demand (CD)*, *Annoyance (A)*, *Habitability (H)* and *Speed (S)*. In the end, the participants were asked to complete a short demographic questionnaire.

The participants were recruited via flyers and mailing lists at the university and received an expense allowance of 10 EUR. Overall, 30 people participated in the study. 16 participants were male and 14 were female, the average age was 25.7 years and most of the participants have never used speech-based assistants.

The results of our user study are shown in Fig. 3. It can be seen that we obtain an overall user satisfaction of 3.74 by calculating the average of all ratings of the SASSI questionnaire. The categories *Likeability (L)* and *Cognitive Demand (CD)* achieve the best ratings. The *Likeability* category includes statements of opinion about the system as well as feeling/affect items. The *Cognitive Demand* category comprises items that summarise both the perceived level of effort needed to use the system and user feelings arising from this effort. Thus, we can conclude that the participants have a good feeling while using the system and that the level of effort needed to use the system is appropriate. The categories *Habitability (H)* and *Speed (S)* received the worst ratings, showing that the participants did not always know what to say and what the system is doing and that the system sometimes responded slowly. One explanation for these two ratings is probably that we encountered problems with the WLAN connection while walking. If the WLAN connection was bad, the response time of the system increased and the Speech Recognition did not work properly. Moreover, the limitation of the system that it only allows the user to ask for the next description or repeat the previous description does not allow for further clarification inquiries of the user (e.g. "Do you mean the blue door?").

6 Conclusion

We have presented an indoor navigation system that is based on natural spoken interaction. The system navigates the user through the University of Ulm based on scripts, supporting three different routes and varying communication styles for the system descriptions. In order to create the scripts, we have carried out a data collection in which 97 native speakers described the routes with the help of videos. Moreover, we have presented the evaluation results of a user study with 30 participants. Using the SASSI questionnaire [2], which has been developed in order to measure the subjective assessment of speech system interfaces, we have obtained an overall user satisfaction of 3.74 for our system. While the categories *Likeability* and *Cognitive Demand* have achieved the best valuation, the ratings in the categories *Habitability* and *Speed* show that there is still room for improvement. So far, the system only allows the user to ask for the next description or repeat the previous description, yet it does not allow for further clarification inquiries of the user (e.g. "Do you mean the blue door?"). In order to allow for such requests, in future work, we will expand the system with semantic information. Moreover, we will investigate the influence of different communication styles during navigation as Miehle et al. [4] have shown that the user satisfaction may be increased by adapting the system's communication style to the user.

Acknowledgements The work is part of the research project "MUNGO—Multi-User Indoor Navigation Using Natural Spoken Dialog and Self-learning Ontologies" that has received funding from the 2018 Google Faculty Research Award. Additionally, this work has received funding within the BMBF project "RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with

Learning Autonomous Robots" and the technology transfer project "Do it yourself, but not alone: Companion Technology for DIY support" of the Transregional Collaborative Research Centre SFF/TRR 62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Frassl M, Angermann M, Lichtenstern M, Robertson P, Julian BJ, Doniec M (2013) Magnetic maps of indoor environments for precise localization of legged and non-legged locomotion. In: 2013 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 913–920
2. Hone KS, Graham R (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat Lang Eng 6(3–4):287–303
3. Miehle J, Feustel I, Hornauer J, Minker W, Ultes S (2020) Estimating user communication styles for spoken dialogue systems. In: Proceedings of the 12th international conference on language resources and evaluation (LREC 2020). European Language Resources Association (ELRA)
4. Miehle J, Minker W, Ultes S (2018) Exploring the impact of elaborateness and indirectness on user satisfaction in a spoken dialogue system. In: Adjunct publication of the 26th conference on user modeling, adaptation and personalization (UMAP). ACM, pp 165–172
5. Pragst L, Minker W, Ultes S (2019) Exploring the applicability of elaborateness and indirectness in dialogue management. In: Eskenazi M, Devillers L, Mariani J (eds) Advanced social interaction with agents, vol 510. Lecture notes in electrical engineering. Springer, Cham, pp 189–198
6. Robertson P, Frassl M, Angermann M, Doniec M, Julian BJ, Puyol MG, Khider M, Lichtenstern M, Bruno L (2013) Simultaneous localization and mapping for pedestrians using distortions of the local magnetic field intensity in large indoor environments. In: International conference on indoor positioning and indoor navigation. IEEE, pp 1–10

Data Collection Design for Dialogue Systems for Low-Resource Languages

Zulipiye Yusupujiang and Jonathan Ginzburg

Abstract This paper presents our plan and initial design for constructing a dialogue corpus for a low resource language, in this case Uyghur, with the ultimate goal of developing a dialogue system for Uyghur. We plan to design and create a Massively multiplayer online role-playing game (MMORPG), using the `RPG Maker MV Game Engine`. We also introduce our initial design of a method for collecting various types of naturally generated questions and answers from native Uyghur speakers. Our method and the design of the game can be used for other low resource languages for collecting a large amount of dialogue data, which is crucial for implementing a dialogue system for such languages.

1 Introduction

Dialogue systems for major languages such as English, French, and Chinese are widely available with high quality performance, whereas for many other languages this remains an unexplored area. A large amount of language data is necessary to train a dialogue system. However, many languages which are under-resourced have very little if any digital resources for such technology. As a result, speakers of low resource languages are compelled to use other majority languages which have wider usage and better communication in the digital society. This can be threatening for the development of low resource languages. It is, therefore, essential to conduct more studies on those languages and make real efforts to digitize language resources and develop advanced NLP technologies for them.

Building an efficient dialogue system requires a large dialogue corpus of that language. However, collecting a sufficient amount of natural dialogue data for low resource languages is challenging, especially if there is no possibility for field work. In view of this, we aim to propose an efficient method for collecting natural dialogue

Z. Yusupujiang (✉) · J. Ginzburg
Université de Paris CNRS, Laboratoire de Linguistique Formelle, Paris, France
e-mail: zulipiye.yusupujiang@linguist.univ-paris-diderot.fr

J. Ginzburg
e-mail: yonatan.ginzburg@univ-paris-diderot.fr

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_30

data from native speakers of such languages. In this study, we focus on collecting natural dialogues in Uyghur, a low resource language, and most importantly, a language under severe threat since it is officially banned in all educational systems in Xinjiang Uyghur Autonomous Region in China. After collecting the dialogue data using the method we propose, we will conduct a detailed study of the corpus to characterize the response space of questions in Uyghur and will model a formal grammar of questions and answers in Uyghur. Ultimately, we aim to implement a dialogue system for Uyghur using the collected dialogue corpus.

The outline of this paper is as follows: in Sect. 2 we introduce our data-collection process, including the design of our game with a purpose; in the concluding section, we describe our future answer-type classification task and the formal analysis of the dialogue corpus and sketch an initial plan for implementing a dialogue system for Uyghur.

2 Dialogue Corpus Construction

As mentioned in Sect. 1, building a dialogue corpus for low resource languages is a fundamental step in the development of a dialogue system for such languages. As for the language we are concerned with in this project, Uyghur, it is one of the Turkic languages spoken by around 20 million people around the world, including 12–15 million people in Xinjiang Uyghur Autonomous Region in China, as well as in Kazakhstan, Kyrgyzstan, Uzbekistan, and in other diaspora communities around the world. The typical word order in an Uyghur sentence is Subject-Object-Verb(SOV), for instance, *"Men Uyghurche Oquymen. ('I Uyghur study.')"* Uyghur is an agglutinative language in which suffixes denoting person, number, case, mood, etc. are usually all attached to one word stem, for example, *"öyingzge" ('home-your-to')*. Besides, Uyghur has vowel and consonant harmony when suffixes and other elements are attached. Overall, Uyghur has very different linguistic characteristics than most of the well-studied languages such as English, French, Chinese, etc.

There is one open source Uyghur speech database, THUYG-20 built by Tsinghua University and Xinjiang University [1]. However, this corpus is a database of voice recordings of literature, newspaper articles, and also other books in different genres, and does not include any conversational dialogue. Thus, this corpus is not suitable for developing a Uyghur dialogue system.

Given this, we decided to compile an Uyghur dialogue corpus which consists of naturally generated conversational dialogues by native Uyghur speakers. Since accessing the Uyghur region for conducting field work is not possible at present, we need to design a new method for collecting conversational dialogues from native Uyghur speakers in diaspora.

2.1 Design a Game with a Purpose

Our basic idea, following much past work [2, 3], is to use a "game with a purpose" (GWAP) whose players will contribute significantly large amounts of dialogue data. GWAPs have also been used for text data for NLP problems, for instance the *Phrase Detectives* [4] for creating anaphorically annotated resources, and also the *Puzzle Racer* and the *Ka-Boom!* GWAPs for word sense disambiguation. Concretely, we plan to design and create a Massively multiplayer online role-playing game (MMORPG), using the **RPG Maker MV Game Engine**[1]. We will create a virtual world in our GWAP, so every player of the game is the citizen of that world. Virtual worlds [5] are computer systems or applications which imitate a real environment. They can be simultaneously affected by an enormous numbers of players, and can exist and develop internally. Thus, virtual worlds are said to be *shared* or *multi-user*, and also *persistent* [5].

Since the Uyghur diaspora are the primary target audience of this game, we would like to attract more Uyghurs to participate through designing the game to their taste. Therefore, we will use some famous Uyghur fairy tales as narratives of the game and also use the original names in these stories as some characters' names in the game. Since most of the Uyghurs are familiar with those fairy tales, it will hopefully be intriguing for them to play the game. In addition, we will develop the game in English and in Uyghur, so Uyghur can also be the language of instruction of the game. To the best of our knowledge, there is almost no game which has Uyghur as the instruction language. Having a game in their mother tongue would be novel and interesting for the target audience, so hopefully there will be more people, especially people who are not familiar with the world's main languages, who will give the game a try. Most importantly, the Uyghur diaspora are well aware that their language is under threat and they have to make a great effort to keep it alive. Thus, participating in such a game with a scientific purpose will potentially impress members of this community.

We have come up with an idea which encourages or in a sense "forces" the players to have a discussion on various topics with other players during the game. Players will be given several topics to choose, or sometimes will be randomly assigned to a specific topic, and they will have a free chat according to the instructions within a time limit. Players must also follow certain specified rules in the virtual world, and failing that, be punished. Here are the initial design of scenarios for various tasks in the game:

- **Role-playing:** in this task, you will be role playing one of the characters in the following story: There was a severe public attack yesterday on the main street of your city. The police has successfully arrested one of the assailants, who is currently being interrogated.
  - If you are the police: you should ask as many as useful questions from the suspect, and try to let him admit his crime, and also force him to tell out his accomplices;

[1] https://store.steampowered.com/app/363890/RPG_Maker_MV/.

  - If you are the criminal suspect: you should try your best to deny your crime, and make the police believe that you are innocent.

- **Planning a given task:** you are invited to participate in a real-life TV show, and you are paired up with a stranger (who is also here to participate the TV show). Your task is to plan a trip together to a totally unfamiliar place. The two of you should work together on planning the entire two week trip. The trip is self-funded so you may want to discuss your financial situation and how to arrange the budget for the trip. Since you will travelling together for the entire two weeks you should start by getting to know your partner well, including his/her basic information, family situation, hobbies etc.

- **Direction giving:** in this task you will be chatting with your partner in order to find out how to get to his/her current address. You should pay close attention to the details and draw a travel plan to your partners's place.
This task should be carried out in two rounds, with each of you playing both roles.

- **Real open discussions:** in this task, you and your partner/partners in the chatroom should freely and openly discuss a topic, you may discuss a currently occurring event from around the world, or news, politics, comedies, education, or indeed anything you may interested in. During the discussion, you should ask each other various questions about the topic.

- **Future ideal society:** in this task you and your partner/partners in the chatroom should discuss the ideal future society you want to live in. You should tell your partner how your ideal future society may like, and your partner should ask questions about that ideal society. You can talk about the social system, education, medical, transportation, and any other aspect of that ideal society.

- **Interviewing:** in this task, you will be role-playing an interviewer or an interviewee.

  - If you are the interviewer, you should ask various questions of the person you are interviewing, including basic information, private information, their current mission, their opinion about some topics, or even their further plans.
  - If you are the interviewee (you will role-play one of the famous person randomly assigned to you from our list), you may choose quite freely how to respond, you may want to answer correctly, or lie to the interviewer, you can refuse to answer or change the topic.

- **Guessing other person's current location:** in this task, you will guess the current location of your partner according to the description of their surrounding environment. You can also ask some questions to verify your partner's location, such as, *'Is there a desk nearby?'*. You should ask as many questions as you need to correctly guess the current location of your partner.

2.2 Data Collection

All conversational tasks above will be take place in the chatroom system which is implemented in the game. Furthermore, in order to collect the conversational data in a more efficient way, we will adapt the Dialogue Experimental Toolkit (DiET)[2] [6] to be usable on the internet and link it with the chatroom system of our game. DiET is in its original form a text-based chat-tool which allows utterances of particular types to be artificially introduced into natural dialogues in a controlled and synchronous manner, without the knowledge of the dialogue participants. DiET has lead to novel findings about dialogue interaction [6, 7]. DiET's ability to surreptitiously insert turns will be important in trying to control for problems of data sparseness. This methodology has been approved by several ethics committees (e.g.., at Stanford), as long as the subjects are debriefed after participation, similarly for online GWAPs[3]. Integrating the chatroom system of our GWAP with DiET helps us improve the data collection accuracy and efficiency.

3 Conclusions and Future Work

In this paper, we present our plan and initial design for constructing a dialogue corpus for the low resource language Uyghur, with the ultimate goal of developing a dialogue system for Uyghur. We plan to design and create a Massively multiplayer online role-playing game (MMORPG), using the **RPG Maker MV Game Engine**. We also introduced our initial design of different task scenarios with the aim of collecting various types of naturally generated questions and answers from native Uyghur speakers. Our method and the design of the game can be used for other low resource languages in order to collect a large amount of dialogue data, which can enable the implementation of dialogue system for such languages. We will then integrate the chat system of the game with the DiET dialogue tool in order to accelerate and improve the data collection process.

After constructing the Uyghur dialogue corpus, we will conduct a detailed corpus study on the questions and answers, and will try to characterize the response space of questions in Uyghur. A question can be responded to in many ways. In the work [8], authors studied one significant component of the response space of questions, which is responding to a question with a question. They conducted their studies on the British National Corpus and three other more genre-specific corpora in English, and characterised the range of question responses into 7 classes (Clarification requests, dependent questions, questions about the form of the answer, requests for underlying motivation, indirect question responses, and two classes of evasion questions), and showed how to model these 7 classes within the framework of Conversation Oriented

[2] https://dialoguetoolkit.github.io/chattool/.

[3] https://www.scienceathome.org/legal/game-consent-form/skill-lab-science-detective-terms-of-service/.

Semantics (KoS in short), which is based on the formalism of Type Theory with Records (TTR) [9]. Another previous related work [10] offered a characterization of the entire response space for English and Polish. Inspired by these works, we aim to take the challenge of characterizing the response space of questions and develop a taxonomy for question response in Uyghur. Ultimately, we aim to provide the formal modelling of various response classes within the framework of KoS.

Given a detailed taxonomy for question responses in Uyghur and also the formal modelling of various response classes within the framework of KoS, this will provide us a good starting point for developing a dialogue system for Uyghur, using [11, 12] for dialogue management.

Acknowledgments We acknowledge the support by a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001.

References

1. Rouzi A, Yin S, Zhang Z, Wang D, Hamdulla A, Zheng F (2017) Thuyg-20: a free uyghur speech database. J Tsinghua Univ (Sci Technol) 57(2):182–187
2. von Ahn L (2006) Games with a purpose. Computer 39(6):92–94
3. Ho C-J, Chang T-H, Lee J-C, Hsu JY, Chen K-T (2009) Kisskissban: a competitive human computation game for image annotation. In Proceedings of the ACM SIGKDD workshop on human computation, pp 11–14
4. Poesio M, Chamberlain J, Kruschwitz U, Robaldo L, Ducceschi L (2013) Phrase detectives: utilizing collective intelligence for internet-scale language resource creation. ACM Trans Interact Intell Syst (TiiS) 3(1):1–44
5. Bartle RA (2004) Designing virtual worlds. New Riders
6. Healey PGT, Purver M, King J, Ginzburg J, Mills G (2003) Experimenting with clarification in dialogue. In: Alterman R, Kirsh D (eds) Proceedings of the 25th annual conference of the cognitive science society, LEA, Mahwah, N.J., pp 539–544
7. Eshghi A, Healey PGT (2016) Collective contexts in conversation: grounding by proxy. Cogn Sci 40(2):299–324
8. Łupkowski P, Ginzburg J (2016) Query responses. J Lang Modell 4(2):245–292
9. Ginzburg J (2012) The interactive stance. Oxford University Press, UK
10. Ginzburg J, Yusupujiang Z, Li C, Ren K, Łupkowski P (2019) Characterizing the response space of questions: a corpus study for English and polish. In Proceedings of the 20th annual SIGdial meeting on discourse and dialogue, pp 320–330
11. Larsson S, Berman A (2016) Domain-specific and general syntax and semantics in the talkamatic dialogue manager. Empir Issues Syntax Sem 11:91–110
12. Maraev V, Ginzburg J, Larsson S, Tian Y, Bernardy J-P (2018) Towards KOS/TTR-based proof-theoretic dialogue management. In: Proceedings of SemDial

Conversational Systems Research in Spain: A Scientometric Approach

David Griol and Zoraida Callejas

Abstract The aim of this paper is to present a preliminary scientometric study of the area of conversational systems in Spain. In order to do so, we have used the Web of Science database to retrieve the papers in the area using a comprehensive list of keywords and considering those papers with at least one author with Spanish affiliation. Our results present an overview of the main topics, authors and institutions involved in conversational system research and show the good status of Spanish research in this discipline.

1 Introduction

Scientometrics is a research area that studies science, technology and innovation using quantitative approaches, e.g. by means of statistical mathematical methods [1]. This is usually achieved by studying peer-reviewed scientific publications and other related documents.

Scientometric indicators are very useful to provide evidence of scientific outcomes, achieve a general perspective of a field of research or even infer hot topics and emerging research fronts [6].

For example, the OECD has a section devoted to scientometric indicators to standardise, collect, report and analyse a wide range of science, technology and innovation activities[1] [4], including trends in scientific production (e.g. top cited publications, recent trends in scientific excellence, quantity and quality of scientific production, citation impact), scientific collaboration (e.g. patterns of international collaboration), and insights on scientific production and funding.

D. Griol (✉) · Z. Callejas
University of Granada, Periodista Daniel Saucedo Aranda sn, Granada, Spain
e-mail: dgriol@ugr.es

Z. Callejas
e-mail: zoraida@ugr.es

[1] https://www.oecd.org/sti/inno/scientometrics.htm.

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_31

The general aim of the Plan for the Advancement of Language Technology[2] of the Digital Agenda for Spain is to promote the development of natural language processing, conversational interfaces and machine translation in Spanish and Spain's co-official languages. The Plan has the following specific goals:

- Increasing the amount, quality and availability of linguistic infrastructure in Spanish and in Spain's co-official languages.
- Fostering the language industry by promoting knowledge transfer from the research field to that industry and the internationalisation of companies and institutions in the sector; improving the reach of current projects.
- Improving the quality and capacity of public services, integrating natural language processing and machine translation technologies, simultaneously driving demand. Supporting creation, standardisation and distribution of language resources created by the management activities performed by the public administrations.

This paper contributes an initial scientometric analysis of research in conversational systems in Spain. In the scope of the Plan for the Advancement of Language Technology, the term conversational system is used to refer to computer applications that can hold a conversation in natural language [3]. This approach emphasises two key ideas. First, the use of natural language (written or spoken) as a support for communication. Second, that the system has conversational abilities and is able to interact in a series of turns in which it exchanges messages with the user in natural language.

There is a wide terminology related to conversational systems. The term chatbot is often used to describe systems that usually interact in text mode (through a chat), with which users can talk about any topic; while conversational and dialogue systems are sometimes used as synonyms. There is however a subtle distinction between a conversational system and a conversational agent. Usually, the term agent is used when the system appears to the user as an identifiable interlocutor. This usually happens with personified conversational agents, which have a physical appearance through avatars or other graphic representations; or robots, for which the term conversational robot is also used [2].

The development of a conversational system involves a large number of components that replicate the functions that human beings carry out to maintain a dialogue: recognise the sequence of words mentioned by the user (automatic speech recognition); extract the meaning of these words, that is, understand the information pieces that are useful in the domain of the system (natural language understanding); perform operations to access services, data repositories or other system resources, in which the information requested by the user is stored or the operations that they want to know are recorded; decide the next system action after each user request (dialogue management); reproduce a spoken message that informs the user about the action the system has selected (text-to-speech synthesis), and that it may be enriched by additional information pieces in other modalities [5].

[2] https://www.plantl.gob.es/.

There is a vast number of applications and tasks in which conversational systems can be applied, for instance: systems that provide information, healthcare systems, electronic banking services, tourism, industrial environments, applications accessible from vehicles, systems that facilitate access to information for people with disabilities, tele-education applications, apps and personal assistants for mobile devices; access to services and control of machines using the telephone, home interaction and home automation control, interaction with robots and wearable devices, systems able of recognizing gestures and users' emotional states, etc.

For this study, we have focused the search of relevant pieces of work in different aspects that include the main terminology and related terms, technologies and components that are unequivocally related to conversational systems (e.g. dialogue management, but not speech recognition), functional systems and specific technology and the main tasks and application domains.

2 Method and Sample

Our contribution is based on primary studies (articles, books and conference papers) indexed in Web of Science (WOS). To retrieve the papers related to conversational systems, we created a list of keywords that, from our point of view, could indicate that the research described falls into the area of conversational systems. The list of 176 keywords contains the main terminology, modules, technologies, approaches, domains and even names of already existing systems and can be found in the Annex.

We performed an advanced search in the Web of Science Core collection indicating that the topic (TS) can contain any of the keywords in the list and the country of affiliation (CU) of at least one author must be Spain (see the exact search query in the Annex). The publication timespan was "all years" (from 1900 to 2020), the date of recovery was 1st January 2020. With this search we retrieved a sample of 646 papers ($SAMPLE\_ES$).

For comparison purposes, we performed a similar query in which we did not restrict the country to Spain (we allowed publications from any country). We obtained a sample of 13,775 papers ($SAMPLE\_INT$).

It is important to note that when restricting to science only indexes (i.e. excluding social science indexes), the results are less significant, as relevant categories such as linguistics, educational research and behavioural sciences are missed. For example, the journal of the Spanish National Society for Language Processing (Procesamiento del Lenguaje Natural) is indexed under WOS category "Linguistics", and so it would not be considered if only science and technology sources are considered.

3 Discussion of Results

The following subsections present an overview of the main areas, publishers and document types, authors and institutions involved in the items of conversational system research retrieved.

3.1 Areas

The top 20 Web of Science areas with more papers published from the Spanish sample are shown in Fig. 1. As can be observed 67.49% (436) of the papers are published in the *Computer Science* area, and the top 5 areas are related to computer science, engineering, communications and robotic and automation. However, also in the top 10 we find other areas related to linguistics, education research and even pharmacology.

It is interesting to see areas related to pharmacology (43 papers), oncology (35 papers), geriatrics (32 papers) and general internal medicine (17 papers). When taking a closer look into the papers that fall into these categories, we find that they do not describe advances in the development of the conversational technology, but rather this technology is used to serve the health science research. For example, interactive voice response systems are used to randomly assign patients to different conditions (e.g. different drugs or drug vs. placebo) and administrate questionnaires.

In other areas (e.g. education and educational research) the conversational systems have a more prominent role, e.g. as key tools for second language learning, pedagogical agents or counsellors.

The results are similar to the international scenario (see Fig. 2), in which 60.12% of the papers are in the *Computer Science* area. The top 3 are the same in both

Fig. 1 WOS areas of papers from $SAMPLE\_ES$

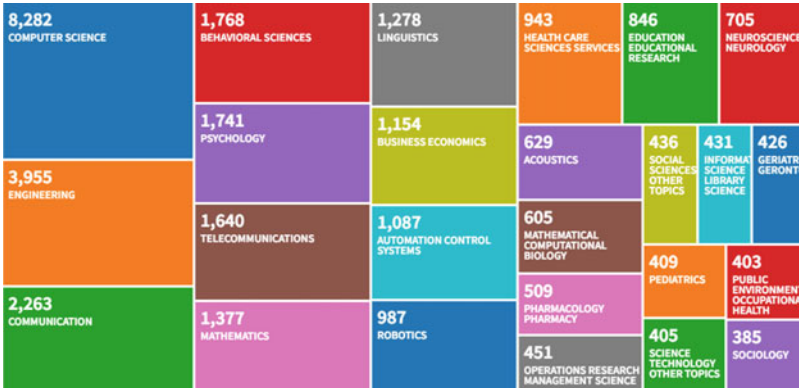


Fig. 2 WOS areas of papers from $SAMPLE\_INT$

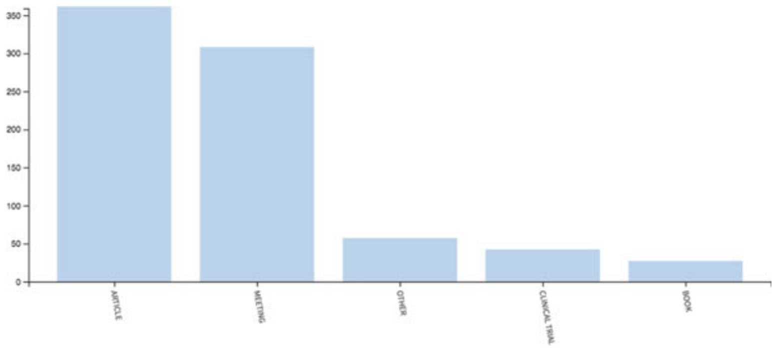


Fig. 3 Document types of papers in $SAMPLE\_ES$

cases, and the main categories appear in the top positions of the ranking in both cases. However, it is interesting to note that internationally Psychology, Behavioural Sciences and Economics play a more relevant role than in the Spanish scenario. In the international scenario also educational research appears at a more discrete position compared to the Spanish scenario.

3.2 Publishers and Document Types

Regarding the type of publications, as shown in Fig. 3, 55.57% are articles in journals and 47.37% publications in conference proceedings. Only 25 books or book chapters (3.87%) resulted from our query.

The percentages for all international publications ($SAMPLE\_INT$) are very similar (54.08% articles, 47.53% proceedings). This shows that Spanish authors publish article vs. proceeding papers at the same balance as international authors.

The sources with more publications in the Spanish list are Lectures Notes in Computer Science and Lecture Notes in Artificial Intelligence, which account for 18.73% of the records. It is very important to highlight that the third source (6.81% of the papers) is the journal Procesamiento del Lenguaje Natural, from the Spanish Society for Natural Language Processing (SEPLN).

Among the journals specialized in the area, the most representative are: Procesamiento del Lenguaje Natural, Speech Communication, Expert Systems with Applications, Ambient Intelligence and Smart Environments and Computer Speech and Language.

From conference series the ones with more papers from the $SAMPLE\_ES$ sample are Advances in Intelligent and Soft Computing, Advances in Intelligent Systems and Computing, Communications in Computer and Information Science. Regarding specific conferences, the most relevant are Text Speech and Dialogue (TSD) and Interspeech. It is also worth noticing the International Conference on Human-Computer Interaction (INTERACT), the Social Robotics International Conference (ICSR), the Artificial Intelligence in Education Conference (AIED), and the International Intelligent Environments (IE). The conferences organized in Spain with more publications from the sample are IberSpeech and the SEPLN annual conference. Also the workshop Future and Emergent Trends in Language Technology had an important impact.

Internationally, in the $SAMPLE\_INT$ sample, there also appear in the top 10 positions the International Conference on Acoustics Speech and Signal Processing (ICASSP) and the IEEE Workshop on Spoken Language Technology (SLT). The Journal of Pragmatics is in the 4th position.

3.3 Authors and Affiliations

From the 21 authors with more than 15 papers (Fig. 4) in $SAMPLE\_ES$, we can see there are strong publication records from authors who have had or have affiliations in Universidad Carlos III de Madrid (D. Griol, J.M. Molina, A. Sanchis), Universidad de Granada (D. Griol, R. López-Cózar, Z. Callejas, G. Espejo, N. Ábalos), Universidad Politécnica de Valencia (D. Griol, E. Sanchis, E. Segarra, L.F. Hurtado), Universidad Politécnica de Madrid (J. Ferreiros) and Universidad del País Vasco (M.I. Torres).

There exist more than 3k different affiliations in $SAMPLE\_ES$, from which only the universities mentioned in the previous paragraph produce more than 5% of the publications, being the most productive Universidad Carlos III de Madrid (22%), Universidad de Granada (17%) and Universidad Politécnica de Valencia (11%).

In the list of top-10 affiliations (see Fig. 5) there also appear Universidad Politécnica de Madrid, Universidad de Sevilla, Universidad Politécnica de Catalunya, Universidad de Barcelona and Universidad de Zaragoza.

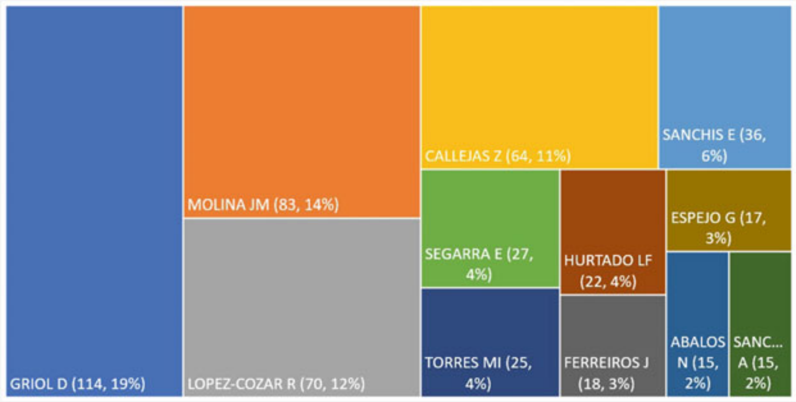


Fig. 4 Most prolific Spanish authors in $SAMPLE\_ES$

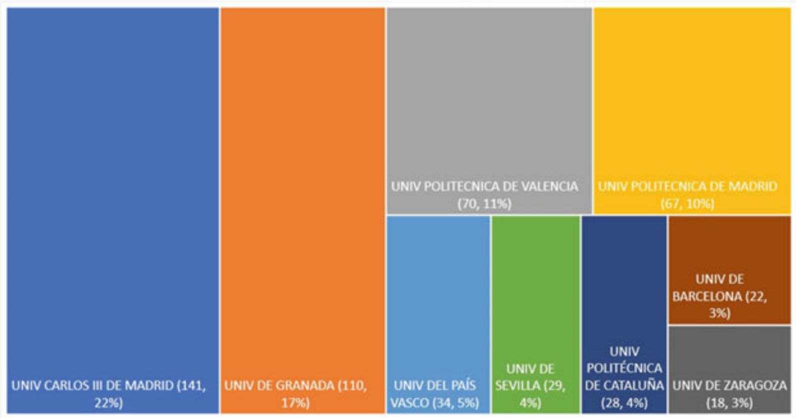


Fig. 5 Most prolific Spanish institutions in $SAMPLE\_ES$

In $SAMPLE\_ES$ there are authors with affiliations from 75 countries (see Fig. 6), which shows that Spanish authors in this field work in international teams with co-authors from other countries. The most representative countries are USA, UK, Italy, Germany and France.

As can be observed in Fig. 7, Spain holds the 7th position at the international level (sample $SAMPLE\_INT$), with 4.70% of the publications. The most prolific authors are form USA, followed by UK, Germany and China.

Regarding individual authors, Fig. 8 shows that 5 Spanish authors are among the 25 most productive internationally, and 4 of them in the first positions.

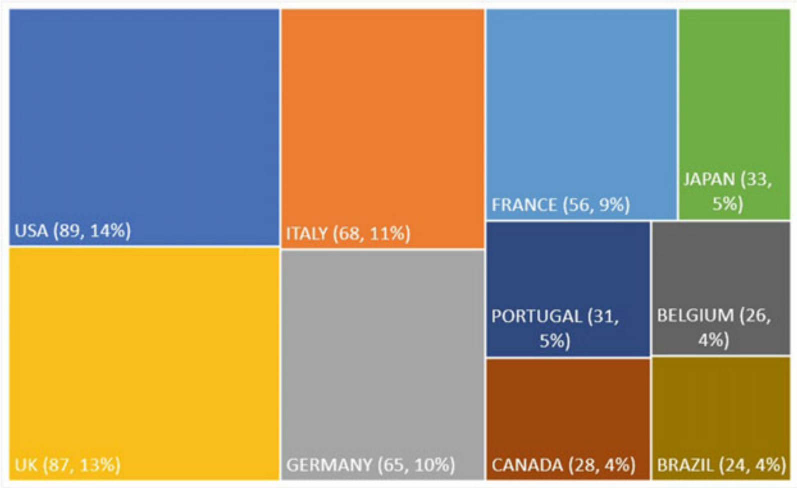


Fig. 6 International collaborations with Spanish authors

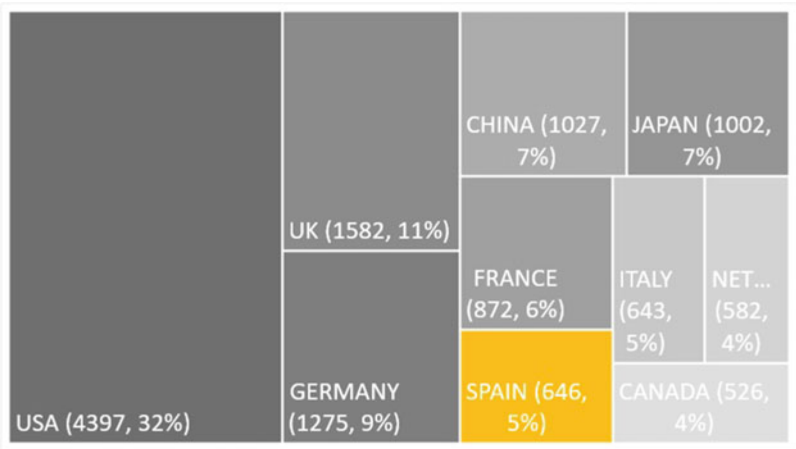


Fig. 7 Most productive countries in the field

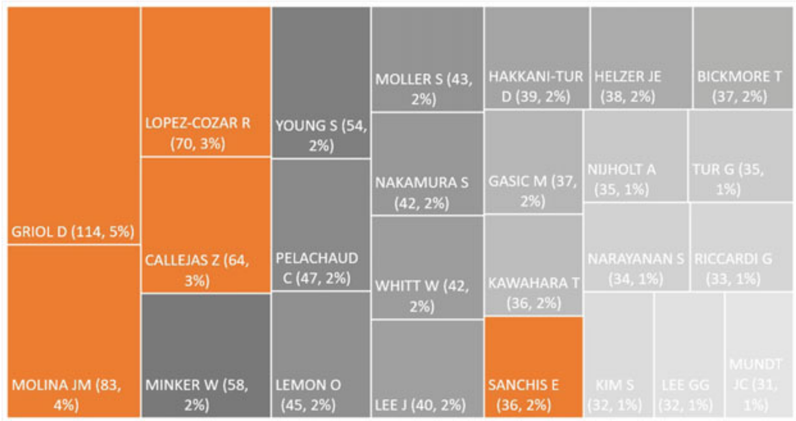


Fig. 8 Most productive authors in the field (25 best in $SAMPLE\_INT$)

4 Conclusions and Future Work

This paper presents a preliminary scientometric study of the area of conversational systems in Spain. In order to do so, we have used the Web of Science database and retrieved the papers in the area using a comprehensive list of keywords and considering those papers with at least one author with Spanish affiliation.

Our results present an overview of the main areas, authors and institutions involved in conversational system research and show that the status of Spanish research in this discipline is significant and Spanish researchers occupy prominent positions in the rankings generated.

For future work we plan to extend the study considering co-citation, co-occurrence and altmetric indicators. We will also investigate the main collaboration networks between Spanish researchers and between Spanish and international researchers. For this new piece of work we will consider additional databases, such as Scopus, DBLP and Semantic Scholar.

Acknowledgements We would like to thank the support of the Spanish Plan for the Advancement of Language Technology (https://www.plantl.gob.es/) and the Secretaría de Estado para la Digitalización e Inteligencia Artificial (SEDIA). This research has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR project: https://menhir-project.eu).

Annex

TS = (Acobot OR Airbud OR "Amazon Alexa" OR "Amazon Echo" OR "Amazon Lex" OR "Ambit AI" OR AmplifyReach OR "Android speech APIs" OR Apex-Chat OR Api.Ai OR Articbot OR AskSid OR Avaamo OR "BOTNATION AI" OR Boost.AI OR BotArtisanz OR Botkit OR Botsify OR BotXO OR "call center" OR "chat with dialogue" OR "chat dialogue" OR "chat with dialog" OR "chat dialog" OR chatbot OR ChatFuel OR "chatter bot" OR Cognigy.AI OR Conversica OR Cortana OR "conversational agent" OR "conversational AI" OR "conversational Artificial Intelligence" OR "conversational assistant" OR "conversational app" OR "conversational interface" OR "conversational applications" OR "conversational platform" OR "conversational robot" OR "conversational system" OR "conversational toys" OR "conversational web interface" OR "corpus-based dialog management" OR "corpus-based dialogue management" OR DAMSL OR "dialog bot" OR "dialogue bot" OR "dialog act" OR "dialogue act" OR "dialogue act modeling" OR "dialogue act classification" OR "dialog act recognition" OR "dialogue act recognition" OR "dialog act theory" OR "dialogue act theory" OR "dialog agent" OR "dialogue agent" OR "dialog assistant" OR "dialogue assistant" OR "dialog system" OR "dialogue system" OR "dialog manager" OR "dialogue manager" OR "dialog management" OR "dialogue management" OR Dialogbot.ai OR Dialogflow OR DialogFlow OR "embodied conversational agent" OR "Embodied Conversational Agent" OR "end-to-end goal-oriented dialog" OR ENCollect OR Faqbot OR "finite state dialog" OR "Flow XO" OR Flow.ai OR "form interpretation algorithm" OR "frame-based dialog" OR "GALAXY architecture" OR "Google assistant" OR "Google Home" OR "Google Now" OR hellomybot OR HeyMojo OR Houndify OR "HubSpot's Chatbot Builder" OR Ideta OR InfinitAI OR "Information State Approach" OR "intelligent personal assistant" OR IntelliTicks OR "interactive speech systems" OR "interactive systems for speech" OR "Interactive Voice Response" OR "IOX bot" OR JeffreyAI OR Joey OR Kommunicate OR Landbot.io OR LiveEngage OR "mixed-initiative dialog" OR "mixed-initiative dialogue" OR MobileMonkey OR "multimodal conversational interfaces" OR "multimodal conversations" OR "multimodal dialog" OR "multimodal dialogue" OR "multimodal systems" OR "multimodal human computer interface" OR "multimodal interaction" OR "multimodal interfaces" OR "multilingual systems" OR nmodes OR Odus OR "Olympus architecture" OR "oral interface" OR PandoraBots OR "PARADISE project" OR "PARADISE paradigm" OR "PARADISE evaluation" OR PolyAI OR Polyfins OR "POMDP-based dialog" OR "POMDP-based dialogue" OR "POMDP dialog" OR "POMDP dialog management" OR "POMDP dialogue" OR "POMDP dialogue management" OR "question-answering from speech" OR "Rasa Stack" OR Ravenclaw OR Recast.AI OR "rule-based dialog management" OR SalesboxAI OR (Siri AND Apple) OR "slot filling dialog" OR "smart speakers" OR Smartloop OR SMARTKOM OR SmatBot OR Solvemate OR "spoken dialog system" OR "spoken dialogue system" OR "spoken dialog systems" OR "spoken dialogue systems" OR "spoken humancomputer interaction" OR "spoken question answering" OR "spoken virtual agents" OR "spoken virtual human toolkit" OR "statistical dialog management" OR Streebo OR supertext.ai OR Surbo OR "system-directed dialog" OR "technology of conversation" OR "Teneo" OR "Tilde.AI" OR "turn taking" OR "turn-taking" OR "Twyla" OR "vernacular.ai" OR "virtual personal assistant" OR "voice actions" OR "voice interface" OR "voice user interface" OR voicebot OR Voiceflow OR VoiceXML OR VOIQ OR Voxeo OR VUI OR "Web speech" OR Yekaliva OR Zynoviq) AND CU = Spain.

References

1. Leydesdorff L, Milojević S (2015) Scientometrics. In: Wright JD (ed) International encyclopedia of the social & behavioral sciences, 2nd edn. Elsevier, Oxford, pp 322–327. https://doi.org/10.1016/B978-0-08-097086-8.85030-8
2. McTear M, Callejas Z, Griol D (2016) The conversational interface: talking to smart devices. Springer. https://doi.org/10.1007/978-3-319-32967-3
3. McTear MF (2016) The rise of the conversational interface: a new kid on the block? In: Proceedings of the second international workshop future and emerging trends in language technology. Machine learning and big data (FETLT), Seville, Spain, pp 38–49
4. OECD, SCImago Research Group (CSIC) (2016) Compendium of bibliometric science indicators. Technical report, OECD, Paris. http://oe.cd/scientometrics
5. Quesada JF, Callejas Z, Griol D (2019) Informe sobre sistemas conversacionales multimodales y multilingües. Tecnologías y arquitecturas para el desarrollo de asistentes virtuales, sistemas de diálogo y otros interfaces conversacionales. Technical report, Plan de Impulso de las Tecnologías del Lenguaje. https://www.plantl.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/sistemas-conversacionales.aspx
6. Úbeda-Sánchez A, Fernández-Cano A, Callejas Z (2019) Inferring hot topics and emerging educational research fronts. Horizon 27(2):125–134

Author Index

A
Aceta, Cristina, 243
Adiba, Amalia Istiqlali, 129
Aguiló, Jordi, 193
Ali Mehenni, Hugues, 257
Álvarez, Aitor, 243
Anantha, Raviteja, 87
Arzelus, Haritz, 243
Azpeitia, Andoni, 289

B
Bailón, Raquel, 193
Banchs, Rafael E., 53
Benítez-Guijarro, Antonio, 201
Bertero, Dario, 129
Bond, Raymond, 201, 369
Booth, Frederick, 201, 369
Brixey, Jacqueline, 41

C
Callejas, Zoraida, 201, 393
Cartuyvels, Ruben, 363
Chappidi, Srinivas, 87

D
D'Haro, Luis Fernando, 15, 53, 71
Dawoodi, William, 87
de la Cámara, Concepción, 193
de los Riscos, Agustín Manuel, 15
del Pozo, Arantza, 289
Develasco, Mikel, 183
deVelasco, Mikel, 369
Devillers, Laurence, 257
Duong, Hong Ngan, 31

E
Ennis, Edel, 201
Esposito, Anna, 201

F
Fernández, Izaskun, 243
Fernández-Bhogal, Mikel D., 289
Feustel, Isabel, 379
Friedrichs, Thomas, 53
Fuchs, Michael, 369
Fujimura, Hiroshi, 117
Funakoshi, Kotaro, 319

G
García, Esther, 193
García-Sardiña, Laura, 289
Georgila, Kallirroi, 145
Ginzburg, Jonathan, 387
González-Docasal, Ander, 243
Gordon, Carla, 145
Griol, David, 393

H
Haro, Josep Maria, 193
Heid, Ulrich, 349
Hemmje, Matthias, 369
Homma, Takeshi, 129

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7

I
Inoue, Koji, 3, 303
Isonishi, Sota, 303
Iwata, Kenji, 117

J
Jokinen, Kristiina, 213, 271
Justo, Raquel, 183

K
Kale, Mihir, 279
Kawahara, Tatsuya, 3, 303
Kildal, Johan, 243
Kittel, Patricia, 349
Klakow, Dietrich, 163
Kobayashi, Yuka, 117
Kobylyanskaya, Sofiya, 257
Kontaxis, Spyridon, 193
Kraus, Matthias, 201

L
Lala, Divesh, 303
Lam, Darryl, 319
Laperrière, Jean-Sébastien, 319
Letaifa, Leila Ben, 183
Li, Haizhou, 53
Lim, Ting Jen, 31
López-Zorrilla, Asier, 183

M
Martínez, Carmen, 193
Matía, Fernando, 71
Mc Kevitt, Paul, 369
McConvey, Gavin, 201
McTear, Michael, 201
Miehle, Juliana, 103, 379
Minker, Wolfgang, 103, 349, 379
Moens, Marie-Francine, 363
Mulvenna, Maurice, 201, 369
Mundhra, Ria, 31

N
Nagamatsu, Kenji, 129
Nakamura, Satoshi, 227
Nakamura, Shizuka, 3
Niculescu, Andreea I., 31
Nishimura, Takuichi, 213
Nishimura, Yuta, 227
Niwa, Shin-ichi, 213

O
Olaso, Javier Mikel, 183
Ortega, Alfonso, 193
Oshiyama, Chiaki, 213

P
Petukhova, Volha, 163
Posadas-de Miguel, Mar, 193
Potts, Courtney, 201
Pragst, Louisa, 201

R
Rodríguez-Cantelar, Mario, 71
Romero, David, 339

S
Salamea, Christian, 339
Schmidt, Maria, 349
Serras, Manex, 289
Sharifullaeva, Firuza, 163
Siddhant, Aditya, 279
Siddi, Sara, 193
Spinks, Graham, 363
Stier, Daniela, 349
Sumiyoshi, Takashi, 129

T
Takanashi, Katsuya, 3, 303
Torres, M. Inés, 183
Traum, David, 41, 145
Turkington, Robin, 201, 369

U
Ultes, Stefan, 103, 379

V
Vasilescu, Ioana, 257
Vázquez, Alain, 183
Vu, Binh, 369

W
Wagner, Nicolas, 103, 201
Wakimoto, Kohei, 227

Y
Yamamoto, Kenta, 3
Yanov, Volodymyr, 145
Yeo, Kheng Hui, 31
Yoshida, Takami, 117
Yoshino, Koichiro, 227
Yusupujiang, Zulipiye, 387

Z
Zhang, Chen, 53
Zheng, Huiru, 201