Identification of Research Data References Based on Citation Contexts

Tomoki Ikoma[1](✉) and Shigeki Matsubara[2]

[1] Graduate School of Informatics, Nagoya University, Nagoya, Japan
ikoma.tomoki@h.mbox.nagoya-u.ac.jp
[2] Information and Communications, Nagoya University, Nagoya, Japan

Abstract. In this paper, a method for the automatic identification of research data references in publications is proposed for automatically generating research data repositories. The International Conference on Language Resources and Evaluation (LREC) requires authors to list research data references separately from other publication references. The goal of our research is to automate the discrimination process. We investigated the reference lists in LREC papers and the citation contexts to find characteristic features that are useful for identifying research data references. We confirmed that key phrases appeared in the citation contexts and the bibliographical elements in the reference lists. Our proposed method uses the presence or absence of key phrases to identify research data references. Experiments on LREC proceedings papers proved the effectiveness of using key phrases in the citation context.

Keywords: Research data · Text classification · Scholarly papers

1 Introduction

The demand for the share and reuse of research data has significantly increased with the spread of open science. In the field of natural language processing, organizations such as Linguistic Data Consortium (LDC) [3], International Standard Language Resource Numbe (ISLRN) [5,7], and Common Language Resources and Technology Infrastructure (CLARIN) [11] have created data repositories. However, these repositories do not thoroughly collect research data, because they are manually maintained.

To enhance the repositories, automatic construction and updates are mandatory. This may be achieved by utilizing the information of research data references in scholarly papers. However, since research data references and bibliographical references are usually mixed in publications, it is required to distinguish research data references from other types of references.

In this paper, a method is proposed for the automatic identification of research data references in publications. Although bibliographical elements in the reference lists contain several useful clues, they are not always available, for example, when the provided information is incomplete. Our proposed method

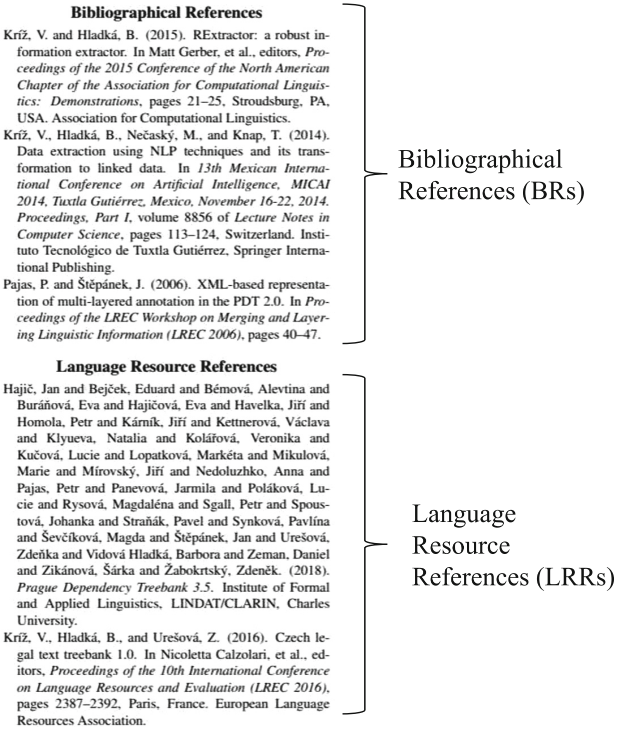
E. Ishita et al. (Eds.): ICADL 2020, LNCS 12504, pp. 149–156, 2020.
https://doi.org/10.1007/978-3-030-64452-9_13

uses key phrases extracted from citation contexts as well as bibliographical elements for classification even without sufficient clues in the bibliographical elements. Experiments on international conference proceedings proved the effectiveness of using clues derived from citation contexts.

This paper is organized as follows: in Sect. 2, we describe how authors list research data references; in Sect. 3, the characteristic features of research data references are investigated; in Sect. 4, a method is proposed for identifying research data references; finally, we describe the experiments for evaluating the proposed model in Sect. 5.

2 Research Data References in Reference Lists

While the format of the bibliography of cited publications is uniformly determined, authors often decide how to list research data references. As a result, different methods for listing research data in reference lists have been applied, such as listing publications that are related to the research data, the URL of the site where the data is available, or the user guide of the research data.



Bibliographical References

Kríž, V. and Hladká, B. (2015). RExtractor: a robust information extractor. In Matt Gerber, et al., editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kríž, V., Hladká, B., Nečaský, M., and Knap, T. (2014). Data extraction using NLP techniques and its transformation to linked data. In *13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I*, volume 8856 of *Lecture Notes in Computer Science*, pages 113–124, Switzerland. Instituto Tecnológico de Tuxtla Gutiérrez, Springer International Publishing.

Pajas, P. and Štěpánek, J. (2006). XML-based representation of multi-layered annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47.

Language Resource References

Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Hajičová, Eva and Havelka, Jiří and Homola, Petr and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Pajas, Petr and Panevová, Jarmila and Poláková, Lucie and Rysová, Magdaléna and Sgall, Petr and Spoustová, Johanka and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Štěpánek, Jan and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. (2018). *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University.

Kríž, V., Hladká, B., and Urešová, Z. (2016). Czech legal text treebank 1.0. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2387–2392, Paris, France. European Language Resources Association.

Bibliographical References (BRs)
Language Resource References (LRRs)

Fig. 1. Example of reference lists in LREC proceedings

Distinguishing research data references from other references is highly beneficial for readers that are interested in using the research data. Since 2016, the International Conference on Language Resources and Evaluation (LREC) has required authors to list references in two divisions, as shown in Fig. 1: Bibliographical References (BRs) for references to publications and Language Resource References (LRRs) for references to language resources (research data in the natural language processing field) [2]. This requirement contributes to the organization of information on research data cited in publications, and the spread of such rules can facilitate the utilization of research data. The goal of our work is to automate the discrimination process for the generation of research data repositories from academic papers [6,8,10].

3 Investigation of Research Data References

We investigated the characteristic features of research data references. In the proceedings of LREC 2016, 2018, and 2020 [4], 416 papers cited language resources. We collected these papers and randomly split them into 10 blocks (blocks 0–9) with equal size as the dataset for our research.

We investigated block 8 of the dataset (investigation data), and the subjects of our investigation were as follows:

Bibliographical elements: Information, such as the title, name of the journal, and where the cited item is available, that was listed in the citation list.

Citation context: The title of the section and the sentences in the text that contain the citation tag.

Table 1. Key phrases in bibliographic elements

Appears in	Key phrase	LRR ratio (%)	
Title	corpus, corpora, dictionary, lexicon, language resources	26.8	(37/138)
Title	data, set, bank	38.7	(36/93)
Title	annotate, construct, build	19.2	(14/73)
Title	name of languages (i.e. English, Chinese)	26.5	(31/117)
Bibliographical elements	University, institute, center	29.9	(20/67)
Bibliographical elements	proceedings, journal	8.0	(25/435)
Bibliographical elements	http(s)://, www	50.0	(29/58)
Bibliographical elements	LDC, CLARIN, ISLRN, LREC	42.7	(41/96)

The investigation data contain 963 references: 841 (87.3%) BRs and 122 (12.7%) LRRs. We extracted words and phrases that can serve as clues for distinguishing BRs and LRRs as key phrases and calculated the ratio of the LRRs to their appearances in the text (LRR ratio) for each of the key phrases.

The classification criteria for BRs and LRRs vary from author to author, as LREC's author guidelines do not define any specific rules. This investigation focuses on understanding the tendency of the classification criteria and citation methods adopted by authors.

Table 2. Key phrases in citation contexts

Appears in	Key phrase	LRR ratio (%)	
Section title	corpus, corpora	16.8	(28/167)
Section title	data, set, bank	45.3	(24/53)
Section title	method, algorithm	18.9	(7/37)
Section title	introduction, conclusion, related work	7.5	(29/387)
Section title	experiment, evaluation	6.1	(5/82)
Citation sentence	corpus, corpora, dictionary, lexicon, word embedding, word2vec, WordNet	18.9	(43/227)
Citation sentence	data, set, bank, collection	14.9	(28/188)
Citation sentence	tool, parser, library, code, repository, resource	8.3	(10/120)
Citation sentence	capitalized words	12.3	(73/594)
Citation sentence	We	11.2	(43/375)
Citation sentence	They	3.4	(2/59)
Citation sentence	(use, apply, utilize, etc.) and names of language resources	17.4	(32/184)
Citation sentence	reference tag at the top of the citation sentence	9.3	(49/528)
Citation sentence	reference tag at the end of the citation sentence	9.6	(60/626)

3.1 Key Phrases Related to Bibliographical Elements

Table 1 summarizes the key phrases extracted from bibliographical elements and the LRR ratio for each key phrase. Examples include:

Language names and language resource categories: Titles of publications on language resource construction often include language names and language resource categories, such as Corpus of Reading Comprehension Exercises in German (CREG).

URL: Bibliographical elements for language resources usually contain the URL of the language resource, whereas it mainly consists of the conference and journal name for publications.

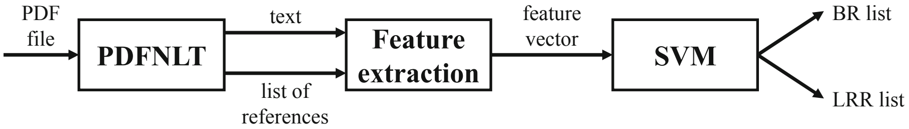
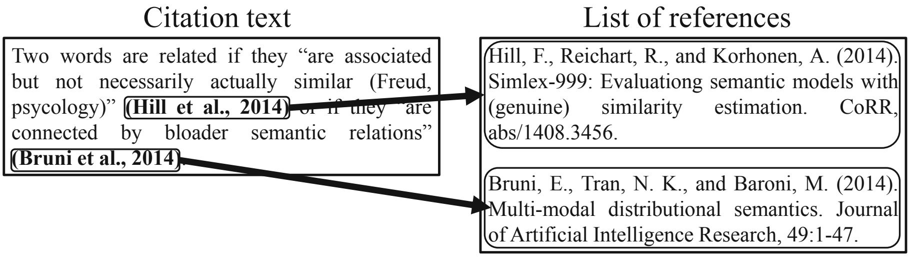


Fig. 2. Configuration of the proposed system

3.2 Key Phrases Related to Citation Contexts

Table 2 summarizes the key phrases extracted from citation contexts and the LRR ratio for each key phrase. Examples include:

Title of the section containing the citation context: The section of the experimental settings describes the language resources used in the experiment. In contrast, citations in the introduction and related work sections mainly describe the proposed ideas and preliminaries.



Two words are related if they "are associated but not necessarily actually similar (Freud, psycology)" (Hill et al., 2014) or if they "are connected by bloader semantic relations" (Bruni et al., 2014)

Hill, F., Reichart, R., and Korhonen, A. (2014). Simlex-999: Evaluationg semantic models with (genuine) similarity estimation. CoRR, abs/1408.3456.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multi-modal distributional semantics. Journal of Artificial Intelligence Research, 49:1-47.

Fig. 3. Citation text to corresponding reference list entries

Language resource categories: The citation tags for language resources often appear after the word that represents the language resource categories. For instance, CREG is cited in the sentence "Second, we provide POS and normalization annotation on top of the CREG Corpus (Merrers et al., 2011).", in which the citation tag (Merrers et al., 2011) appears after the word "Corpus".

Citation tag at the beginning of the sentence: Citation tags often serve as the subject of the sentences that cite publications by appearing at the beginning of the sentence. For instance, the sentence "Selinker (1972) coined the term interlanguage for these language variants of individual learners." describes the idea of interlanguage presented in the publication with the citation "Selinker (1972)".

Thus, the key phrases appear in the citation contexts and the bibliographical elements.

4 Method

Our method identifies research data references based on the following steps (see Fig. 2):

1. Extract the text and the reference list from the PDF file using PDFNLT [1].
2. Extract citation tags from the text and associate them with the corresponding item in the reference list (see Fig. 3).
3. Extract the features for identification from the bibliographical elements and the citation contexts for each item.
4. Classify each item as either research data or publication using support vector machine (SVM).

The method employs the presence or absence of each key phrase listed in Tables 1 and 2.

5 Experiments

5.1 Experimental Setting

We conducted experiments to evaluate the effectiveness of citation contexts for identifying research data references. We implemented the SVM classifier using the SVM module of scikit-learn [9]. In the development stage, we used blocks 0–8 of the dataset described in Sect. 3 to train the model and block 9 to evaluate the model performance. We used precision, recall, and F-score as the evaluation metrics.

Although the classification criteria for BRs and LRRs are not standardized in LREC, the items cited as LRRs are the research data that the authors used in their works. We expect that models that show high performances in discriminating LRRs can appropriately identify research data used in the studies.

5.2 Negative Sampling

We compared the performances of models trained on different sample sizes. We set the sample size N based on the ratio of the BR samples to the LRR samples. For each N, we trained the model on N randomly sampled BRs and all 1,407 LRRs in blocks 0–8 and evaluated the performance on block 9.

We repeated the procedures for training and evaluating the model 100 times for each N and compared the averages of the corresponding F-scores. The best performance was observed for $N = 2{,}110$, which is 1.5 times the size of the LRRs.

5.3 Cross-validation Test

We assessed the proposed method by 10-fold cross-validation. At each step, we trained the model with nine blocks of the dataset and evaluated it with the other one block (i.e., blocks 1–9 for training and block 0 for evaluation and so on). To train the model, we used 2,110 randomly sampled BRs and all LRRs in the training blocks based on the result of the negative sampling experiment.

Blocks 0–7 were used for evaluation in the cross-validation test, and blocks 8 and 9 were used for key phrase investigation and model development, respectively. For each metric, we calculated the average of the values recorded in each of the eight steps.

We compared the proposed model performance to a baseline model that uses only bibliographical element-related features (listed in Table 1). We performed the cross-validation test 10 times and compared the average of each attempt as the final result, because the performance of the model fluctuates due to the random negative sampling.

5.4 Experimental Result

Table 3. Cross-validation result

	Precision (%)	Recall (%)	F1 score
Baseline (without citation context)	40.2	46.0	42.9
Proposed method (with citation context)	45.0	51.5	48.0

Table 3 summarizes the final result of the cross-validation test. Our proposed model outperforms the baseline model, proving the effectiveness of using citation context.

We describe an LRR instance that only the proposed model correctly classified as a language resource below:

Title of the cited item: Novel word-sense identification
Title of the section with the cited item: 3.1.2. Novel sense Dataset
Citation context: Here we use the dataset provided by (Cook et al., 2014).

The title of this instance does not include any key phrases. Meanwhile, the bibliographical elements include only the title, while lacking other information, such as the name of the journal or conference. With no clues in the bibliographical elements, the baseline model could not correctly classify this instance as a language resource.

On the contrary, the section title and citation context included the word dataset, which is a key phrase that enabled the proposed model to classify the instance as a language resource. Thus, our model is capable of identifying LRRs even without sufficient clues in the bibliographical elements.

6 Conclusion

In this paper, we proposed a method for the automatic identification of research data references in publications. Firstly, we described the reference list division rule in LREC and stated that the purpose of this work is to automate the discrimination process.

Furthermore, we investigated the reference lists in LREC papers and the citation contexts to find useful characteristic features for identifying research data references. We confirmed that key phrases appeared in the citation contexts and the bibliographical elements in the reference lists.

Our proposed method uses the presences or absences of key phrases to identify research data references. Experiments on LREC proceedings papers proved the effectiveness of using key phrases in the citation context.

In future work, we will evaluate the proposed method using datasets labeled based on the definite classification criteria for BRs and LRRs, which the author guidelines of LREC does not specify. Some authors still list research data used in their research in BRs, especially when they cite publications on constructing the research data. We investigated 36 randomly sampled papers without research data references and found that half of them included citations that the authors should have listed as LRRs. This result suggests that a more definite guideline is needed, which is easier for authors to follow.

Additionally, we will verify the effectiveness of citation contexts in other research fields and consider a method for automatic key phrase extraction.

References

1. PDFNLT. https://github.com/KMCS-NII/PDFNLT-1.0
2. LREC Author's kit (2016). https://www.lrec2016.lrec-conf.org/en/submission/authors-kit/
3. Ahtaridis, E., Cieri, C., DiPersio, D.: LDC language resource database: Building a bibliographic database. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 1723–1728. European Language Resources Association (ELRA), Istanbul, May 2012
4. Calzolari, N. et al. (eds.): Proceedings of LREC 2016, 2018, and 2020. http://www.lrec-conf.org/proceedings/
5. Choukri, K., Arranz, V., Hamon, O., Park, J.: Using the international standard language resource number: practical and technical aspects. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 50–54. European Language Resources Association (ELRA), Istanbul, May 2012
6. Kozawa, S., Tohyama, H., Uchimoto, K., Matsubara, S.: Collection of usage information for language resources from academic articles. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta, May 2010
7. Mapelli, V., Popescu, V., Liu, L., Choukri, K.: Language resource citation: the ISLRN dissemination and further developments. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1610–1613. European Language Resources Association (ELRA), Portorož, May 2016
8. Namba, H.: Construction of an academic resource repository. In: Proceedings of Toward Effective Support for Academic Information Search Workshop, pp. 8–14 (2018)
9. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
10. Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S., Isahara, H.: Construction of an infrastructure for providing users with suitable language resources. In: Coling 2008: Companion volume: Posters, pp. 119–122. Coling 2008 Organizing Committee, Manchester, August 2008
11. Zinn, C.: Squib: The language resource switchboard. Comput. Linguist. **44**(4), 631–639 (2018)