Using Word Embeddings for Library and Information Science Research: A Short Survey

Marie Katsurai
Doshisha University

In the research field of library and information science (LIS), pattern recognition and machine learning approaches have often been adopted to analyze massive amounts of textual data in digital libraries. In recent natural language processing, word embeddings have attracted much attention because they can capture semantic meanings and contexts. This paper presents a short survey of how word embeddings have been used in LIS research, especially focusing on articles published in LIS journals between 2009 and 2019. Our simple bibliographic analysis showed that at least 15 LIS journal papers used word embeddings. These 15 papers were briefly described under the following categories: knowledge extraction and visualization from scholarly data, classification in scholarly data, and others that focus on more general corpora.

DOI: 10.1145/3387726.3387730 http://doi.acm.org/10.1145/3387726.3387730

1. INTRODUCTION

Library and information science (LIS) has been recognized as an interdisciplinary field covering a wide range of topics, such as information retrieval, organizing, and data mining. These topics are also of interest in ACM SIGWEB community, and sharing knowledge, theories, tools, and techniques brings benefits to both research fields. Such digital library research involves many text classification or knowledge extraction problems. Solving these via pattern recognition and machine learning approaches often requires representing words/sentences as numerical vectors. One of the modern semantic representations is word embedding, which learns a vector space considering the contexts in which words appear. The most famous model is called word2vec [Mikolov et al. 2013]. In word embeddings, a word can correspond to a vector, and similar words are likely to have similar vectors. This property has been adopted in many natural language processing tasks, such as word relatedness measurement and textual feature extraction. This paper presents a short survey of how word embeddings have been used in LIS research, especially focusing on articles published in top-ranked LIS journals.

For this survey, we first chose 25 journals as an information source of data collection. These journals were ranked by Google Scholar Metrics on the basis of h-index in the category of "Social Sciences – Library& Information Science." We extracted from the Scopus Abstract and Citation database[1] the titles and abstracts of papers published in these journals between 2009 and 2019. Note that we only collected articles, excluding reviews,

[1]https://www.scopus.com

Table I. LIS journals and their number of papers used for our bibliographic analysis.

Journal	Num. papers
Journal of the Association for Information Science and Technology	1,065
Journal of the American Society for Information Science and Technology	895
Scientometrics	3,114
Journal of Informetrics	798
Journal of Academic Librarianship	701
Online Information Review	593
Journal of Information Science	618
College and Research Libraries	341
Journal of Documentation	605
Portal	317
Electronic Library	604
The Electronic Library	28
Aslib Journal of Information Management	225
Aslib Proceedings: New Information Perspectives	170
Aslib Proceedings	21
Information Development	462
Learned Publishing	316
Journal of the Medical Library Association	375
Library and Information Science Research	1,090
Library Hi Tech	429
Journal of Librarianship and Information Science	339
Information Research	457
Information and Learning Science	128
New Library World	357
Library Philosophy and Practice	1,552

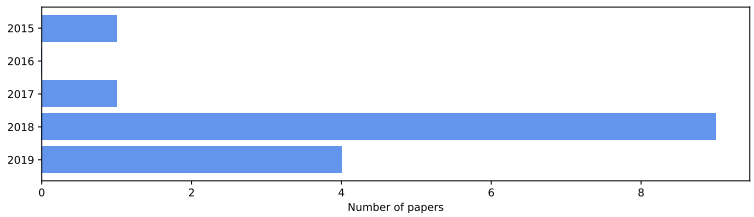

Fig. 1. Number of LIS papers whose titles or abstracts include the term “word embedding” or “word2vec.”

letters, editorials, and other types of documents. Table I presents a list of the 25 journals we used and their number of papers collected.

From the paper collection, we further extracted papers whose abstracts include the terms “word embedding” or “word2vec.” Figure 1 shows the number of papers extracted for each publication year. As shown, the total number of papers that used word embeddings was at least 15. The next section describes the motivation and methodology of each of these papers.

Table II. References introduced in Section 2 and their tasks that used word embeddings.

Reference	Task
[Sun and Ding 2018]	Visualization of the temporal changes of word meanings
[Zhang et al. 2018]	Topic extraction from bibliometric data
[Hu et al. 2018a]	Topic visualization from bibliographic data
[Hu et al. 2018b]	Topic visualization from bibliographic data
[Choi 2018]	Biomedical entity classification
[Mao and Cui 2018]	Biological entity classification
[Kholghi et al. 2017]	Clinical concept classification
[Heffernan and Teufel 2018]	Problem/solution classification
[Zhao et al. 2018]	Reviewer field classification
[Li et al. 2018]	Patent classification
[Zhang and He 2015]	Sentiment classification
[Ma and Luo 2019]	Opinion classification
[Lee et al. 2019]	Semantic relatedness measurement
[Qian et al. 2019]	New word detection
[Mahdaouy et al. 2019]	Query expansion in information retrieval

2. USAGE OF WORD EMBEDDINGS IN LIS JOURNAL PAPERS

This section briefly describes how the LIS papers included in Fig. 1 used word embeddings. The references and their tasks are summarized in Table II. We found that the literature can be roughly divided into the following categories: knowledge extraction and visualization from scholarly data, classification in scholarly data, and other tasks based on more general corpora.

2.1 Knowledge extraction and visualization from scholarly data

Knowledge extraction and visualization from scholarly data, also known as science mapping, have been actively studied in LIS. The vector space of word embeddings often provides a visualization of domain knowledge by placing related words close to one another. [Sun and Ding 2018] used word embeddings to understand the semantic changes of scientific and technological knowledge *memes* in a target research domain. The authors trained word2vec using the titles of scholarly data to capture the semantic meanings of memes. They showed that the semantic meanings of some memes were shifting over time using the form of networks, in which nodes were words and edges represent strong relatedness between words.

To extract topics from bibliographic data, [Zhang et al. 2018] used word embeddings as features for clustering. After training word2vec on the basis of a target corpus, they calculated a vector representation of each paper by simply averaging the vectors of all words within the paper. The word vectors were grouped using k-means clustering. An empirical study conducted on bibliometric journal papers showed that word embeddings improved topic extraction performance compared with other text features. However, the authors also concluded that word embeddings lead to a lack of interpretability and explainability of the topics extracted.

In [Hu et al. 2018b], with the aim of visualizing the knowledge structure of a target research domain, word2vec was used to merge keywords into a semantic unit. Using a TF-IDF like approach, infrequent but important units were identified as the domain-specific research

topics. For a similar purpose, in [Hu et al. 2018a], word2vec was used to map scientific keywords into a two-dimensional space constructed using the t-SNE algorithm. The resulting map with keyword citation counts contributed to a new understanding of previous key topics in the literature related to geographical natural hazards.

2.2 Classification in scientific data

Extracting scientific knowledge from bibliographic data has been sometimes formulated as a specific classification task. [Choi 2018] constructed a convolutional neural network (CNN) to extract pairs of protein names that interact with each other from a biomedical paper collection. Using a public word embedding model, sentence embedding features that consider word positions were calculated. The vector output by feeding the features into a convolution layer was concatenated to manually defined linguistic features. The final classification layer determines whether two proteins interact or not. This approach achieved remarkably superior performance compared with conventional methods that used handcrafted features only. [Mao and Cui 2018] extracted bacterial biotope entities from scientific documents using machine learning techniques, which is a named entity recognition task in biology and bioinformatics. To identify the types of entities in the text, they presented several textual features; one of these was calculated using biomedical word2vec provided by bio.nlplab.org[2]. Word vectors were grouped using K-means clustering, and the identifiers of clusters that a given token belongs to were used as new embedding features. The experiments demonstrated that word embeddings were the most helpful features when a large and more general external corpus is available. [Kholghi et al. 2017] presented an active learning approach based on word and sequence representations for clinical information extraction. In this method, word2vec was used to semantically cluster words. Specifically, each word was represented by a 100-dimensional vector produced by word2vec, which was further concatenated to an n-gram based vector. In experiments for extracting concepts, such as clinical problems, tests, and treatments, the authors demonstrated that the proposed clustering can be effective for reducing manual annotation effort.

[Heffernan and Teufel 2018] used supervised machine learning to classify a given textual description into a scientific problem or a solution. To train a binary classifier, such as a support vector machine, they presented 14 types of features; most of these were handcrafted features, while some were designed based on word/document embeddings that were trained using an external biomedical corpus. Experiments were conducted using 2,000 sentences labeled by the authors themselves, demonstrating that combining all features achieved the best classification result.

[Zhao et al. 2018] presented a method that automatically recommends reviewers for a submitted paper. This problem assumed that a submission has multiple tags assigned by its authors, while a reviewer candidate is described with his/her research interests. Zhao et al. first trained word2vec using Google News corpus [Mikolov et al. 2013] and utilized it to convert all words to 300-dimensional vectors. A submission or a reviewer was then characterized by vectors corresponding to its tags. The distance between a reviewer's tags and a submission's tags was calculated in an optimization framework, which corresponds to the mining relationships between papers and reviewers. Finally, they trained a classifier

[2]http://bio.nlplab.org/

that predicts research field labels for reviewer candidates on the basis of the learnd distance metrics.

In addition to academic papers, patents are also important academic accomplishments. [Li et al. 2018] constructed CNNs for multi-label patent classification. Each word in the dataset was first represented by a 200-dimensional vector via word2vec. A patent document was converted into a dense text matrix using the word vectors, which was fed to the CNNs to output the prediction probabilities over labels. Experiments using a new dataset constructed by the authors and CLEP-IP competition dataset [Piroi and Hanbury 2019] showed the effectiveness of the proposed method compared with traditional machine learning algorithms.

2.3 Other tasks in diverse corpora

The scope of LIS journals is, of course, not only limited to scholarly data analysis, such as that explained above, but also covers more general digital libraries. Sentiment classification has been one of the topics in LIS journals. [Zhang and He 2015] used topic models and word embeddings to calculate semantic features for determining the sentiment polarity of sentences from small training data. Specifically, a testing sentence was enriched with a set of related words by using word2vec trained on a large external dataset. Experiments showed that an ensemble approach based on both topics and word2vec for calculating such additional features improved sentiment classification performance. Related to this task, [Ma and Luo 2019] classified opinion types of tweets about rumour topics. They used word2vec trained on Wikipedia corpus and averaged the word vectors of each tweet to produce a single feature vector.

Measuring the semantic relatedness between two words has been a basic component in several text-based applications. To improve the performance of this task, [Lee et al. 2019] presented the linear combination of the cosine similarity between word vectors with path-based features extracted from a lexical ontology, such as WordNet [Miller 1995]. They further used support vector regression to select meaningful features in a supervised setting. Experiments conducted using eight benchmark datasets showed the effectiveness of such a semantic combination with feature selection. [Qian et al. 2019] addressed a problem of detecting new words from Chinese sentences, which is usually difficult because of the absence of a separator between words. All word units in a target domain corpus were first mapped to a vector space using word2vec, in which some word units might be inappropriately separated. After the extraction of frequent n-grams as new word candidates, word units that have highly similar vectors were merged into a single word string, which avoids the oversegmentation of new words. Experiments based on four corpora showed the effectiveness of this approach compared with traditional unsupervised methods. There are many language-specific problems similar to these and are not limited to Chinese sentences; [Mahdaouy et al. 2019] studied word representations for query expansion in Arabic information retrieval. Their experiments demonstrated that word embedding-based term similarity measurement significantly improved retrieval performance.

3. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presented a short survey on the use of word embeddings in LIS research, especially focusing on LIS-related journal papers. We found that the presented tasks are diverse, and such real-word applications pose a great need to further develop semantic representations of textual content. Although word embeddings have often been adopted in several research fields related to textual analysis, our simple bibliographic analysis only found 15 LIS journal papers that used embedding models, which is an interesting phenomenon.

We are currently developing several systems that use scholarly data, including a system for author matching across multiple academic databases [Katsurai and Ohmukai 2019], research trend visualization [Katsurai and Ono 2019; Katsurai 2017], and research interest visualization [Katsurai et al. 2016; Araki et al. 2017; Nishizawa et al. 2018]. It would be valuable to integrate word embeddings to these systems, which will be the focus of our future work. In addition, we will design new embedding models for learning the meanings of technical terms in scholarly data.

ACKNOWLEDGMENTS

This research has been partly supported by JSPS KAKENHI grant number 17K12794, JST ACT-I grant number JPMJPR18UC, and JST ACT-X grant number JPMJAX1909. We thank Dr. Soohyung Joo for providing the bibliographic data used in this study.

REFERENCES

ARAKI, M., KATSURAI, M., OHMUKAI, I., AND TAKEDA, H. 2017. Interdisciplinary collaborator recommendation based on research content similarity. *IEICE Trans. Information and Systems E99-D,* 4 (April), 785–792.

CHOI, S.-P. 2018. Extraction of proteinprotein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings. *Journal of Information Science 44,* 1, 60–73.

HEFFERNAN, K. AND TEUFEL, S. 2018. Identifying problems and solutions in scientific text. *Scientometrics 116,* 2 (Aug), 1367–1382.

HU, K., QI, K., YANG, S., SHEN, S., CHENG, X., WU, H., ZHENG, J., MCCLURE, S., AND YU, T. 2018a. Identifying the "Ghost City" of domain topics in a keyword semantic space combining citations. *Scientometrics 114,* 3 (Mar), 1141–1157.

HU, K., WU, H., QI, K., YU, J., YANG, S., YU, T., ZHENG, J., AND LIU, B. 2018b. A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. *Scientometrics 114,* 3 (Mar), 1031–1068.

KATSURAI, M. 2017. Bursty research topic detection from scholarly data using dynamic co-word networks: A preliminary investigation,. In *Proc. IEEE Int. Conf. Big Data Analysis (ICBDA)*. 115–119.

KATSURAI, M. AND OHMUKAI, I. 2019. Author matching across different academic databases: Aggregating simple feature-based rankings. In *2019 ACM/IEEE Joint Conf. Digital Libraries (JCDL)*. 279–282.

KATSURAI, M., OHMUKAI, I., AND TAKEDA, H. 2016. Topic representation of researcher's interests in a large-scale academic database and its application to author disambiguation. *IEICE Trans. Information and Systems E99-D,* 4 (April), 1010–1018.

KATSURAI, M. AND ONO, S. 2019. TrendNets: Mapping emerging research trends from dynamic co-word networks via sparse representation. *Scientometrics 121,* 3 (Dec), 1583–1598.

KHOLGHI, M., DE VINE, L., SITBON, L., ZUCCON, G., AND NGUYEN, A. 2017. Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings. *Journal of the Association for Information Science and Technology 68,* 11, 2543–2556.

LEE, Y.-Y., KE, H., YEN, T.-Y., HUANG, H.-H., AND CHEN, H.-H. 2019. Combining and learning word embedding with wordnet for semantic relatedness and similarity measurement. *Journal of the Association for Information Science and Technology n/a,* n/a.

LI, S., HU, J., CUI, Y., AND HU, J. 2018. DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics 117,* 2 (Nov), 721–744.

MA, J. AND LUO, Y. 2019. The classification of rumour standpoints in online social network based on combinatorial classifiers. *Journal of Information Science 0,* 0, 1–14.

MAHDAOUY, A. E., ALAOUI, S. O. E., AND GAUSSIER, E. 2019. Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. *Journal of Information Science 45,* 4, 429–442.

MAO, J. AND CUI, H. 2018. Identifying bacterial biotope entities using sequence labeling: Performance and feature analysis. *Journal of the Association for Information Science and Technology 69,* 9, 1134–1147.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. Advances in Neural Information Processing Systems (NIPS)*. 3111–3119.

MILLER, G. A. 1995. WordNet: A lexical database for English. *Commun. ACM 38,* 11 (Nov.), 39–41.

NISHIZAWA, H., KATSURAI, M., OHMUKAI, I., AND TAKEDA, H. 2018. Measuring researcher relatedness with changes in their research interests. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 149–152.

PIROI, F. AND HANBURY, A. 2019. *Multilingual Patent Text Retrieval Evaluation: CLEF–IP*. Springer International Publishing, Cham, 365–387.

QIAN, Y., DU, Y., DENG, X., MA, B., YE, Q., AND YUAN, H. 2019. Detecting new Chinese words from massive domain texts with word embedding. *Journal of Information Science 45,* 2, 196–211.

SUN, X. AND DING, K. 2018. Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents. *Scientometrics 116,* 3 (Sep), 1735–1748.

ZHANG, P. AND HE, Z. 2015. Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science 41,* 4, 531–549.

ZHANG, Y., LU, J., LIU, F., LIU, Q., PORTER, A., CHEN, H., AND ZHANG, G. 2018. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics 12,* 4, 1099 – 1117.

ZHAO, S., ZHANG, D., DUAN, Z., CHEN, J., ZHANG, Y., AND TANG, J. 2018. A novel classification method for paper-reviewer recommendation. *Scientometrics 115,* 3 (Jun), 1293–1313.

Marie Katsurai is currently an assistant professor in the Department of Intelligent Information Engineering and Sciences, Doshisha University. Her research interests include scholarly data mining and multimedia information retrieval.