Bib2vec: Embedding-based Search System for Bibliographic Information

Takuma Yoneda **Koki Mori** **Makoto Miwa** **Yutaka Sasaki**
Department of Advanced Science and Technology
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Japan
{sd14084,sd15435,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

Abstract

We propose a novel embedding model that represents relationships among several elements in bibliographic information with high representation ability and flexibility. Based on this model, we present a novel search system that shows the relationships among the elements in the ACL Anthology Reference Corpus. The evaluation results show that our model can achieve a high prediction ability and produce reasonable search results.

1 Introduction

Modeling relationships among several types of information, such as nodes in information network, has attracted great interests in natural language processing (NLP) and data mining (DM), since their modeling can uncover hidden information in data. Topic models such as author-topic model (Rosen-Zvi et al., 2004) have been widely studied to represent relationships among these types of information. These models, however, need a considerable effort to incorporate new types and do not scale well in increasing the number of types since they explicitly model the relationships between types in the generating process.

Word representation models, such as skip-gram and continuous bag-of-word (CBOW) (Mikolov et al., 2013), have made a great success in NLP. They have been widely used to represent texts, but recent studies started to apply these methods to represent other types of information, e.g., authors or papers in citation networks (Tang et al., 2015).

We propose a novel embedding model that represents relationships among several elements in bibliographic information, which is useful to discover hidden relationships such as authors' interests and similar authors. We built a novel search system that enables to search for authors and words related to other authors based on the model using the ACL Anthology Reference Corpus (Bird et al., 2008). Based on skip-gram and CBOW, our model embeds vectors to not only words but also other elements of bibliographic information such as authors and references and provides a great representation ability and flexibility. The vectors can be used to calculate distances among the elements using similarity measures such as cosine distance and inner products. For example, the distances can be used to find words or authors related to a specific author. Our model can easily incorporate new types without changing the model structure and scale well in the number of types.

2 Related works

Most previous work on modeling several elements in bibliographic information is based on topic models such as author-topic model (Rosen-Zvi et al., 2004). Although the models work fairly well, they have comparably low flexibility and scalability since they explicitly model the generation process. Our model employs word representation-based models instead of topic models.

Some previous work embedded vectors to the elements. Among them, large-scale information network embedding (LINE) (Tang et al., 2015) embedded a vector to each node in information network. LINE handles single type of information and prepares a network for each element separately. By contrast, our model simultaneously handles all the types of information.

3 Method

We propose a novel method to represent bibliographic information by embedding vectors to elements based on skip-gram and CBOW.

3.1 Task definition

We assume the bibliographic data set has the following structure. The data set is composed of bib-

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 112–115

liographic information of papers. Each paper consists of several categories. Categories are divided into two groups: a textual category Ψ (e.g., titles and abstracts[1]) and non-textual categories Φ (e.g., authors and references). Figure 1 illustrates an example structure of bibliographic information of a paper. Each category has one or more elements; the textual category usually has many elements while a non-textual category has a few elements (e.g., authors are not many for a paper).

3.2 Proposed model

Our model focuses on a *target* element, and predicts a *context* element from the target element. We use only the elements in non-textual categories as contexts to reduce the computational cost. Figure 1 shows the case when we use an element in a non-textual category as a target. For the black-painted target element in category Φ^2, the shaded elements in the same paper are used as its contexts.

When we use elements in the textual category as a target, instead of treating each element as a target, we consider that the textual category has only one element that represents all the elements in the category like CBOW. Figure 1 exemplifies the case that we consider the averaged vector of the vectors of all the elements in the textual category as a target.

We describe our probabilistic model to predict a context element e^j_O from a target e^i_I in a certain paper. We define two d-dimensional vectors v^i_t and ω^i_t to represent an element e^i_t as a target and context, respectively. Similarly to the skip-gram model, the probability to predict element e^j_O in the context from input e^i_I is defined as follows:

$$p(e^j_O|e^i_I) = \frac{\exp(\omega^j_O \cdot v^i_I + \beta^j_O)}{\sum_{(\omega^j_s,\beta^j_s)\in S^j} \exp(\omega^j_s \cdot v^i_I + \beta^j_s)}, \quad e^j_O \in \Phi, \;\; e^i_I \in \Psi \cup \Phi, \tag{1}$$

where β^j_s denotes a bias corresponds to ω^j_s, and S^j denotes pairs of ω^j_s and β^j_s that belong to a category Φ^j. As we mentioned, our model considers that the textual category Ψ has only one averaged vector. The vector v^j_{rep} can be described as:

$$v^j_{rep} = \frac{1}{n}\sum_{q=1}^{n} v^j_q, \; e^j \in \Psi \tag{2}$$

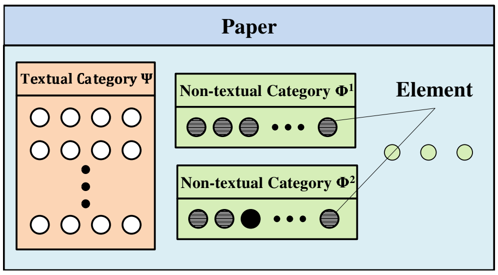

Figure 1: Example of the bibliographic information of a paper when the target is the element in the non-textual category. The black element is a target and the shaded elements are contexts.

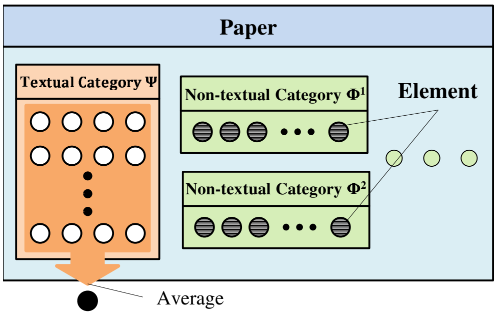

Figure 2: Example when the target is the elements in the textual category

Our target loss can be defined as:

$$-\sum_{(e_a,e_b)\in D} \log p(e_b|e_a), \tag{3}$$

where D denotes a set of all the correct pairs of the elements in the data set. To reduce the cost of the summation in Eq. (1), we applied the noise-contrastive estimation (NCE) to minimize the loss (Gutmann and Hyvärinen, 2010).

3.3 Predicting related elements

We predict the top k elements related to a query element by calculating their similarities to the query element. We calculate the similarities using one of three similarity measures: the linear function in Eq. (1), dot product, and cosine distance.

4 Experiments

4.1 Evaluation settings

We built our data set from the ACL Anthology Reference Corpus version 20160301 (Bird et al., 2008). The statistics of the data set and our model settings are summarized in Table 1.

As pre-processing, we deleted commas and periods that sticked to the tails of words and removed non-alphabetical words such as numbers

[1] Note that we have only one textual category since the categories for texts are usually not distinguished in most word representation models.

		#Elements		Min.
Category	Type	Original	Processed	Freq.
text	textual	59,276	10,994	20
author	non-textual	17,260	2,609	5
reference	non-textual	10,871	10,871	1
year	non-textual	16	16	1
paper-id	non-textual	19,475	19,475	1

Table 1: Summary of our data set and model

and brackets from abstracts and titles. We then lowercased the words, and made phrases using the word2phrase tool[2].

We prepared 5 categories: *author*, *paper-id*, *reference*, *year* and *text*. *author* consists of the list of authors without distinguishing the order of the authors. *paper-id* is an unique identifier assigned to each paper, and this mimics the paragraph vector model (Le and Mikolov, 2014). *reference* includes the paper ids of reference papers in this data set. Although ids in paper-id and reference are shared, we did not assign the same vectors to the ids since they are different categories. *year* is the publication year of the paper. *text* includes words and phrases in both abstracts and titles, and it belongs to the textual category Ψ, while each other category is treated as a non-textual category Φ^i. We regard elements as unknown elements when they appear less than minimum frequencies in Table 1.

We split the data set into training and test. We prepared 17,475 papers for training and the remaining 2,000 papers for evaluation. For the test set, we regarded the elements that do not appear in the training set as unknown elements.

We set the dimension d of vectors to 300 and show the results with the linear function.

4.2 Evaluation

We automatically built multiple choice questions and evaluate the accuracy of our model. We also compared some results of our model with those of author-topic model.

Our method models elements in several categories and allows us to estimate relationships among the elements with high flexibility, but this makes the evaluation complex. Since it is tough to evaluate all the possible combinations of inputs and targets, we focused on relationships between authors and other categories. We prepared an evaluation data set that requires to estimate an author from other elements. We removed an (not unknown) author from each paper in the evaluation set to ask the system to predict the removed author considering all the other elements in the paper. To choose a correct author from all the authors can be insanely difficult, so we prepared 10 selection candidates. In order to evaluate the effectiveness of our model, we compared the accuracy on this data set with that by logistic regression. As a result, when we use our model, we got 74.3% (1,486 / 2,000) in accuracy, which was comparable to 74.1% (1,482 / 2,000) by logistic regression.

Table 2 shows the examples of the search results using our model. The leftmost column shows the authors we input to our model. In the rightmost two columns, we manually picked up words and authors belonging to a certain topic described in Sim et al. (2015) that can be considered to correspond to the input author. This table shows that our model can predict relative words or similar authors favorably well although the words are inconsistent with those by the author topic model.

Figure 3 shows the screenshot of our system. The lefthand box shows words in the word cloud related to the query and the righthand box shows the close authors. We can input a query by putting it in the textbox or click one of the authors in the righthand box and select a similarity measure by selecting a radio button.

4.3 Discussion

When we train the model, we did not use elements in category Ψ as context. This reduced the computational costs, but this might disturbed the accuracy of the embeddings. Furthermore, we used the averaged vector for the textual category Ψ, so we do not consider the importance of each word. Our model might ignore the inter-dependency among elements since we applied skip-grams. To resolve these problems, we plan to incorporate attentions (Ling et al., 2015) so that the model can pay more attentions to certain elements that are important to predict other elements.

We also found that some elements have several aspects. For example, words related to an author spread over several different tasks in NLP. We may be able to model this by embedding multiple vectors (Neelakantan et al., 2014).

5 Conclusions

This paper proposed a novel embedding method that represents several elements in bibliographic information with high representation ability and

[2] `https://github.com/tmikolov/word2vec`

	Our Model		Author Topic-Model	
Input Author	Relevant Words	Similar Authors	Topic Words	Topic Authors
Philipp Koehn	machine translation	Hieu Hoang	alignment	Chris Dyer
	hmeant	Alexandra Birch	translation	Qun Liu
	human translators	Eva Hasler	align	Hermann Ney
Ryan McDonald	dependency parsing	Keith Hall	parse	Michael Collins
	extrinsic	Slav Petrov	sentense	Joakim Nivre
	hearing	David Talbot	parser	Jens Nilson

Table 2: Working examples of our model and author topic-model

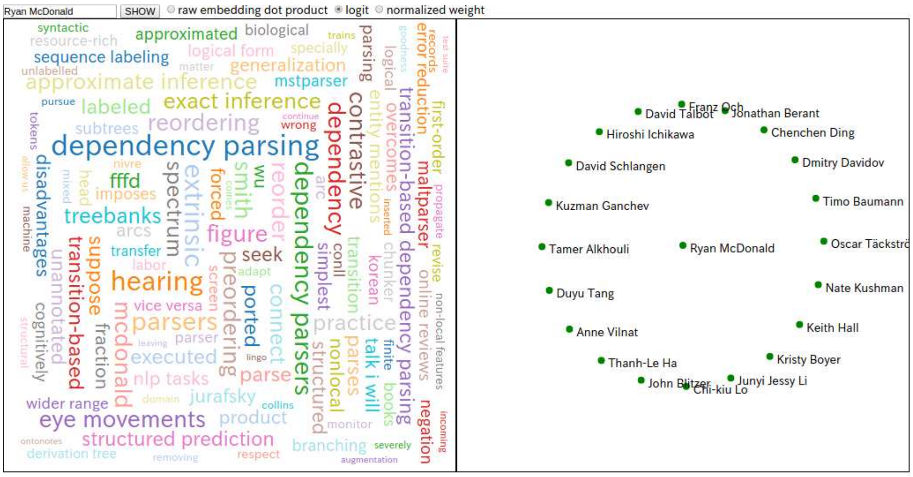
Ryan McDonald SHOW ○ raw embedding dot product ● logit ○ normalized weight

Figure 3: Screen shot of the system with the search results for the query "Ryan McDonald".

flexibility, and presented a system that can search for relationships among the elements in the bibliographic information. Experimental results in Table 2 show that our model can predict relative words or similar authors favorably well. We plan to extend our model by other modifications such as incorporating attention and embedding multiple vectors to an element. Since this model has high flexibility and scalability, it can be applied to not only papers but also a variety of bibliographic information in broad fields.

Acknowledgments

We would like to thank the anonymous reviewer for helpful comments and suggestions.

References

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, pages 297–304.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. 2015. Not all contexts are created equal: Better word representations with variable attention. In *EMNLP*, pages 1367–1372.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*, pages 1059–1069.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI*, pages 487–494.

Yanchuan Sim, Bryan R. Routledge, and Noah A. Smith. 2015. A utility model of authors in the scientific community. In *EMNLP*, pages 1510–1519.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: large-scale information network embedding. In *WWW*, pages 1067–1077.