Target-Topic Aware Doc2Vec for Short Sentence Retrieval from User Generated Content

Kosuke Kurihara
Aoyama Gakuin University
Kanagawa, Japan
kurihara@sw.it.aoyama.ac.jp

Yoshiyuki Shoji
Aoyama Gakuin University
Kanagawa, Japan
shoji@it.aoyama.ac.jp

Sumio Fujita
Yahoo JAPAN Corporation
Tokyo, Japan
sufujita@yahoo-corp.jp

Martin J. Dürst
Aoyama Gakuin University
Kanagawa, Japan
duerst@it.aoyama.ac.jp

ABSTRACT

This paper proposes a new method of supplementing the context of short sentences for the training phase of Doc2Vec. Since CGM (Consumer Generated Media) sites and SNS sites become widespread, the importance of similarity calculation between a given query and a short sentence is increasing. As an example, a search by the query "sad" should find actual expressions such as "I needed a handkerchief" on a movie review site. Doc2Vec is one of the most widely used methods for vectorization of queries and sentences. However, Doc2Vec often exhibits low accuracy if the training data consists of short sentences, because they lack context. We modified Doc2Vec with the hypothesis that other posts for the same topic (*i.e.* reviews for the same movie in online movie review sites) may share the same background. Our method uses target-topic IDs instead of sentence IDs as the context in the training phase of the Doc2Vec with the PV-DM model; this model estimates the next term from a few previous terms and context. The model trained with item IDs vectorizes a sentence more accurately than a model trained with sentence IDs. We conducted a large-scale experiment using 1.2 million movie review posts and a crowdsourcing-based evaluation. The experimental result demonstrates that our new method achieves higher precision and nDCG than previous Doc2Vec variants and traditional topic modeling methods.

KEYWORDS

Online Review Sites, Doc2Vec, Information Retrieval, Impression Retrieval

1 INTRODUCTION

The present-day Web can be seen both as a collection of fragmented sentences and as a set of documents. Due to the rise of CGM (Consumer Generated Media) sites and SNS (Social Networking Service) sites, many people frequently post a large number of short sentences on the Web. The Web contains many by-products of communication and short posts of self-expression. Following such change of property of the Web resources, the importance of similarity calculation between short sentences and queries is increasing. Most of the existing search algorithms mainly target larger documents, but there are few suitable algorithms for searching Web resources with a lot of fragmented sentences.

Doc2Vec [6] is one of the most popular methods to vectorize sentences and to calculate the similarity between sentences and a query. However, Doc2Vec exhibits low accuracy in learning short sentences because short sentences do not provide enough context. This paper proposes a method to supplement for the lack of context during the training phase of Doc2Vec when learning short sentences from the social sites. Posts in such kinds of sites generally have a target-topic. For instance, a social media post often contains one or more hashtags, a review in an online review site has a target item, and a comment in an online discussion forum has a news article as the topic. Here we introduce the term "target-topic" for the objects or referents of these sentences. We modify Doc2Vec by using target-topics as an additional context for training the Doc2Vec network. The expectation is that this allows to exploit the similarity between sentences that relate to the same target-topic, and will therefore improve search accuracy.

A typical application is movie search using reviews: Imagine you want to watch a movie that makes you feel nostalgic. Using the search term "nostalgic", it should be possible to find reviews with contain text fragments such as "brings back memories" or "reminds me of my childhood". Using vectorization including target-topics, we significantly improve the accuracy of the similarity calculation between queries and target sentences.

Details of the new method are given in Section 3. Section 4 describes the experimental setup and Section 5 provides the results to show the effectiveness of using target-topics.

2 RELATED WORK

We introduce related work from the viewpoint of our proposed method and our application.

Our method is an instance of information retrieval with distributed expressions. There are many methods using Word2Vec or Doc2Vec to find information from social sites. Van Gysel *et al.* [10] also used Doc2Vec models for short sentences in the social sites. It is active research field that tackle short sentences in social sites, Trieu *et al.* [9] propose a method for tagging and classifying news information posted on Twitter and searching for similar news. Neither of these use target-topics to improve search accuracy. Zuo [13] *et.al* uses the external information to vectorize short sentences in social sites for probabilistic topic models. Our method also uses external context for Doc2Vec-based vectorization. The main difference is that we use target-topic as an external information; our method does not need any ontologies or dictionaries.

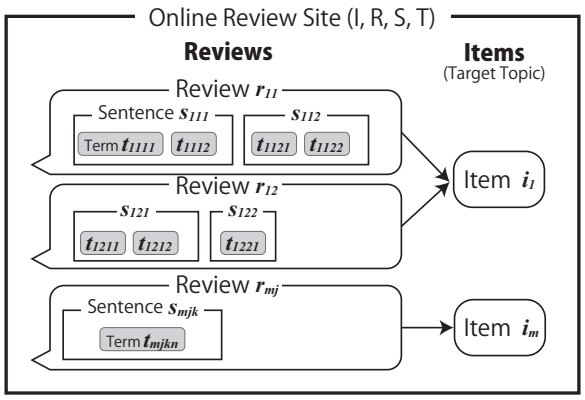

Figure 1: Site structure of a typical online review site. Every sentence points to one item (target-topic).

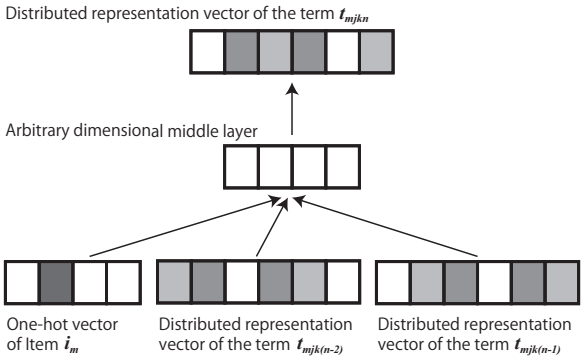

Figure 2: Input and output of training with window size 2. The color depth of each cell reflects the value of each dimension.

Online review analysis is the second research field strongly related to our proposal. Online reviews are a powerful information resource for item retrieval [5], recommendation [11], decision support [3, 12], and so on. Singh *et al.* [7] and Bader *et al.* [1] focus on expressions and sentiments in reviews. Jo *et al.* [4] propose a method that automatically detects the combination of various aspects and polarities in reviews. Tan *et al.* [8] propose another way to find short sentences which have the similar sentiment. There are two advantages of our method when compared to previous research. It can search for sentences with any aspects, not only sentiments but also story patterns or genres. Also, it does not need preparing a sentiment label dictionary in advance.

3 TARGET-TOPIC AWARE DOC2VEC

Overall, our proposed method is a variant of the PV-DM (Paragraph Vector Distributed Memory) model of Doc2Vec. First, a two-layer network is trained through the task of estimating the next term in a sentence from its context. In the original PV-DM model, a separate context is used for each sentence. We modify this by using the target-topic as the context for all sentences about this target-topic. Second, each sentence is vectorized using the trained network. This step is the same as in the original model, but as a result, the granularity is different for the training step and the vectorization step.

To explain the details of our new method, we use an online review site as an example; it is one of the most typical examples of a CGM site. Online review sites have a common structure as shown in Figure 1. In this example, we call target-topic of a review as "item" for clarity. A review site consists of items (target-topics) I, reviews R, sentences S and terms T. Each item i_m is discussed by a number of reviews. Each review r_{mj} for item i_m consists of a number of sentences. Each sentence s_{mjk} is a sequence of terms denoted by t_{mjkn}. The goal of our method is to vectorize s_{mjk} accurately. For that purpose, our method uses item i_m as the context of s_{mjk}.

As shown in Figure 2, we modified the input vector for training as follows:

$$v(t_{mjkn}) = \left(v_{\text{onehot}}(i_m), \text{w2v}(t_{mjk(n-w)}), \cdots, \text{w2v}(t_{mjk(n-1)})\right) \tag{1}$$

where w is the window size, $\text{w2v}(t)$ is a distributed expression of term t obtained by using Word2Vec, and $v_{\text{onehot}}(i_m)$ is the one-hot vector for item i_m. The p-th dimension of $v_{\text{onehot}}(i_m)$ is defined as follows:

$$v_{\text{onehot}}(i_m)_p = \begin{cases} 1 & \text{if} \quad p = m, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It uses one-hot vector about the item instead of about the sentence.

Next, our method vectorizes all the sentences in the dataset by using the trained network. The vectorization procedures are the same to the previous Doc2Vec. The values of the middle layer are used as the vector of a sentence in the vectorization phase. When it was used for the search, a query is vectorized with the same trained model. To vectorize a certain short text, it uses the input vector $v(t_{mjkn})$ that contains an m-dimensional zero vector instead of one-hot vector. The vectorizations of the sentences in the dataset and the query are then used for similarity calculation (*e.g.* cosine similarity).

4 EXPERIMENT

We verified the usefulness of the proposed method by an experiment using movie reviews. Movie reviews are a typical application which can benefit from the proposed method. We compared our new method against three baseline methods. Since our final goal is to make a comfortable item search system based on reviewers' opinions, we evaluated methods with the metrics and the measurement used in information retrieval research. For 10 queries prepared in advance, we retrieved actual movie review sentences that have similar meanings to the queries. After ranking the sentences, a crowdsourcing questionnaire was used to evaluate the degree of matching between the query and a selection of sentences. In addition, we qualitatively evaluated the variability of expression in the actual search results.

4.1 Dataset

We used actual movie review data posted on Yahoo! Movies, one of the biggest online movie review site in Japan. We created movie review dataset which consist of 3,245 movies. Each movie in the

dataset has 300 or more reviews. The dataset contains approximately 1.3 million reviews, and approximately 12 million sentences in total.

We preprocessed the data to make it suitable for Doc2Vec. All sentences were separated into words using a Japanese morphological analyzer. This is an important step because Japanese text is written without spaces between words. Word classes were limited to nouns, adjectives, and verbs. This is different from most other research, where only nouns are used. The reason for this difference is that information such as emotions and impressions at the time of watching a movie, and the situation suitable for watching the movie are also important.

4.2 Evaluated Methods

Four methods were prepared for evaluation. The details of the four evaluated methods are as follows:

- **TTA-D2V**: Target-Topic-Aware Doc2Vec is the proposed method described in Section 3.
- **plain-D2V**: Plain Doc2Vec [6] is a baseline method using Doc2Vec without any changes.
- **LSI**: Latent Semantic Indexing [2] is a baseline method using topic modeling. We selected this method because LSI is considered more suitable for short sentences than probabilistic topic models (*e.g.* pLSA and LDA).
- **random**: Random Extracting ranks sentences randomly from the dataset for any query.

As a Doc2Vec implementation, we used gensim[1] for the proposed method and **plain-D2V**. The vector size is 200 and the window size is 7. All other learning parameters are the default values of gensim. The number of topics (vector size) for **LSI** is 200, equal to the methods using Doc2Vec.

Note that, because of high memory requirements, it was not possible to calculate **plain-D2V** and **LSI** on the whole dataset. For these two methods, the data was randomly reduced to a 10% subset.

4.3 Queries

For the evaluation experiment, we selected 10 queries. The selected queries were derived from tags on movie review sites, categories of movie information sites, and feature articles about movies. Table 1 shows the selected queries and their features. Each query can be roughly classified into three types: representing the scene drawn by the movie, representing emotions and impressions that people have when watching the movie, and representing a situation suitable for watching the movie. These queries were used with each method, and a review sentences with high similarity were retrieved.

4.4 Relevance Labeling with Crowdsourcing

We used Yahoo! Japam Crowdsourcing, that is a well-known Japanese crowdsourcing service, for labeling the search results. The participants were asked to score the similarity between the shown query and the sentence on a scale from 1 to 4: completely different, slight different, slight similar, and completely similar.

The number of questions was 100 for each of 10 queries and for four methods, resulting in a total of 4,000 questions. Because the aim of the method is to find expressions different from the query, sentences which contain the query term itself were removed from the search results.

[1]https://radimrehurek.com/gensim/

Table 1: Queries used for movie review search task (translated from Japanese)

Query	Type
surrealistic	movie contents
surprise ending	
familial love	
near-futuristic	
makes me relaxed	viewer sentiment
makes me sad	
makes me nostalgic	
makes me want to go on a trip	
suitable for a date	situation
rewatchable	

Sentences used for the questions were sampled from the 500 top-ranked results for each of the four methods. The sampling rate was set high in the top part of the ranking, and lower further down. In addition to all of the top 30 sentences, 30 sentences from rank 31 to rank 100 and 40 sentences from rank 101 to rank 500 were randomly selected.

5 RESULTS

This section describes the experimental results from the view points of precision, ranking, and expression diversity. In the experiment, 16,000 answers were collected from 298 crowd workers.

The precision of rankings retrieved by each method is compared. A sentence with an average score of 2.75 points or more for the four answers is defined as relevant. Table 2 shows the precision at 100 and precision at 500 of each method. A precision of 0.07 for random extraction means that the dataset contains around seven percent of correct answers. The method proposed and **LSI** archived higher precision than the other methods. The new method shows significantly higher precision than **LSI** ($p = 0.00$ on Welch's t test). Figure 3 shows the precision in each section of the rankings. The proposed method has the highest precision both in the top ranks and in total.

The ranking accuracy of each method was evaluated using nDCG (normalized Discounted Cumulative Gain). As table 2 shows, the proposed method has a higher nDCG score than the other methods.

Expression diversity of the sentences in the results is also an important factor. As an example, Table 3 shows the top five search results of our new method and of **LSI** for the query "makes me sad". The precision of both methods is almost the same for this query, but our method found more diverse expressions.

6 DISCUSSION

The experimental results show that our method is significantly more accurate than other methods. This can be attributed to one of two reasons, or a mixture of both. First, our new method may be more precise because it uses target-topics as contexts. Second, **LSI** and

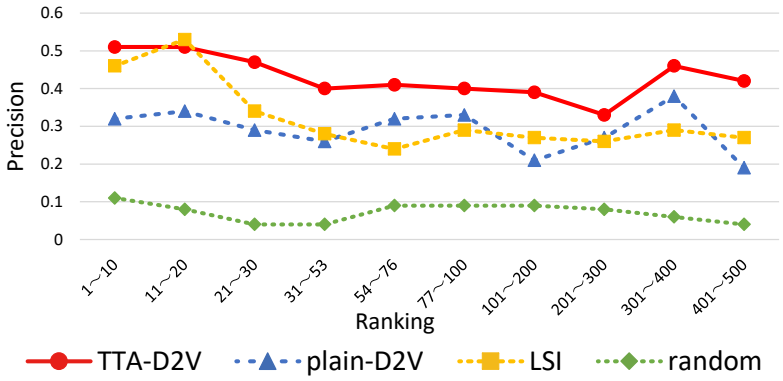

Figure 3: Precision for each method and ranking section

Table 2: Precision and nDCG (Average of 10 queries)

	p@100	p@500	nDCG
TTA-D2V	**0.45**	**0.43**	**0.74**
plain-D2V	0.31	0.29	0.65
LSI	0.36	0.32	0.67
random	0.08	0.07	0.51

Table 3: Top 5 results for proposed method (TTA-D2V) and LSI for the query "makes me sad" (translated from Japanese).

Method	Rank	Sentence	Relevance
TTA-D2V	1	Sentimental people and those who like dogs need a handkerchief.	**3.75**
	2	This moved me to tears.	**3.25**
	3	I sobbed.	**3.75**
	4	You need a handkerchief.	**3.50**
	5	People are in tears, but I can't understand.	2.25
LSI	1	Let's all just cry.	**3.25**
	2	I cried, laughed, and was impressed.	**3.00**
	3	Cry!!	2.00
	4	I cried.	**4.00**
	5	I cried during the song.	**3.50**

plain-D2V may be less precise because they use a smaller dataset, which results from an explosion of the number of dimensions. A more detailed analysis is needed. In any case, our proposed method is likely to be useful for the vectorization of short sentences with target-topics.

With respect to the type of query, Table 4 shows that LSI obtained higher precision for three queries. In the case of the queries "surprise ending" and "familial love", LSI found a few effective synonyms for the query (*i.e.*, "last scene" for "surprise ending"). This shows that LSI is useful for tasks which can be solved by using synonyms. For one query, "makes me want to go on a trip", the precision of the proposed method was not very high. Table 5 shows the example of sentences in the top search results for the query "makes me want to go on a trip". For queries which consists of a sequence of general terms, Doc2Vec and its variants tend to produce a vague vector. Further improvements may be possible by selecting the method based on the type of query.

7 CONCLUSION

This paper proposed a Doc2Vec-based method to vectorize short sentences included in social sites. By focusing on the target-topic, the proposed method can supplement the context of short sentences for the training phase of Doc2Vec. Through a large-scale evaluation with actual movie review data, we showed that our method produces better ranking quality and more diverse results.

A more detailed analysis is needed to understand the advantages of our new method better. This includes comparing the methods by using a dataset of the same size for all methods. The application is also important. We plan to verify the usefulness of our method in other fields. We will also seek advanced ways to use the vectorized sentences, such as embedding for deep learning.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grants Number 18K18161 and 18H03243.

REFERENCES

[1] Nadeem Bader, Osnat Mokryn, and Joel Lanir. 2017. Exploring Emotions in Online Movie Reviews for Online Browsing. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces Companion (IUI '17 Companion)*. ACM, New York, NY, USA, 35–38. https://doi.org/10.1145/3030024.3040982

[2] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391–407.

[3] Nan Hu, Paul A. Pavlou, and Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-mouth Communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC '06)*. ACM, New York, NY, USA, 324–330. https://doi.org/10.1145/1134707.1134743

[4] Yohan Jo and Alice H. Oh. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 815–824. https://doi.org/10.1145/1935826.1935932

Table 4: nDCG and p@k for each method and individual queries

	TTA-D2V		plain-D2V		LSI		random	
Query	nDCG	p@500	nDCG	p@500	nDCG	p@500	nDCG	p@500
surrealistic	**0.74**	**0.28**	0.70	0.17	0.69	0.16	0.64	0.09
surprise ending	0.71	0.28	0.62	0.19	**0.74**	**0.40**	0.53	0.08
familial love	**0.64**	0.25	0.55	0.10	0.62	**0.28**	0.49	0.05
near-futuristic	**0.70**	**0.38**	0.69	0.37	0.63	0.18	0.49	0.03
makes me relaxed	**0.81**	**0.69**	0.74	0.51	0.63	0.37	0.49	0.05
makes me sad	**0.83**	**0.78**	0.66	0.52	0.78	0.77	0.46	0.08
makes me nostalgic	**0.74**	**0.48**	0.69	0.43	**0.74**	0.46	0.48	0.08
makes me want to go on a trip	**0.77**	0.30	0.70	0.06	0.63	**0.33**	0.56	0.03
suitable for a date	**0.77**	**0.36**	0.64	0.17	0.56	0.05	0.57	0.07
rewatchable	**0.76**	**0.50**	0.53	0.39	0.72	0.23	0.45	0.16

Table 5: Top 5 results for proposed method (TTA-D2V) and LSI for the query "makes me want to go on a trip" (translated from Japanese).

Method	Rank	Sentence	Relevance
TTA-D2V	1	I want to go there.	**3.25**
	2	I watched this movie while I was travelling.	2.50
	3	I would better go together with somebody.	**3.25**
	4	It contains a trip scene.	**3.00**
	5	This movie was played in my trip destination.	2.50
LSI	1	I want to go there after a long time	2.25
	2	I wanted to visit Machu Picchu.	**4.00**
	3	I wanted so much to go fishing.	2.25
	4	I wanted to go to the airport.	**3.25**
	5	I wanted to eat noodles, and I eat.	**3.00**

[5] Kenji Sugiki and Shigeki Matsubara. 2007. A product retrieval system robust to subjective queries. In *2007 2nd International Conference on Digital Information Management*, Vol. 1. 351–356. https://doi.org/10.1109/ICDIM.2007.4444248

[6] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, II–1188–II–1196. http://dl.acm.org/citation.cfm?id=3044805.3045025

[7] Vivek Kumar Singh, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*. 712–717. https://doi.org/10.1109/iMac4s.2013.6526500

[8] Jiaxing Tan, Alexander Kotov, Rojiar Pir Mohammadiani, and Yumei Huo. 2017. Sentence Retrieval with Sentiment-specific Topical Anchoring for Review Summarization. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 2323–2326. https://doi.org/10.1145/3132847.3133153

[9] Lap Q. Trieu, Huy Q. Tran, and Minh-Triet Tran. 2017. News Classification from Social Media Using Twitter-based Doc2Vec Model and Automatic Query Expansion. In *Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT 2017)*. ACM, New York, NY, USA, 460–467. https://doi.org/10.1145/3155133.3155206

[10] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Mix 'N Match: Integrating Text Matching and Product Substitutability Within Product Search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 1373–1382. https://doi.org/10.1145/3269206.3271668

[11] Libing Wu, Cong Quan, Chenliang Li, and Donghong Ji. 2018. PARL: Let Strangers Speak Out What You Like. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 677–686. https://doi.org/10.1145/3269206.3271695

[12] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie Review Mining and Summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*. ACM, New York, NY, USA, 43–50. https://doi.org/10.1145/1183614.1183625

[13] Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic Modeling of Short Texts: A Pseudo-Document View. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 2105–2114. https://doi.org/10.1145/2939672.2939880