A Latent Topic Extracting Method based on Events in a Document and its Application

Risa Kitajima
Ochanomizu University
kitajima.risa@is.ocha.ac.jp

Ichiro Kobayashi
Ochanomizu University
koba@is.ocha.ac.jp

Abstract

Recently, several latent topic analysis methods such as LSI, pLSI, and LDA have been widely used for text analysis. However, those methods basically assign topics to words, but do not account for the events in a document. With this background, in this paper, we propose a latent topic extracting method which assigns topics to events. We also show that our proposed method is useful to generate a document summary based on a latent topic.

1 Introduction

Recently, several latent topic analysis methods such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Probabilistic LSI (pLSI) (Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been widely used for text analysis. However, those methods basically assign topics to words, but do not account for the events in a document. Here, we define a unit of informing the content of document at the level of sentence as an "Event" [1], and propose a model that treats a document as a set of Events. We use LDA as a latent topic analysis method, and assign topics to Events in a document. To examine our proposed method's performance on extracting latent topics from a document, we compare the accuracy of our method to that of the conventional methods through a common document retrieval task. Furthermore, as an application of our method, we apply it to a query-biased document summarization (Tombros and Sanderson, 1998; Okumura and Mochizuki, 2000; Berger and Mittal, 2000) to verify that the method is useful for various applications.

2 Related Studies

Suzuki et al. (2010) proposed a flexible latent topics inference in which topics are assigned to phrases in a document. Matsumoto et al. (2005) showed that the accuracy of document classification will be improved by introducing a feature dealing with the dependency relationships among words.

In case of assigning topics to words, it is likely that two documents, which have the same word frequency in themselves, tend to be estimated as they have the same topic probablistic distribution without considering the dependency relation among words. However, there are many cases where the relationship among words is regarded as more important rather than the frequency of words as the feature identifying the topics of a document. For example, in case of classifying opinions to objects in a document, we have to identify what sort of opinion is assigned to the target objects, therefore, we have to focus on the relationship among words in a sentence, not only on the frequent words appeared in a document. For this reason, we propose a method to assign topics to Events instead of words.

As for studies on document summarization, there are various methods, such as the method based on word frequency (Luhn, 1958; Nenkova and Vanderwende, 2005), and the method based on a graph (Radev, 2004; Wan and Yang, 2006). Moreover, several methods using a latent topic model have been proposed (Bing et al., 2005; Arora and Ravin-

[1] For the definition of an Event, see Section 3.

Proceedings of the ACL-HLT 2011 Student Session, pages 30–35,
Portland, OR, USA 19-24 June 2011.

dran, 2008; Bhandari et al., 2008; Henning, 2009; Haghighi and Vanderwende, 2009). In those studies, the methods estimate a topic distribution on each sentence in the same way as the latent semantic analysis methods normally do that on each document, and generate a summary based on the distribution. We also show that our proposed method is useful for the document summarization based on extracting latent topics from sentences.

3 Topic Extraction based on Events

In this study, since we deal with a document as a set of Events, we extract Events from each document; define some of the extracted Events as the index terms for the whole objective documents; and then make an Event-by-document matrix consisting of the frequency of Events to the documents. A latent topic distribution is estimated based on this matrix.

3.1 Definition of an Event

In this study, we define a pair of words in dependent relation which meets the following conditions: (Subject, Predicate) or (Predicate1, Predicate2) , as an Event. A noun and unknown words correspond to Subject, while a verb, adjective and adjective verb correspond to Predicate. To extract these pairs, we analyze the dependency structure of sentences in a document by a Japanese dependency structure analyzer, CaboCha [2]. The reason why we define (Predicete1, Predicate2) as an Event is because we recognized the necessity of such type of an Event by investigating the extracted pairs of words and comparing them with the content of the target document in preliminary experiments, and could not extract any Event in case of extracting an Event from the sentences without subject.

3.2 Making an Event-by-Document Matrix

In making a word-by-document matrix, high-frequent words appeared in any documents, and extremely infrequent words are usually not included in the matrix. In our method, high-frequent Events like the former case were not observed in preliminary experiments. We think the reason for this is because an Event, a pair of words, can be more meaningful than a single word, therefore, an Event is particularly a good feature to express the meaning of a document. Meanwhile, the average number of Events per sentence is 4.90, while the average number of words per sentence is 8.93. A lot of infrequent Events were observed in the experiments because of the nature of an Event, i.e., a pair of words. This means that the same process of making a word-by-document matrix cannot be applied to making an Event-by-document matrix because the nature of an Event as a feature expressing a document is different from that of a word. In concrete, if the events, which once appear in documents, would be removed from the candidates to be a part of a document vector, there might be a case where the constructed document vector does not reflect the content of the original documents. Considering this, in order to make the constructed document vector reflect the content of the original documents, we do not remove the Event only itself extracted from a sentence, even though it appears only once in a document.

3.3 Estimating a Topic Distribution

After making an Event-by-document matrix, a latent topic distribution of each Event is estimated by means of Latent Dirichlet Allocation. Latent Dirichlet Allocation is a generative probabilistic model that allows multiple topics to occur in a document, and gets the topic distribution based on the idea that each topic emerges in a document based on a certain probability. Each topic is expressed as a multinomial distribution of words.

In this study, since a topic is assigned to an Event, each topic is expressed as a multinomial distribution of Events. As a method to estimate a topic distribution, while a variational Bayes method (Blei et al., 2003) and its application (Teh et al., 2006) have been proposed, in this study we use Gibbs sampling method (Grififths and Steyvers, 2004). Furthermore, we define a sum of topic distributions of the events in a query as the topic distribution of the query.

4 Performance Evaluation Experiment

Through a common document retrieval task, we compare our method with the conventional method and evaluate both of them. In concrete, we regard the documents which have a similar topic distribu-

[2] http://chasen.org/ taku/software/cabocha/

tion to a query's topic distribution as the result of retrieval, and then examine whether or not the estimated topic distribution can represent the latent semantics of each document based on the accuracy of retrieval results. Henceforth, we call the conventional word-based LDA "wordLDA" and our proposed event-based LDA "eventLDA".

4.1 Measures for Topic Distribution

As measures for identifying the similarity of topic distribution, we adopt Kullback-Leibler Divergence (Kullback and Leibler, 1951), Symmetric Kullback-Leibler Divergence (Kullback and Leibler, 1951), Jensen-Shannon Divergence (Lin, 2002), and cosine similarity. As for wordLDA, Henning (2009) has reported that Jensen-Shannon Divergence shows the best performance among the above measures in terms of estimating the similarity between two sentences. We also compare the performance of the above measures when using eventLDA.

4.2 Experimental Settings

As for the documents used in the experiment, we use a set of data including users' reviews and their evaluations for hotels and their facilities, provided by Rakuten Travel[3]. Each review has five-grade evaluations of a hotel's facilities such as room, location, and so on. Since the data hold the relationships between objects and their evaluations, therefore, it is said that they are appropriate for the performance evaluation of our method because the relationship is usually expressed in a pair of words, i.e., an Event. The query we used in the experiment was "a room is good". The total number of documents is 2000, consisting of 1000 documents randomly selected from the users' reviews whose evaluation for "a room" is 1 (bad) and 1000 documents randomly selected from the reviews whose evaluation is 5 (good). The latter 1000 documents are regarded as the objective documents in retrieval. Because of this experiment design, it is clear that the random choice for retrieving "good" vs. "bad" is 50%. As for the evaluation measure, we adopt 11-point interpolated average precision.

In this experiment, a comparison between the both methods, i.e., wordLDA and eventLDA, is conducted from the viewpoints of the proper number of topics and the most useful measure to estimate similarity. At first, we use Jensen-Shannon Divergence as the measure to estimate the similarity of topic distribution, changing the number of topics k in the following, $k = 5$, $k = 10$, $k = 20$, $k = 50$, $k = 100$, and $k = 200$. Next, the number of topics is fixed based on the result of the first process, and then it is decided which measure is the most useful by applying each measure to estimate the similarity of topic distributions. Here, the iteration count of Gibbs Sampling is 200. The number of trials is 20, and all trials are averaged. The same experiment is conducted for wordLDA to compare both results.

4.3 Result

Table 1 shows the retrieval result examined by 11-point interpolated average precision, changing the number of topics k. High accuracy is shown at $k = 5$ in eventLDA, and $k = 50$ in wordLDA, respectively. Overall, we see that eventLDA keeps higher accuracy than wordLDA.

number of topics	wordLDA	eventLDA
5	0.5152	0.6256
10	0.5473	0.5744
20	0.5649	0.5874
50	0.5767	0.5740
100	0.5474	0.5783
200	0.5392	0.5870

Table 1: Result based on the number of topics.

Table 2 shows the retrieval result examined by 11-point interpolated average precision under various measures. The number of topics k is $k = 50$ in wordLDA and $k = 5$ in eventLDA respectively, based on the above result. Under any measures, we see that eventLDA keeps higher accuracy than wordLDA.

similarity measure	wordLDA	eventLDA
Kullback-Leibler	0.5009	0.5056
Symmetric Kullback-Leibler	0.5695	0.6762
Jensen-Shannon	0.5753	0.6754
cosine	0.5684	0.6859

Table 2: Performance under various measures.

4.4 Discussions

The result of the experiment shows that eventLDA provides a better performance than wordLDA, there-

[3]http://travel.rakuten.co.jp/

fore, we see our method can properly treat the latent topics of a document. In addition, as for a property of eventLDA, we see that it can provide detail classification with a small number of topics. As the reason for this, we think that a topic distribution on a feature is narrowed down to some extent by using an Event as the feature instead of a word, and then as a result, the possibility of generating error topics decreased.

On the other hand, a proper measure for our method is identified as cosine similarity, although cosine similarity is not a measure to estimate probabilistic distribution. It is unexpected that the measures proper to estimate probabilistic distribution got the result of lower performance than cosine similarity. From this, there are some space where we need to examine the characteristics of topic distribution as a probabilistic distribution.

5 Application to Summarization

Here, we show multi-document summarization as an application of our proposed method. We make a query-biased summary, and show the effectiveness of our method by comparing the accuracy of a generated summary by our method with that of summaries by the representative summarization methods often used as benchmark methods to compare.

5.1 Extracting Sentences by MMR-MD

In extracting important sentences, considering only similarity to a given query, we may generate a redundant summary. To avoid this problem, a measure, MMR-MD (Maximal Marginal Relevance Multi-Document), was proposed (Goldstein et al., 2000). This measure is the one which prevents extracting similar sentences by providing penalty score that corresponds to similarity between a newly extracted sentence and the previously extracted sentences. It is defined by Eq. 1 (Okumura and Nanba, 2005).

$$MMR\text{-}MD \equiv argmax_{C_i \in R \setminus S}[\lambda Sim_1(C_i,Q) - (1-\lambda) max_{C_j \in S} Sim_2(C_i,C_j)] \quad (1)$$

We aim to choose sentences whose content is similar to query's content based on a latent topic, while reducing the redundancy of choosing similar sentences to the previously chosen sentences. Therefore, we adopt the similarity of topic distributions for Sim_1 which estimates similarity between a sentence and a query, and adopt cosine similarity based on Events as a feature unit for Sim_2 which estimates the similarity with the sentences previously chosen. As the measures to estimate topic distribution similarity, we use the four measures explained in Section 4.1. Here, as for the weighting parameter λ, we set $\lambda = 0.5$.

C_i : sentence in the document sets
Q : query
R : a set of sentences retrieved by Q from the document sets
S : a set of sentences in R already extracted
λ : weighting parameter

5.2 Experimental Settings

In the experiment, we use a data set provided at NTCIR4 (NII Test Collection for IR Systems 4) TSC3 (Text Summarization Challenge 3) [4] .

The data consists of 30 topic sets of documents in which each set has about 10 Japanese newspaper articles, and the total number of the sentences in the data is 3587. In order to make evaluation for the result provided by our method easier, we compile a set of questions, provided by the data sets for evaluating the result of summarization, as a query, and then use it as a query for query-biased summarization. As an evaluation method, we adopt precision and coverage used at TSC3 (Hirao et al., 2004), and the number of extracted sentences is the same as used in TSC3. Precision is an evaluation measure which indicates the ratio of the number of correct sentences to that of the sentences generated by the system. Coverage is an evaluation measure which indicates the degree of how the system output is close to the summary generated by a human, taking account of the redundancy.

Moreover, to examine the characteristics of the proposed method, we compare both methods in terms of the number of topics and the proper measure to estimate similarity. The number of trials is 20 at each condition. 5 sets of documents selected at random from 30 sets of documents are used in the trials, and all the trials are totally averaged. As a target for comparison with the proposed method, we also conduct an experiment using wordLDA.

[4] http://research.nii.ac.jp/ntcir/index-en.html

5.3 Result

As a result, there is no difference among the four measures — the same result is obtained by the four measures. Table 3 shows comparison between eventLDA and wordLDA in terms of precision and coverage. The number of topics providing the highest accuracy is $k = 5$ for wordLDA, and $k = 10$ for eventLDA, respectively.

number of topics	wordLDA		eventLDA	
	Precision	Coverage	Precision	Coverage
5	0.314	0.249	0.404	0.323
10	0.264	0.211	0.418	0.340
20	0.261	0.183	0.413	0.325
50	0.253	0.171	0.392	0.319

Table 3: Comparison of the number of topics.

Furthermore, Table 4 shows comparison between the proposed method and representative summarization methods which do not deal with latent topics. As representative summarization methods to compare our method, we took up the Lead method (Brandow et al., 1995) which is effective for document sumarization of newspapers, and the important sentence extraction-based summarization method using TF-IDF.

method	Precision	Coverage
Lead	0.426	0.212
TF-IDF	0.454	0.305
wordLDA (k=5)	0.314	0.249
eventLDA (k=10)	0.418	0.340

Table 4: Comparison of each method.

5.4 Discussions

Under any condition, eventLDA provides a higher accuracy than wordLDA. We see that the proposed method is useful for estimating a topic on a sentence. As the reason for that the accuracy does not depend on any kinds of similarity measures, we think that an estimated topic distribution is biased to a particular topic, therefore, there was not any influence due to the kinds of similarity measures. Moreover, the proper number of topics of eventLDA is bigger than that of wordLDA. We consider the reason for this is because we used newspaper articles as the objective documents, so it can be thought that the topics onto the words in the articles were specific to some extent; in other words, the words often used in a particular field are often used in newspaper articles, therefore, we think that wordLDA can classify the documents with the small number of topics. In comparison with the representative methods, the proposed method takes close accuracy to their accuracy, therefore, we see that the performance of our method is at the same level as those representative methods which directly deal with words in documents. In particular, as for coverage, our method shows high accuracy. We think the reason for this is because a comprehensive summary was made by latent topics.

6 Conclusion

In this paper, we have defined a pair of words with dependency relationship as "Event" and proposed a latent topic extracting method in which the content of a document is comprehended by assigning latent topics onto Events. We have examined the ability of our proposed method in Section 4, and as its application, we have shown a document summarization using the proposed method in Section 5. We have shown that eventLDA has higher ability than wordLDA in terms of estimating a topic distribution on even a sentence or a document; furthermore, even in case of assigning a topic on an Event, we see that latent topics can be properly estimated. Since an Event can hold a relationship between a pair of words, it can be said that our proposed method, i.e., eventLDA, can comprehend the content of a document more deeper and proper than the conventional method, i.e., wordLDA. Therefore, eventLDA can be effectively applied to various document data sets rather than wordLDA can be. We have also shown that another feature other than a word, i.e., an Event is also useful to estimate latent topics in a document. As future works, we will conduct experiments with various types of data and query, and further investigate the characteristic of our proposed method.

Acknowledgments

We would like to thank Rakuten, Inc. for permission to use the resources of Rakuten Travel, and thank the National Institute of Informatics for providing NTCIR data sets.

References

Adam Berger and Vibhu O. Mittal. 2000. Query-relevant summarization using FAQs. In *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*:294–301.

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*:2–10.

Ani Nenkova and Lucy Vanderwende. 2005. The Impact of Frequency on Summarization. Technical report, Microsoft Research.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-Document Summarization. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*:362–370.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*,3:993–1022.

Dragomir R. Radev. 2004. Lexrank: graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR.*

Harendra Bhandari, Masashi Shimbo, Takahiko Ito, and Yuji Matsumoto. 2008. Generic Text Summarization Using Probabilistic Latent Semantic Indexing. In *Proceedings of the 3rd International Joint Conference on Natural Langugage Proceeding*:133-140.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAALP-ANLP Workshop on Automatic Summarization*:40–48.

Jianhua Lin. 2002. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Leonhard Henning. 2009. Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis. *Recent Advances in Natural Language Processing*:144–149.

Manabu Okumura and Eiji Nanba. 2005. *Science of knowledge: Automatic Text Summarization.*(in Japanese) ohmsha.

Manabu Okumura and Hajime Mochizuki. 2000. Query-Biased Summarization Based on Lexical Chaining. *Computational Intelligence*,16(4):578–585.

Qin Bing, Liu Ting, Zhang Yu, and Li Sheng. 2005. Research on Multi-Document Summarization Based on Latent Semantic Indexing. *Journal of Harbin Institute of Technology*,12(1):91–94.

Rachit Arora and Balaraman Ravindran. 2008. Latent dirichlet allocation based multi-document summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data.*

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management: an International Journal - Special issue: summarizing text*,31(5):675–685.

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 of Information Science*, 41(6):391–407.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees. In *Proceedings of the 9th Pacific-Asia International Conference on Knowledge Discovery and Data Mining*:301–310.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *Annuals of Mathematical Statistics*, 22:49–86.

Thomas L. Grififths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*,101:5228–5235.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*:50–57.

Tsutomu Hirao, Takahiro Fukusima, Manabu Okumura, Chikashi Nobata, and Hidetsugu Nanba. 2004. Corpus and evaluation measures for multiple document summarization with multiple sources. In *Proceedings of the 20th International Conference on Computational Linguistics*:535–541.

Xiaojun Wan and Jianwu Yang. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*

Yasuhiro Suzuki, Takashi Uemura, Takuya Kida, and Hiroki Arimura. 2010. Extension to word phrase on latent dirichlet allocation. Forum on Data Engineering and Information Management,i-6.

Yee W. Teh, David Newman, and Max Welling. 2006. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems Conference*,19:1353–1360.