EPiC Series in Computing

Volume 81, 2022, Pages 41–52

Proceedings of 11th International Congress
on Advanced Applied Informatics

Automatic extractive summarization for Japanese documents by LDA

Hideyuki Sawahata and Tetsuro Nishino

Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu, Tokyo, Japan

Abstract

The demand for automatic summarization of newspaper headlines and article summaries has increasing with various studies on automatic summarization being currently conducted. However, there are only a few studies on Japanese documents as compared English documents.

In this paper, wheter existing summarization methods can be effective for academic papers written in Japanese is verified. First, we demonstrate the effectiveness of topic-based extractive summarization methods Latent Semantic Analysis (LSA). Then, a more effective topic-based extractive summarization is possible by using Latent Dirichlet Allocation (LDA) is demonstrated.

Keywords— Natural Language Processing, Automatic Summarization, Extractive summarization, LDA, LSA

1 Introduction

1.1 Background

Currently, the demands for automatic summary generation can widely vary from automatic generation of headlines and summaries of newspapers to automatic generation of abstracts of academic papers. In addition, document summarization is required in many situations, such as summarization of business books and novels.

Through numerous studies on English summarization, automatic summarization methods have been developed using both supervised and unsupervised learning.[1, 2] For example, supervised summarization methods are using neural networks, such as the encoder-decoder model[3]. Recently, a pre-training model for automatic summarization called PEGASUS[4] was developed.

Unsupervised learning mainly consists of graph-based and topic-based methods in English summarization. LexRank[5] is an example of graph-based summarization. This method uses degree centrality for automatic summarization. Latent Semantic Analysis (LSA)[6] is an example of topic-based methods. Methods using LSA use topic data and singular values obtained from LSA to generate a summary. Various LSA summarization methods depends on how these two types of information are used to generate a summary.

As aforementioned above, there have been many studies on the automatic generation of summaries. However, most of the studies have been conducted for English summaries, and only a few of them have been conducted on documents written in Japanese. In particular, there are very few studies on academic papers written in Japanese.

1.2 Purpose

Based on the above-mentioned background, the purpose of this study is to verify whether existing automatic summarization methods are effective for Japanese papers. However, there are only a few Japanese corpora for training summarization. Therefore, in this study, we focused on topic-based extractive summarization methods, that are capable of unsupervised auto-summarization. Especially, we focused on an extractive automatic summarization method using Latent Semantic Analysis (LSA), which has been suggested to work for multiple languages[7].

In this study, we first verified whether the extractive summarization method based on LSA was effective for Japanese language. In addition, we verify whether using Latent Dirichlet Allocation (LDA)[8], the same topic model as LSA that improves the flexibility of topic representation by using a probabilistic model, is more effective than LSA.

2 Related works

2.1 LSA

Latent Semantic Analysis (LSA) is the first method used in statistical latent semantic analysis. LSA uses singular value decomposition to extract co-occurrence of words, including latent ones. LSA is used in clustering of words and calculating the similarity between documents.

In LSA, words are extracted from documents by morphological analysis and a word-document matrix M is generated with words in rows and documents in columns. After obtaining word-document matrix M, we perform a singular value decomposition, as in Equation 1:

$$M = U\Sigma V^T \tag{1}$$

We call the following for the obtained U, Σ, V^T.

U Left singular value matrix

Σ Singular value matrix

V^T Right singular value matrix

The three matrices obtained can be used for various applications in LSA.

2.2 LDA

Latent Dirichlet Allocation is statistical latent semantic analysis method similar to LSA and a probabilistic model. LDA defines documents as a collection of words and assumes that words are generated by topics, and that document is created as a collection of words. Specifically, LDA sets up the process of document generation in the following order:

1. Number of words N is decided by Poisson distribution.

2. Determine the parameter θ of the topic distribution from the Dirichlet distribution $Dir(\alpha)$ with α as a parameter

Table 1: Example of matrix X

	s_1	s_2	s_3
He	1	1	1
has	1	0	0
dog	1	1	0
walked	0	1	0
went	0	0	1
park	0	0	1

3. For each N words w_n
   - (a) Determine the latent topic z_n for w_n from the multinomial distribution $Multi(\theta)$
   - (b) Determine the word w_n from the multinomial conditional probability $p(w_n|z_n, \beta)$ on the latent topic z_n

The parameter β in the above generation process is a matrix representing the selection probability of a word for each topic, and each element represents the probability that a word appears in a certain topic.

According to the above generation process, we define the probability distribution and learn the parameters based on it.

2.3 Extractive summarization method using LSA

2.3.1 LSA in extractive summarization

There are methods that use LSA in extractive summarization methods. This method extracts topic from each sentence in the document to be summarized and decides which sentence to select as the summary sentence based on the topic. In the extractive summarization method based on LSA, matrix X is generated for the document summary, whose elements are the number of occurrences of each word in each sentence. Suppose we have a sentence such as the following.

s_1 He has a dog.

s_2 He walked with the dog.

s_3 He went to the park.

We extract only nouns and verbs from these matrices. Matrix X with words in rows and sentences in columns is shown in Table 1.

For matrix X, LSA extractive summarization applies singular value decomposition. When singular value decomposition is used, matrix X is decomposed into three matrices as shown in the equation 1.

Extractive summarization using LSA is often performed using mainly Σ and V. As V corresponds to sentences in rows and topics in columns, V is often used as a summary in many methods. Therefore, V is used as a criterion for selecting sentences as summary sentences in many methods because sentences correspond to rows and topics to columns. For selecting summary sentences using topics, it is possible to consider the potential co-occurrence of words among sentences in the selection, and thus it is possible to select summary sentences considering the whole document.

Table 2: Example of right singular matrix

	topic0	topic1
sent1	0.457	-0.77
sent2	0.728	0.037
sent3	0.51	0.637

Table 3: Example of singular value matrix

topic0	5.033
topic1	1.574

In LSA, the singular value matrix is used as a criterion to determine the topic to focus on. Extractive summarization methods based on LSA can be divided into several methods depending on the management of the right singular value matrix V and the singular value matrix Σ described above. In this study, we introduce the five methods. Tables 2 and 3, obtained by singular value decomposition from Table 1 are used in the description of these methods.

2.3.2 Gong & Liu's method

Gong & Liu's method[9] is one of the extractive summarization methods using LSA. In this method, for the right singular value matrix obtained by LSA, sentences are selected up to a default upper limit by the following procedure, and a summary is generated.

1. Generate a word-document matrix for the target document, where each sentence is a document.
2. Obtain the right singular value matrix of the word-document matrix by using LSA.
3. For the right singular value matrix, select the sentences with the highest value in each column from a row, starting from the top of the row, until the default number of sentences is satisfied.
4. The selected sentences are determined as the summary sentences.

In this method, the higher the value of the right singular matrix, the better the sentence represents each topic. Therefore, we assume that from each topic, we can summarize the entire document by selecting the sentence with the highest value.

In Table 2, we first select a sentence from topic0. The sentence with the highest value in the row of topic0 is sent2, so we select sent2 as the summary sentence. Next, we select the sentence with the highest value in the row of topic1, sent3, as the summary sentence, and output these two sentences as a summary.

2.3.3 Steinberger & Jezek's method

Steinberger & Jezek's method[10] uses the right singular matrix and singular matrix to select sentences used for summary. In this method, the sentence's "length" in document is calculated

Table 4: ”length” of each sentences

	length
sent1	1.043
sent2	1.93
sent3	1.89

Table 5: Percentages of each topic’s singular

topic0	0.76
topic1	0.24

using Eq. 2.

$$length_i = \sqrt{\sum_j V_{ij}^T \Sigma_{jj}} \tag{2}$$

As you can be seen from Equation 2, entences for which both the value of the singular value matrix and the value of the right singular value matrix are large are more likely to be selected. In other words, the more the topic of the document matches the topic of the sentence, the more likely that sentence will be selected.

Using Equation 2, the length of each sentence was calculated from Tables 2 and 3 as shown in Table 4. In Table 4, when the summary consists of two sentences, sent2 and sent3 are selected as summary sentences.

2.3.4 Murray et al’s method

Murray et al.’s method[11] selects sentences for summarization from each topic as in Gong & Liu’s method, however the difference is that the number of sentences that can be selected in a topic is determined by the ratio of the singular values in each topic to the sum of all singular values. In this method, we focus on the main topic of the document and select sentences so that sentences that are likely to be the main topic appear more often in the summary. For example, in Table 3, the rough percentages for each topic are listed as in Table 5.

If the summary consists of three sentences, we select two sentences for the summary by rounding $3 \times 0.76 = 2.28$ from topic0.

2.3.5 Cross method

Cross method[12] is similar to that of Steinberger & Jezek’s method. For the right singular value matrix, we calculate the average value for each topic column; and for each column, we set the value of the element to be less than the average value of the column to 0. The selection trend is the same as Steinberger & Jezek’s method, but modified so that it is not influenced by topics that do not match. In the case of Table 2, the average of each column in Table 6.

From this result, for each column, if we assume that the mean of the column is less than zero, we obtain the following results as shown in Table 7.

Table 6: Average of each columns

	topic0	topic1
sent1	0.457	-0.77
sent2	0.728	0.037
sent3	0.51	0.637
Avg.	0.565	-0.032

Table 7: Set less than average to 0 for each column

	topic0	topic1
sent1	0	0
sent2	0.728	0.037
sent3	0	0.637
Avg.	0.565	-0.032

Subsequently, the sum of each column was computed, and the length of the sentence is calculated. In the case of Table 7, we obtain Table 8, and the first sentence selected as the summary is sent2.

2.3.6 Topic method

Similar to the Cross method, the Topic method[12] sets each row of the right singular value matrix to 0 if the element is less than the mean value of that row. After setting, "strength" *strength* between each topic is calculated, and a topic-topic matrix is generated, which shows the strength between each topic. The $strength_{ik}$ of a topic i and topic k is derived from Equation 3.

$$strength_{ik} = \begin{cases} \sum_j V^T_{ij} & (i = k) \\ \sum_j \left(V^T_{kj} + V^T_{ij}\right) & (i \neq k) \end{cases} \tag{3}$$

However, if $V^T_{kj} = 0$ or $V^T_{ij} = 0$, then $V^T_{kj} + V^T_{ij} = 0$.

The above formula, in other words, shows the policy of selecting sentences from topics that are likely to co-occur with other topics in each document. In case of Table8, it looks in Table9.

After calculating the strength between each topic, we compute the sum of each row and select sentences in the same way as Gong & Liu's method in order of the topic with the highest value.

Table 8: Each sentences' "length" in Cross method

	topic0	topic1	length
sent1	0	0	0
sent2	0.728	0.037	0.765
sent3	0	0.637	0.637

Table 9: topic-topic matrix

	topic0	topic1
topic0	0.728	0.765
topic1	0.765	0.674

Table 10: Sum of each topics' strength

	topic0	topic1	sum
topic0	0.728	0.765	1.533
topic1	0.765	0.674	1.439

From Table 9, we obtained the results shown in Table 10. In this case, the selection was made from topic0.

2.4 ROUGE

ROUGE[13] evaluates the words that match reference and generated summaries. Recall and Precision are calculated in ROUGE, and their harmonic mean is used to evaluate the generated summary. In other words, the $score_{f1}$ for the generated summary can be derived from the following equation, where $score_{recall}$ is the recall and $score_{precision}$ is the precision.

$$score_{f1} = \frac{2(score_{recall} \times score_{precision})}{score_{recall} + score_{precision}} \tag{4}$$

In ROUGE, there are several variants that depends on the method of calculating precision and recall. We introduce ROUGE-L. ROUGE-L calculates the recall and the precision based on the ”Longest Common Subsequence” (LCS). In this case, the LCS is defined as the sequence of words that are completely consistent with each other. It is assumed that there is no problem even if the words in between are different, as long as the sequence matches.

The recall and precision of ROUGE-L can be obtained using the following equations.

$$score_{recall} = \frac{LCS(Ref_{words}, Sys_{words})}{|Ref_{words}|} \tag{5}$$

$$score_{precision} = \frac{LCS(Ref_{words}, Sys_{words})}{|Sys_{words}|} \tag{6}$$

LCS indicates the number of words in the LCS, and Ref_{words} and Sys_{words} indicate the number of words in the reference summary and the generated summary, respectively.

3 Experiments

3.1 Experimental methods

In this study, we first investigated the effectiveness of the extractive summarization method using LSA for Japanese sentences, and then we investigated the effectiveness of using LDA as a feature of sentences. In other words, we conducted two experiments.

Experiment 1 Evaluation of the applicability of the extractive summarization method using LSA in Japanese.

Experiment 2 Evaluation of the effectiveness of LDA for extractive summarization for Japanese.

In each experiment, we used five extractive summarization methods.

- Gong & Liu's method
- Steinberger & Jezek's method
- Murray et al's method
- Cross method
- Topic method

We used five methods for experiment because we avoid that it is difficult to judge whether a particular method is effective or not when only one method is used.

ROUGE-L was used as the evaluation method in the verification experiments. The summaries generated by the extractive summarization method were evaluated using the manual summaries set up in each experiment.

Throughout the experiment, we set a limit on the number of sentences to be generated to ”10% of the input sentences”. In addition, stop words were used to remove some words.

3.2 Experiment 1

3.2.1 Overview

Extractive summarization using LSA is stated to be effective not only for English but also for other languages such as Turkish. Therefore, we investigated whether it is possible to achieve the same performance in Japanese. For this verification, we compared the results in English with those results in Japanese. For English, we quoted the evaluation results for the Summac dataset from [7]. Summac is a collection of summaries of articles on computer science in ACL-sponsored conferences. We used the corpus of the Journal of the Association for Natural Language Processing [14] for the evaluation of automatic summarization in Japanese.

In this experiment, the main text was set as the target sentence for the summary, and the following items in the text were deleted.

- Equation, Figures and tables in the text
- Ornamental descriptions such as bf
- Various environments except for the document enclosed by begin and end

The abstract described in the paper was treated as a refference summary.

3.2.2 Results

Table 11 shows the results of the extractive summarization by LSA for English and Japanese. Owing to the difference in the target corpora, the scores of Japanese and English are different for each method, however the scores are relatively close. However, Steinberger & Jezek's method showed a large difference.

Table 11: Score in English and Score in Japanese

	English	Japanese
Gong & Liu's method	0.180	0.181
Steinberger & Jezek's method	0.138	0.193
Murray et al' method	0.180	0.184
Cross method	0.182	0.200
Topic method	0.180	0.190

Table 12: Comparison of LSA and LDA

	LSA	LDA
Gong & Liu's method	0.181	0.197
Steinberger & Jezek's method	0.193	0.208
Murray et al' method	0.184	0.202
Cross method	0.200	0.190
Topic method	0.190	0.199

3.2.3 Observation

In Experiment 1, we checked whether LSA's extractive summarization performs as well in Japanese as it does in English. IIn this experiment, some words were removed as stop words. This removal of words based on frequency of occurrence and part-of-speech is thought to have enabled extractive summarization in Japanese as well as in English. The results shown in Table 11 confirmed that the scores of several methods were relatively close. The slight difference in the scores was probably due to the difference in the target corpora.

The reason why the performance in Japanese was almost the same as that in English is that LSA extracts topics based only on the co-occurrence of words, without considering the order of words, which is a major difference between English and Japanese. Therefore, a method such as LSA, which does not consider word order, can be used regardless of the language used.

3.3 Experiment 2

3.3.1 Overview

As we were able to confirm that extractive summarization using LSA is applicable to Japanese, we next examined whether LDA was more effective than LSA in extractive summarization. The corpus, extractive summarization method used for comparison, and the limitation on the generated summaries were unchanged from Experiment 1. We replaced LSA with LDA in the extractive summarization method used for comparison.

3.3.2 Results

Table 12 presented the results of this study. It was confirmed that the score increased by replacing the topic extraction method with LDA in various extractive summarization methods, while it was confirmed that the score decreased in the Cross method.

3.3.3 Observation

In Experiment 2, we examined whether LDA was effective for extractive summarization methods. As a result, the score of LSA is higher than that of LDA only for the Cross method, whereas that of LDA is higher for the other methods. This may be because of the algorithm of the Cross method. In the Cross method, as described in Section 2.3.5, there is a process before calculating the "length" of a sentence. In this process, for each column of the topic-document matrix, the elements below the average of the row were set to 0. In addition, when calculating the length, the sum of each column is calculated; however, in LDA, the sum of topic distributions is 1. Therefore, setting less than the mean to 0 does not have a significant effect, and the reason may be that the calculation does not consider the topic distribution. However, the reason of LDA score being higher than that of the other methods is probably due to the fact that the obtained values are probability values. Since negative values appear in LSA, it is assumed that some important sentences may be judged as unimportant because of the influence of the negative values.

4 Discussion

The following is a conclusion of the above experiments.

- Topic-based extractive summarization by LSA works as well for Japanese as for English.

- LDA is more effective than LSA for topic-based automatic summarization.

The performance of the extractive summarization by LSA in Japanese was verified and confirmed in Section 3.2 of this study. Furthermore, in Section 3.3, we were confirmed the improvement by using LDA. Thus, it is assumed that the topic-based extractive summarization method may be applicable for various languages. In addition, as mentioned in Section 2, LSA and LDA are unsupervised methods that do not require correct answer data for training. This can be an advantage in summarize documents with a small corpus of language.

The reason why LDA was more effective than LSA seems to be that LDA is a topic model as a probabilistic model. In LSA, the values obtained from singular value decomposition include both positive values and negative values.

However, in LDA, basically probability values are obtained, which are greater than 0 and less than 1. Therefore, in a method that calculates sums and products, a value close to 0 is less likely to affect the calculation, and it making it easier to select important sentences. However, as only values between 0 and 1 are obtain, methods such as the Cross method, in which some values are set to 0 based on the average, may not be effective because of their small value. In other words, although LDA is effective in extractive summarization methods, it may not be effective in some cases.

In the experiments in this paper, we compared several extractive summarization methods for LDA and LSA respectively. According to the results of Experiment 2, the Cross method is the best method for extractive summarization in Japanese, when LSA is used. On the other hand, the best extractive summarization method for LDA is Steinberger & Jezek's method. However, the experimental results of LDA do not show any significant difference from the other methods, so all methods are considered to be effective.

5 Conclusions

In this study, we focused on unsupervised extractive auto-summarization, which can work with a small corpus for summarizing Japanese papers, while there is a demand for auto-summarization. However, as most of the automatic summarization methods had been applied to English documents, it was unclear whether they performed equally well in Japanese. Therefore, we focused on the extractive auto-summarization method using LSA, which can work in multiple languages, and verified whether it works for Japanese or not. In addition, we examined whether LDA, which is a statistical latent semantic analysis method similar to LSA, is more effective than LSA.

On several verifications, we found the following.

- An extractive automatic summarization method using LSA works well for Japanese.
- LDA is more effective than LSA in topic-based extractive summarization.

Therefore, the features obtained by unsupervised learning without considering word order, such as LSA and LDA, are effective for many languages. Furthermore, because LDA obtains probability values as features, it seems to ignore unimportant sentences. However, LDA is not effective for Cross method. As a hypothesis, in a method using the average of features such as the Cross method, the difference between the average and the probability value is small, and it is considered that the method may be unable to exclude unimportant sentences and conversely exclude important sentences. Therefore, although LDA is effective, it is necessary to consider the use of LSA depending on the method. In future works, it will be necessary to verify whether the automatic summarization of Japanese papers works with other unsupervised methods as well.

References

[1] Radityo Eko Prasojo, Mouna Kacimi, and Werner Nutt. Modeling and summarizing news events using semantic triples. In *European Semantic Web Conference*, pages 512–527. Springer, 2018.

[2] Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616. Citeseer, 2014.

[3] Parag Jain, Anirban Laha, Karthik Sankaranarayanan, Preksha Nema, Mitesh M. Khapra, and Shreyas Shetty. A mixed hierarchical attention based encoder-decoder approach for standard table summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 622–627, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[4] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR, 13–18 Jul 2020.

[5] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

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

[7] Ilyas Cicekli Makbule Gulcin Ozsoy, Ferda Nur Alpaslan. Text summarization using latent semantic analysis. *Journal of Information Science*, 2011.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, 2001.

[10] Josef Steinberger and Karel Jezek. Using latent semantic analysis in text summarization and summary evaluation. *Proceedings of ISIM'04*, pages 93–100, 01 2004.

[11] Gabriel Murray, Steve Renals, and Jean Carletta. Extractive summarization of meeting recordings. In *in Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596, 2005.

[12] Makbule Ozsoy, Ilyas Cicekli, and Ferda Alpaslan. Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 869–876, 2010.

[13] Chin-Yew Lin. Rouge : A package for automatic evaluation of summaries. *Proc. Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*, 2004.

[14] The Association for Natural Language Processing. LaTeX corpus of the Transactions of the Association for Natural Language Processing. `https://www.anlp.jp/resource/journal_latex/index.html`, 2020.