First Asia Pacific Conference on Advanced Research (APCAR-2015)
ISBN: 978 0 994 3656 68
www.apiar.org.au

SIMILAR DOCUMENT DETERMINATION METHOD BY USING COMPOUND NOUNS

Kyoko Yanagihori & Kazuhiko Tsuda
University of Tsukuba, Japan
Corresponding Authorkyoko@gssm.otsuka.tsukuba.ac.jp

Abstract

In order to distinguish parts of speech, it is necessary to use a technique of morphological analysis to determine documents' similarity for Japanese documents, which are written with no spaces in between the words.

However, depending on the targeted documents, using the morphological method could lead to the conclusion of all the documents being similar to one another since they contain similar words. These results show that it is not necessarily an efficient method to search for similar document determination. This is particularly shown in patent documents, which contain many distinctive nouns and compound nouns that are constructed with connected nouns. These compound nouns in patent documents often describe the invention itself or its method and are very important. These compound nouns should be treated as they are, as meaningful form, without being broken down to morpheme. Although it is efficient to use compound nouns to determine the documents' similarity in order to correctly understand the contents of the document, the drawback to this method is that there is a low rate of finding the same compound nouns in the other documents. In this paper, a method to summarize the similar compound nouns was used to supplement this issue. This method was inspected for its accuracy to see whether or not it was correctly analyzing the similarity of the documents.

Keywords: compound noun, patent documents, similar document search

Introduction

Unlike English, Japanese writing doesn't use spaces in between words. We humans would know, by the experiences we have, where to use punctuation when reading sentences. However, systems cannot tell where to punctuate, so the method to divide words by each part of speech, the technique called 'Wakachigaki (word segmentation)', is used to analyze Japanese writings. This method is called morphological analysis. Morphological analysis is also often used to measure the similarity of documents.

For example, this method uses the frequency of the same words' appearance in documents A and B to calculate the similarity in the two documents.

However, this method doesn't necessary work well for ones such as patent documents, since they use a particular writing system.

On the other hand, compound nouns, which are constructed by a connection of more than two nouns, hold correct information. For instance, if morphological analysis was used for the word 'information processing equipment', it will be divided to 'information', 'processing' and 'equipment'. This will make

ISBN: 978 0 994 3656 68

analysis to be done correctly difficult to identify the other documents, which hold the similar meanings as 'information processing equipment'. If the word was left as the original compound noun, it should be able to identify the document, which holds the meaning of 'information processing equipment'. If the compound noun is being broken down into individual words by morphological analysis, it makes it difficult to correctly understand what the document is written about.

Therefore, in order to solve this problem, a similar document analysis method using compound nouns was used for this research.

Research problem

Patent documents that were written in Japanese were used as an object of this research to compare and determine document similarity. Especially important parts in patent documents, Claims, have some frequently appearing words such as 'Tokucho (character)', 'Souchi (system)', 'Step (step)', 'Shudan (method)' and 'Houhou (technique)'. These words appear in every patent document. For that reason, they are often analyzed as similar documents even though the content they are written about is different. As mentioned in the introduction, using compound nouns will compensate for the shortcoming of drawing a wrong similar document determination result by morphological analysis as shown in Figure 1, and correctly identify the similarity in the documents.

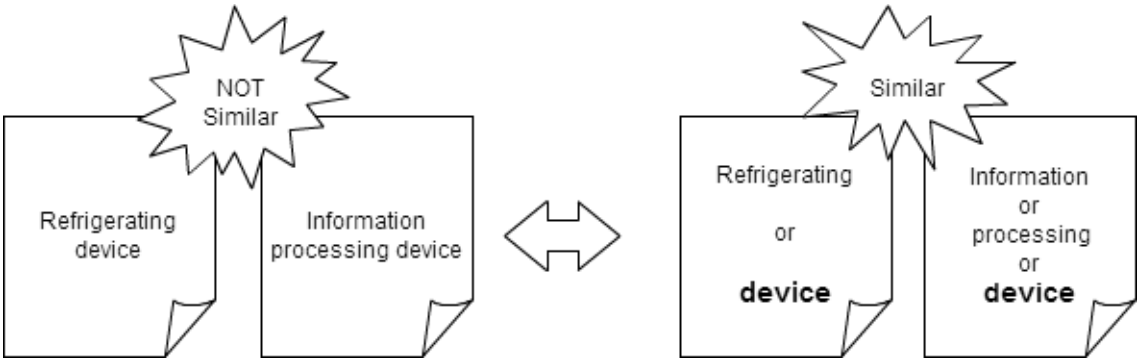

Figure 1 : Device × Device / Information Processing device × Refrigerating device

However, there is a shortcoming in using compound nouns. There is a low rate in finding the same compound nouns in other documents. As a result, the number of documents, which contain the targeted compound nouns, will go down, and it draws the result of low similarity in the documents. To compensate for this shortcoming, this research used the method of gathering all the similar compound nouns to one, even though they are not necessary the same.

Review of the relevant literature

There is research done by Nakagawa et al., which used the method to extract the compound nouns from Japanese documents [1, 2]. This research used the tool '*termmi*' which was created by Nakagawa et al. As a search method that used compound nouns, Hammouda and Kamel performed document clustering using nouns and compound nouns on HTML documents. It disclosed higher similarity results in documents by using compound nouns rather than just using nouns [3].

There is research in Japanese on the method of clustering website search results by Hirao and Takeuch [4]. These researches show that there is significance in using compound nouns for performing searches. Therefore, this paper narrowed the target down to patent documents to perform

verification.

Patent document analysis is also taking place all over the world. This is because the nature of patent documents requires novelty and if there is a similar patent already existing, then it makes the acquisition of rights impossible. Therefore precedent research investigation is essential. However, it is impossible to read each individual document of enormous quantity. That is why the technique to mechanically process natural language is used often in patent document analysis. In order to supplement precedent research investigations, the other methods to classify documents using such things as text mining are also known well [5].

Method

There are 116 documents to be compared individually. It will be a total of 6,670 pairs to be investigated, when they are paired up one to one from document 1 to 116.

Extraction of the targeted documents

Patent application documents, which include 'Biometric Authentication Device' in part of its title. Moreover, the ones, which hold International Patent Classification (IPC) G06F21/20 that, are given for the invention field in patent documents. This is an invention field of security devices for protecting calculators from fraud in the electrical digital data process.

116 documents were selected. The main claim, 'Claim 1', was extracted from 'Claim' section, which specifies the invention rights limit, from the patent documents. Then the common unnecessary words in patent documents such as 'aforesaid', 'the' and 'said' were deleted from those selected documents. These were defined as the extracted documents.

The measurement of the cosine similarity by using morphological analysis

Each of the documents was morphologically analyzed. The parts of speech which correspond to 'nouns' were selected out of those extracted documents, then the cosine similarity was calculated based on the frequency of the noun appearance.

Cosine similarity is a similarity calculation method that is used to compare documents in vector space model. To express the closeness of the angle defined by the each vector, according to the regular cosine of the trigonometric function, closer to 1 means that they are similar and closer to 0 means that they are not similar. Documents' similarity is indicated by the row of the words' appearance frequency. The vector is created according to the words' appearance existence 1 and nonexistence 0.

The cosine similarity measurement by a compound noun analysis

On the other hand, *termmi* was used to extract compound nouns from extracted documents [6]. Each extracted document's cosine similarity was calculated based on those selected compound nouns. This calculation was done the same way as morphological analysis, but changed nouns to compound nouns.

Compound noun symmetrization method

All the 9,223 patent documents which hold IPC G06F21/20 invention field (as of 03/21/2015) were selected as the target documents, and its 'Claim' were extracted. Then the word vectors for compound nouns were extracted, and the close cosine similarity to each was calculated. Theword2vec tool was used for its method [7]. The word2vec tool learns to predict five to ten words that will appear before

and after the word that it took as the basic word by an artificial neural network. It will do it to all the words within the document that was given to it. It is very likely to find the words with similar meanings nearby each other in a process of learning; it has a characteristic to gradually become a closer direction of a vector.
The symmetrization process was completed from the compound nouns that were extracted in 'The cosine similarity measurement by a compound noun analysis' to the higher similarity compound nouns. It targeted the compound nouns that held higher than 0.5 in cosine similarity. This is because it is commonly known that a higher number than 0.5 in cosine similarity is very close in resemblance. Figure 2 indicates the example in extraction.

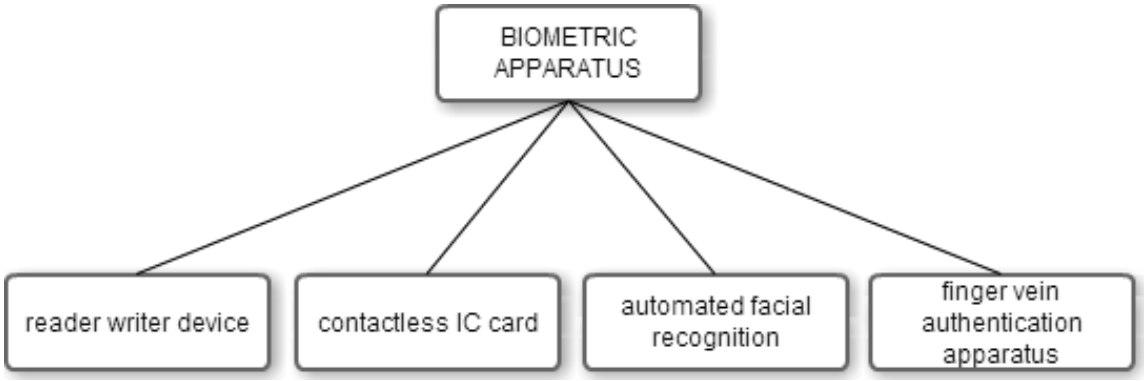

Figure 2: Image of compound noun summarization

Cosine similarity measurement by morphological analysis after the summarization

1016 compound nouns that appeared in the beginning were summarized down to 85%. The cosine similarity on each of the extracted documents was again calculated based on the summarized 865 compound nouns.

Verification of summarized compound nouns

Linear discriminant analysis was used to verify the accuracy of the summarization and the summarized compound nouns, not to just rely on the *word2vec's* function. A description of the invention is always found at the end of the extracted documents. 91 out of 116 documents were written about a 'Biometric Authentication Device'. The other 25 documents had something to do with or used Biometric Authentication Device, however, the invention itself was not about the Biometric Authentication Device. Among the 116 documents, the inventions that were about Biometric Authentication Device were marked as 'Y', and the ones not about the Biometric Authentication Device were marked as 'N' to be inputted into the system as solution data. Thereafter, all the targeted documents and the series of compound nouns were analyzed by linear discriminant analysis. It was used to verify whether or not they were correctly divided into 91 and 25 documents.

Data analysis

Comparison between morphological analysis and compound noun analysis

The compound noun analysis results that were used in this section were targeting the summarized compound nouns in 'Cosine similarity measurement by morphological analysis after the summarization'. Using ' Verification of summarized compound nouns ' as a reference, it is fine to find the similarity in $Y \cap Y$. On the contrary, it is better to have no similarity in $N \cap N$ and $Y \cap N$.

‘

All the document comparison results came out as similar by morphological analysis. The results showed the similarity in all, even though strictly speaking, 25 documents, which were written about the different invention, had been included among these documents. This means that it picked up noise, which is called ' garbage of the search' in those 25 documents, 300 pairs. Therefore, morphological analysis shows the result that it picked up 100% noise in N ∩ N and Y ∩ N.

On the other hand, compound noun analysis results showed the similarity in all Y ∩ Y when it paired up two documents that were written about the Biometric Authentication Device as the correct documents. It also showed the results on 116 out of 300 pairs of N ∩ Na's zero similarity as it corresponds to the actual contents. Y ∩ N's results also showed as zero similarity on 526 out of 2,275 pairs. These results were put in Table 1.

Table1: Comparison between morphological analysis and compound noun analysis

	count	dissimilarity count	Noise count	Noise rate
morphological analysis				
N∩N	300	0	300	100
Y∩N	2275	0	2275	100
compound analysis				
N∩N	300	116	184	61.33
Y∩N	2275	526	1749	76.88

Linear discriminant analysis on the documents using compound nouns

The compound nouns that were extracted in 'The cosine similarity measurement by a compound noun analysis' were summarized down to 85% in 'Compound noun summarization method'. However, if that summarization wasn't done correctly, it may draw a wrong result that shows the non-resembled documents as similar.
It was analyzed by discriminant analysis to check the accuracy of the compound noun summarization. Table 2 indicates that the document classification results in between, before and after the compound noun summarization have slightly moved towards the correct direction.

Table 2: Linear discriminant analysis result

	Morphological analysis		Compound noun analysis(before)		Compound noun analysis(after)		
	n	y	n	y	n	y	
n	17	8	18	7	19	6	25
y	4	87	1	90	1	90	91

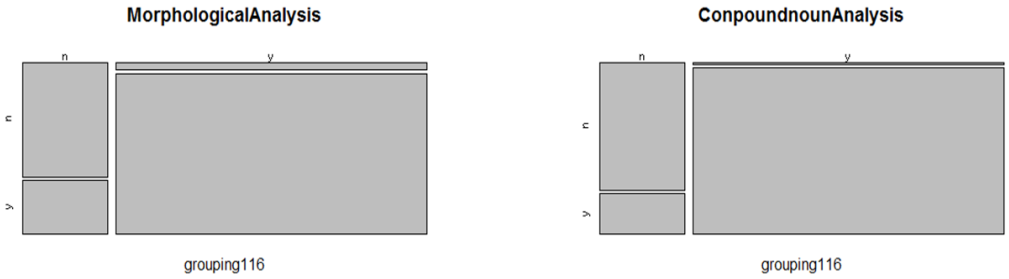

Figure3: The figure of linear discriminant analysis

Furthermore, Figure 3 visualized the result that indicates, in comparison to morphological analysis, compound noun analysis after the compound noun summarization analyzed the documents' similarity more correctly.

Conclusion

Document similarity measurement using a morphological analysis, which has a great amount of noise, is not perfect. Since it picks up a lot of noise, the precision will go down. However, if a compound noun analysis was used for a document similarity search, then the recall ratio will go down.

As forthcoming challenges, in order to raise the recall ratio, the extracted documents' expansion to the entire claim, and having many compound nouns be extracted should be considered. Moreover, it is considered to be able to raise the recall ratio to some extent if summarizing the compound nouns with a better summarization ratio is employed.

References

[1] Nakagawa, H., Mori, T. and Yumoto, H., 2003, Term extraction based on occurrence and concatenation frequency, *Journal of natural language processing,* 10, 27-45.

[2] Nakagawa, H. and Mori, T., "A simple but powerful automatic term extraction method," in *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14*, 1-7, 2002.

[3] Hammouda, K. M. and Kamel, M. S. 2004 Efficient phrase-based document indexing for web document clustering, *Knowledge and Data Engineering, IEEE Transactions on,* 16, 1279-1296.

[4] Hirao, K. and Takeuchi, K., Web Search Result Clustering Based on Structure of Compound Nouns, IPSJ SIG Technical Report 2006, 35-42, 2006.

[5] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, 2007, Text mining techniques for patent analysis, *Information Processing & Management,* 43, 1216-1247.

[6] *termmi*
http://gensen.dl.itc.u-tokyo.ac.jp/termmi.html (*in Japanese)*

[7] *word2vec*
https://code.google.com/p/word2vec/