Thesaurus Expansion using Similar Word Pairs from Patent Documents

Yoshimi Suzuki, Fumiyo Fukumoto

Department of Computer Science and Media Engineering
University of Yamanashi
ysuzuki@yamanashi.ac.jp, fukumoto@yamanashi.ac.jp

Abstract

In both written and spoken languages, we sometimes use different words in order to describe the same meaning. For instance, we use "constraint" (*seigen*) and "restriction" (*seiyaku*) as the same meaning. This makes text classification and text summarization difficult. In order to deal with this problem, dictionaries especially thesauri are used. However, in technical paper and patent documents, a lot of new words which are not given in the dictionary. In this paper, we propose a method to accurately extract words which are semantically similar to each other. Using this method, we extracted similar word pairs from patent documents. We also expand a thesaurus using the extracted similar words.
Index Terms: thesaurus, similar word pair.

1. Introduction

We often access a great deal of machine readable documents. Consequently, the necessity of text classification increases. Many approaches and techniques are used for text classification. For instance, the following techniques are proposed.

- statistical approach: e.g. methods using χ^2 values or TF·IDF (Term Frequency $\times$ Inverse Document Frequency) values as term weights, etc.
- machine learning: e.g. Naive Bayes, Support Vector Machines [1], Maximum Entropy [2], etc.

In patent documents, there are various technical terms. When different terms are used for the same meaning in different texts, term frequency or value of TF·IDF is not very useful for sentence extraction. In this paper, we deal with the problem of similar words for text classification.

Figure 1 illustrates similar words in patent documents.

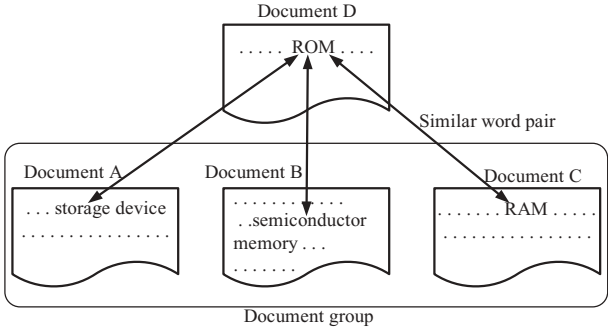

Figure 1: **Similar words in patent documents**

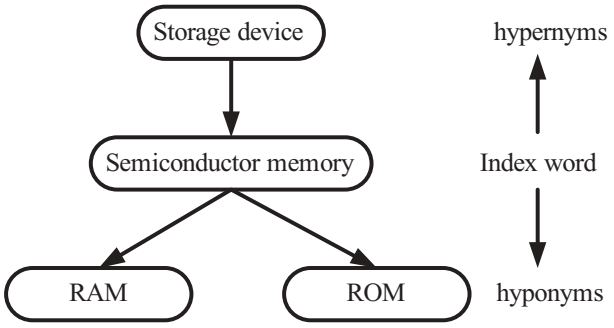

Figure 2: **Semantic hierarchy of "semiconductor memory"**

In Figure 1, each arrow points to similar word pair. "similar words" are defined as a noun pair which has the same role. For instance, let us consider "ROM", "storage device", "semiconductor memory" and "RAM".

Figure 2 illustrates hypernyms and hyponyms of "semiconductor memory", and these words correlate with each other.

We consider that Documents A,B and C are same group which are similar to each other and we have to decide the new document D is similar to the documents of the group. The word "ROM" appears in document D. However, "ROM" does not appear in the documents in the group. The words "storage device", "semiconductor memory" and "RAM" which appear in the documents in the group dose not appear in documents D. Therefore it is difficult to decide documents D is similar to the group. In order to deal with this problem, it is necessary to use thesauri. However, in patent documents and technical papers there are many new words which do not appear in thesauri. In this paper, we propose a method to extract similar word pairs for extending thesaurus.

There are many approaches for extracting similar word pairs in the study of word sense disambiguation and automatic thesaurus construction. Hindle proposed a method using predicate-argument structure. His method is based on the distribution of subject, verb and object [3]. Lin proposed a method based on the distributional pattern of words [4]. He defined the "local context" of each word and used dependency relationship of each word pair in local context. Lin also proposed a method for constructing thesaurus automatically [5]. Uramoto [6] proposed a method to expand an existing thesaurus using statistical information from corpus. Kanzaki [7] proposed a method based on self-organization semantic map for classification of adjectival noun and non-adjectival noun. They obtained similar word pairs accurately. However there are no studies for text classification using extracted similar words.

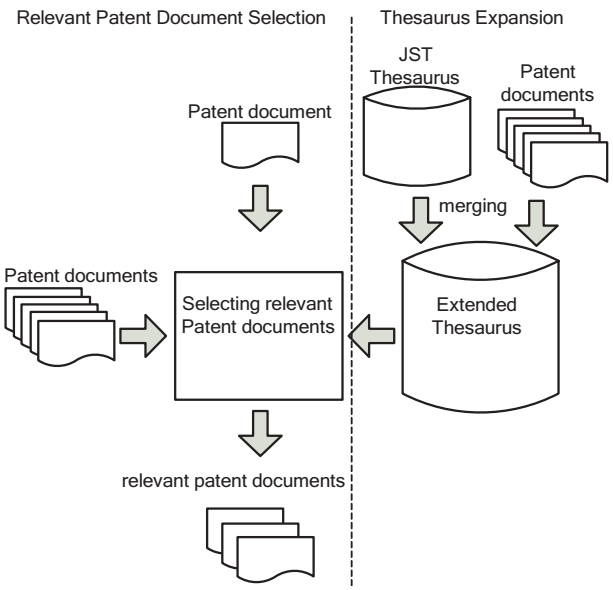

Figure 3: **System diagram of thesaurus expansion for relevant patent document selection**

We propose a method to extract similar noun pairs. More specifically,

- using many types of dependency relationship between noun and verb (Hindle used two types : subject and object)
- using patent documents which have many technical terms

Suzuki claimed Lin's method is better than Hindle's method for Japanese documents [8]. We also demonstrate that the number of types of dependency relationship affects accuracy of extracting similar nouns.

2. System Design

Figure 3 illustrates the diagram of our system. It consists of 2 phases: "Thesaurus Expansion" and "Relevant Patent Document Selection". In "Thesaurus Expansion" phase, similar noun pairs are extracted using a Japanese Dependency Structure Analyzer : CaboCha [9] and Lin's method. Using CaboCha, we obtained nouns and noun phrases which modify a predicate. Using Lin's method, we extracted similar noun pairs. We also expand a thesaurus using the extracted similar noun pairs. In "Relevant Patent Document Selection" phase, we will extract relevant documents of the input document using both TF·IDF and the extended thesaurus.

Figure 3 illustrates the diagram of our system.

3. Extracting Similar Word Pairs

We calculate similarity between nouns by using Lin's method [5].

Lin defined "dependency triple" as consisting of two words: w, w' and the grammatical relationship between them:r in the input sentence. $||w, r, w'||$ denotes the frequency count of the dependency triple (w, r, w'). $||w, r, *||$ denotes the total occurrences of $w - r$ relationships in the corpus, where $*$ indicates wild card.

First, $I(w, r, w')$ is calculated using Formula (1).

Table 1: **Similar nouns of "sound" (*oto*)**

Rank	Similar noun	Similarity
1	discord (*souon*)	0.313
2	voice (*onsei*)	0.267
3	noise (*noizu*)	0.198

$$
\begin{aligned}
& I(w, r, w') \\
&= -\log(P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)) \\
&\quad -(-\log P_{MLE}(A, B, C)) \\
&= \log \frac{||w, r, w'|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||} \qquad (1)
\end{aligned}
$$

where P_{MLE} is the maximum likelihood estimation of a probability distribution.

Let $T(w)$ be the set of pairs (r, w') such that $\log \frac{||w,r,w'||\times||*,r,*||}{||w,r,*||\times||*,r,w'||}$ is positive. The similarity $Sim(w_1, w_2)$ between two words: w_1 and w_2 is defined by Formula (2).

$$
\begin{aligned}
& Sim_L(w_1, w_2) \\
&= \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)} \qquad (2)
\end{aligned}
$$

For example, the top-3 similar nouns for the word "sound" are shown in Table 1.

For word expansion, we used pairs of respective nearest neighbors (RNNs) as similar word pairs. If word A and word B are RNN, word A is the most similar words of word B and also word B is the most similar words of word A.

4. Experiments

We applied our method extracting similar word pairs to patent documents.

4.1. Experimental Data

4.1.1. Dependency Relationship

We used the following 17 kinds of post-positional particles (which frequently appear in newspaper articles). Each pair of particle and verb corresponds to a type of dependency relationship. Table 2 illustrates the particles we use for calculating similarity between nouns. In Table 2 "frequency of appearance" illustrates the frequency of appearance of each particle in the newspaper of Mainichi Newspapers Co. from 1991 to 2000. ◯ means that the post-positional particle is registered in the particle set. We used 3 kinds of particle sets. *Set A* consists of 17 particles, *Set B* consists of 6 particles and *Set C* consists of 2 particles. *Set C* consists of particles "*wo*" and "*ga*" which corresponds to object and subject.

4.1.2. Patent Documents

We used patent documents taken from Japanese patent applications published in 1993 and 1994 which are provided by Patent Retrieval Task of NTCIR-4 [10]. There are 689,238 documents in

Table 2: **Post-positional particles in the experiments**

Post-positional particle	Frequency of appearance	A (17)	B (6)	C (2)
wo	7.6M	○	○	○
ni	6.1M	○	○	
ga	5.0M	○	○	○
ha	4.6M	○		
de	2.4M	○	○	
to	1.8M	○		
mo	1.3M	○		
kara	0.8M	○	○	
toshite	0.3M	○		
made	0.2M	○		
nitsuite	0.2M	○		
he	74K	○	○	
nitaishi	60K	○		
niyotte	46K	○		
demo	42K	○		
nitaishite	24K	○		
wotsuujite	18K	○		

Table 3: **Post-positional particles in the experiments**

filename	# of documents	amount of documents
1993A	164,982	2.2GB
1993B	182,345	2.6GB
1994A	181,384	2.7GB
1994B	160,527	2.6GB
Total	689,238	10.1GB

the training data. Table 3 illustrates amount of patent documents. In Table 3, 1993A means the documents in the first half of 1993.

4.1.3. JST Thesaurus

In order to expand words, we used JST Thesaurus [11]. It has about 430,000 index words which are terms for science and technology. Each index word has some categories, some synonyms, some hypernyms, some hyponyms and some coordinate terms. Figure 4 illustrates an example of index in JST Thesaurus. JST Thesaurus is well organized. However, intersection of the index words and words in patent documents in 1993 is 19,537. It means half of index words do not appear in patent documents, and most of words in patent documents do not appear in the JST Thesaurus. In order to deal this problem, we expanded the thesaurus using similar word pairs from patent documents.

Figure 5 illustrates how to merge new words from patent documents into the thesaurus.

4.2. Experimental Results

First, we calculated similarity between nouns. The extracted similar words are classified into 4 types : synonyms, antonyms, hypernyms and coordinate terms. Table 4 illustrates examples of each type of similar words. All of these examples are extracted as similar word pairs from patent documents. In Table 4, coordinate terms mean that both word A and word B have same hypernyms.

Table 5 lists 10 examples of RNNs on patent documents with

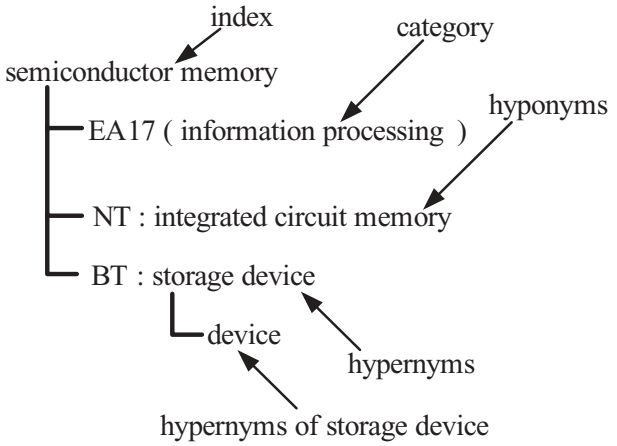

Figure 4: **An example of index in JST Thesaurus**

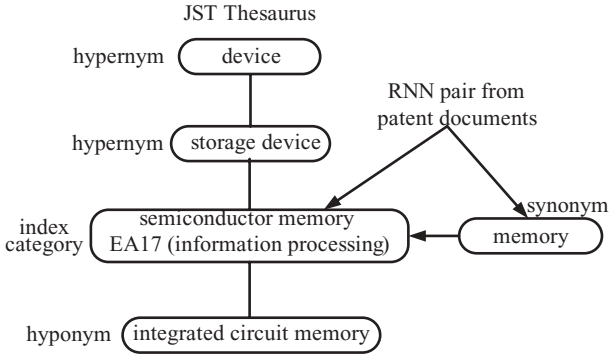

Figure 5: **Structure of JST Thesaurus and RNNs word pairs**

the proposed method. Noun A and noun B represent the pair of similar words. In table 5, there are some examples of similar noun pairs. Nitrogen gas is an inert gas, and alminum is a kind of metal. Protrusion and projection are synonyms. Both transistor and condenser are electronic parts, but they have different functions. We also obtained many paraphrase pairs: e.g. (the United States, the US).

Table 6 illustrates the number of RNNs. In Table 6 '# of appropriate pairs' means the number of RNN pairs obtained. 'particle set' corresponds to 3 kinds of particle sets of Table 2. Set A, B and C consist of 17, 6 and 2 particles respectively. Evaluation as to whether nouns in a pair of RNNs are similar to each other is performed by humans. Using Set A, we obtained better results than that using other sets. Table 6 suggest that for extracting many RNNs we use many kinds of post-positional particles as dependency relationships.

Figure 6 shows coverage for 1995A by # of words in JST Thesaurus and RNNs. In Figure 6, coverage illustrates $\frac{\#\ of\ nouns\ which\ appear\ both\ in\ the\ extended\ thesaurus\ and\ in\ 1995A}{\#\ of\ nouns\ in\ 1995A}$. The coverage becomes high as the quantity of data increases.

5. Conclusions

We proposed a method for calculating similarity between nouns from Japanese documents. Since we use 15 types of dependency

Table 4: **4 types of extracted similar words**

Type	Examples Noun A	Noun B	Sim
synonyms	command	instruction	0.401
	device	system	0.367
antonyms	read	write	0.520
	input data	output data	0.443
hypernyms	size	height	0.404
	cable	optical fiber	0.323
coordinate terms	infra-red ray	ultraviolet ray	0.455
	transistor	condenser	0.305

Table 5: **Examples of RNNs (with all categories, Set A, from Patent Documents)**

Noun A	Noun B	Similarity
inert gas	nitrogen gas	0.47
CPU	processor	0.45
fluid	liquid	0.48
air	gas	0.50
production cost	cost	0.52
ultraviolet ray	infra-red ray	0.46
metal	alminum	0.38
protrusion	projection	0.28
dust	inert matter	0.35
transistor	condenser	0.31

relationship in addition to the object and subject, we can extract similar noun pairs accurately. We used Japanese patent documents as training data, and found that Lin's method is useful for Japanese patent documents.

We also performed extracting similar word pairs from Japanese patent, and we found there are many paraphrases in such documents, and that paraphrases and other similar word pairs are successfully extracted by our method.

Future work includes (i) associative patent retrieval using the extended thesaurus, (ii) automatic noun classification into 4 types of similar words (Table 4).

6. References

[1] Fukumoto, F. and Suzuki, Y., "Correcting Category Errors in Text Classification", Proceedings of COLING2004, pp.868-874, 2004.

[2] Nigam, K., Lafferty, J. and MacCallum, A., "Using Maximum Entropy for Text Classification", Proceeding of the 16th International Joint Conference Workshop on Machine Learning for Information Filtering, pp61-67, 1999.

[3] Hindle, D., "Noun Classification from Predicate-Argument Structures", Proceedings of 28th Annual Meeting of the Association for Computational Linguistics, pp.268-275, 1990.

[4] Dekang, Lin., "Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity",Proceedings of ACL/EACL-97, pp.64-71, 1997.

[5] Dekang, Lin., "Automatic Retrieval and Clustering of Similar Words", Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference, pp.768-774, 1998.

[6] Uramoto, N., "Corpus-based Thesaurus – Positioning Words in Existing Thesaurus Using Statistical Information from a Corpus", Journal of Information Processing Society of Japan, Vol.37, No.12, 2182-2189, 1996.

[7] Ma, Q., Kanzaki, K., Murata, M., Uchimoto, K. and Isahara, H., "Self-Organization Semantic Maps of Japanese Nouns in Terms of Adnominal Constituents", Proceedings of IJCNN2000, Vol.6, pp.91-96, 2000.

[8] Suzuki, Y. and Fukumoto, F., "Clustering Similar Nouns for Selecting Related News Articles", Proceedings of INTERSPEECH2004, ThB2202p.21, 2004.

[9] Kudo, T. and Matsumoto, Y., "Japanese Dependency Analysis using Cascaded Chunking", CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002, pp.63-69, 2002.

[10] Fujii, A., Iwayama, M. and Kando, N., "Test Collection for Patent-to-Patent Retrieval and Patent Map Generation in NTCIR-4 Workshop", Proceedings of the 4th International Conference on Language Resources and Evaluation, pp.1643-1646, 2004.

[11] "JST Thesaurus 1999", http://jois.jst.go.jp/JOIS/html/thesaurus_index.htm, 1999.

Table 6: **Results of extracting RNNs**

Particle set	# of appropriate pairs / # of RNNs
Set C (2)	191/202 = 94.6%
Set B (6)	195/205 = 95.1%
Set A (17)	193/201 = 96.0%

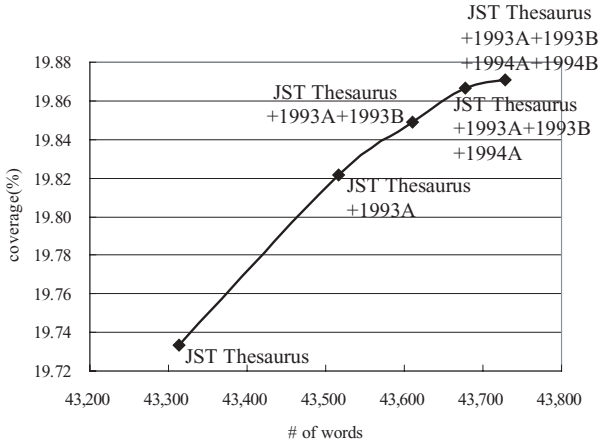

Figure 6: **Coverage for 1995A by # of words in JST Thesaurus and extracted RNNs**