Automatic Construction of a Bilingual Thesaurus using Citation Analysis

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
+81-82-830-1584
nanba@hiroshima-cu.ac.jp

Saori Mayumi
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
mayumi@ls.info.hiroshima-cu.ac.jp

Toshiyuki Takezawa
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
+81-82-830-1768
takezawa@hiroshima-cu.ac.jp

ABSTRACT
We propose a method for constructing a bilingual thesaurus automatically from patents. First, we extract hypernym-hyponym relations from Japanese and US patents by using the pattern "A such as B". Second, we align terms between these thesauri by combining statistical machine translation and citation analysis techniques. To confirm the effectiveness of our method, we conducted some experiments. The results showed that our best method obtained Recall of 79.4%, Precision of 77.5%, and F-measure of 78.3%.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval**]: Search process

H.3.4 [**Systems and Software**]: Performance evaluation

H.3.5 [**Online Information Services**]: Data sharing

General Terms
Measurement, Performance, Experimentation

Keywords
bilingual thesaurus, machine translation, cross-lingual patent retrieval, citation analysis.

1. INTRODUCTION
We propose a method for constructing a bilingual thesaurus from Japanese and US patents. Thesauri are used as information sources for writing and searching technical documents including patents. They also serve as useful resources for natural language processing. However, they are costly to construct and maintain manually, and several methods for automatic construction of monolingual [4, 6, 7, 11, 13, 14] and bilingual [1, 8, 16] thesauri have been studied.

A typical method for constructing a monolingual thesaurus is to apply a pattern "A such as B" to a text corpus, and to extract "A" and "B" as a hypernym-hyponym pair [4]. We apply Hearst's method to Japanese and US patents, and obtain hypernym-hyponym relations. Then, we use citation analysis techniques [5, 15] to align the English and Japanese terms, and finally construct a bilingual thesaurus. This thesaurus enables us to write and search English and Japanese patents. It can also be used for various natural language processing tasks, such as patent translation [3], cross-lingual information retrieval for patents [2, 13], and cross-lingual patent mining [9].

This paper is organized as follows. Section 2 describes related work. Section 3 explains the automatic construction of a bilingual thesaurus using citation analysis techniques. To investigate the effectiveness of our method, we conducted experiments, as reported in Section 4. We discuss the experimental results in Section 5. We present our conclusions in Section 6.

PaIR'11, October 24, 2011, Glasgow, Scotland, UK.

2. RELATED WORK
Several methods for thesaurus construction have been proposed, and they are divided into three categories: (1) extraction of hypernym-hyponym relations [4, 11, 14]; (2) collection of related terms [6, 7]; and (3) translation of technical terms [1, 8, 16]. In the following, we summarize these methods, and describe their relationship to our work.

2.1 Hypernym-Hyponym Extraction
Several methods for extracting hypernym-hyponym relations from text corpora have been proposed. Shinzato and Torisawa [13] extracted such relations from Web documents using HTML structure, while Oishi et al. [11] utilized terms and their definitions[1] that automatically extracted from Web documents. Hearst [4] proposed a method for extracting hyponyms from text corpora using a set of patterns. For example, "magnetic tape" and "floppy disc" are extracted as hyponyms of "magnetic recording media" from the following sentence, using the pattern "NP_0 such as {NP_1, NP_2, (and/or)} NP_n".

> Methods for manufacturing magnetic recording media such as magnetic tapes and floppy discs are well known in the art …

As there are many sentences containing "等の" (such as) or "such as" expressions in both Japanese and US patent documents, we use this pattern for hypernym-hyponym extraction.

2.2 Collection of Related Terms
Lin [7] and Lee [6] proposed "distributional similarity" methods for calculating the similarity between terms. They focused on the contexts in which terms are used, and defined the similarity between two terms as the amount of information contained in the

[1] Terms are often defined by their hypernyms. Therefore, words used in definitions tend to be hypernyms. For example, "mammal" is a hypernym of "lion", and "lion" can be defined as "a mammal."

commonality between the terms, divided by the amount of information in the contexts of the terms.

Our method of using citation analysis techniques for aligning English and Japanese thesauri can also be considered as a kind of distributional similarity. When calculating the similarity between terms, we use their hypernyms and hyponyms, which were extracted using Hearst's method.

2.3 Translation of Technical Terms

Tonoike et al. [16] devised a compositional translation method for translating technical terms. Fujii and Ishikawa [1] applied the compositional translation method to cross-lingual information retrieval, and confirmed its effectiveness.

Morishita et al. [8] proposed a method for translating technical terms by combining a phrase table (statistical machine translation), a manually created bilingual dictionary, and compositional translation using support vector machines, and confirmed the effectiveness of the phrase table. We also use a phrase table as an initial alignment between English and Japanese term pairs.

3. AUTOMATIC ALIGNMENT BETWEEN ENGLISH AND JAPANESE HYPERNYM-HYPONYM PAIRS

We identify correct English-Japanese pairs of hypernym-hyponym relations. We do not align between English and Japanese terms because we wish to avoid the problem of ambiguity of word sense. For example, the Japanese term "カッター" (cutter) has two senses: (1) cutter shirt, and (2) cutter knife, and it is impossible to align the term with an appropriate English term. However, when a hypernym-hyponym pair "衣類 > カッター" (clothing > cutter) is given, the sense of "cutter" is uniquely identified to "cutter shirt." In this section, we describe a phrase table, an English-Japanese bilingual dictionary in Section 3.1. Then, we explain the procedure for aligning English-Japanese hypernym-hyponym pairs using citation analysis techniques in Section 3.2.

3.1 Alignment Using a Phrase Table

To align English and Japanese hypernym-hyponym pairs, we use a statistical machine translation technique. First, we obtain a phrase table (an English-Japanese bilingual dictionary) from 3,185,254 pairs of sentences from English and Japanese patents [3] using Giza++ [10], a statistical machine translation toolkit. Second, we align a hypernym-hyponym pair in Japanese with one in English.

In the following, we explain the procedure for aligning a Japanese hypernym-hyponym pair "A > B" to the corresponding English pair, using Figure 1. Here, "A > B" indicates that A is a hypernym of B.

(Step 1) Translating Japanese terms

We translate the Japanese hypernym "金属" and hyponym "Ａｌ" into English individually using a phrase table.

(Step 2) Creating English hypernym-hyponym pairs

Consider that "金属" is translated into "metal", "iron", and "metallic", and "Ａｌ" is translated into "Al", "aluminium", and "aluminum". Now, we combine them and create English hypernym-hyponym pairs: "metal > Al", "metal > aluminium", "metal > aluminum", "iron > Al", and so on.

(Step 3) Aligning candidates with an English thesaurus

When the English pairs obtained in Step 2 exist in an English thesaurus, we select them as candidate English hypernym-hyponym pairs, corresponding to "金属 > Ａｌ".

Our method using citation analysis, which we will explain in the next section, is applied to the candidates, and valid English hypernym-hyponym pairs are selected.

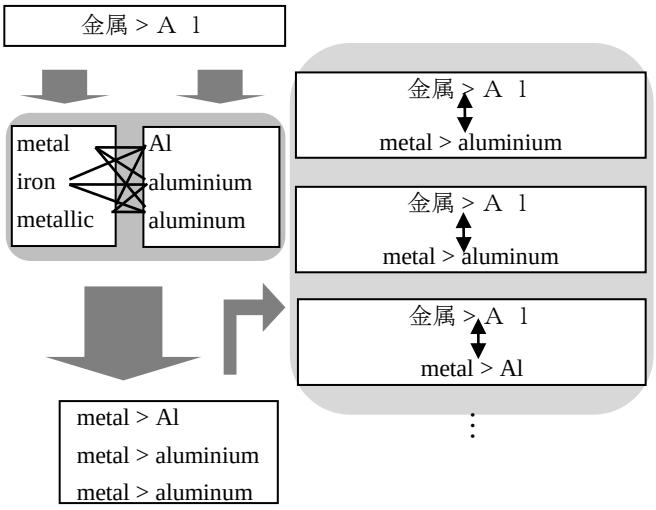

Figure 1. Alignment between English and Japanese terms using a phrase table

3.2 Alignment of English-Japanese Hypernym-Hyponym Pairs

To align English-Japanese hypernym-hyponym pairs, we used the following five features (see also Figure 2). In Figure 2, continuous lines indicate hypernym-hyponym relations, while dotted lines indicate English-Japanese term pairs.

① Translation probability

② The number of common hypernyms ("electronic component" in Figure 2) between English and Japanese hypernyms ("semiconductor device" and "半導体素子")

③ The number of common hyponyms ("IC") between English and Japanese hypernyms

④ The number of common hypernyms ("active device" and "能動素子") between English and Japanese hyponyms ("transistor" and "トランジスタ")

⑤ The number of common hyponyms ("FET") between English and Japanese hyponyms ("FET")

Here, the values of ① are calculated by multiplying the translation probability from Japanese to English with that from English to Japanese. The values of ②, ③, ④, and ⑤ are normalized by dividing by the maximum number in the thesaurus. We use the phrase table described in Section 3.1 for the alignment between English and Japanese terms. In this alignment, when a Japanese term can be aligned to multiple English terms, we select the English term with the highest translation probability.

The idea to use features ②, ③, ④, and ⑤ is based on citation analysis. It is well known that using citation analysis makes it possible to obtain topical collections of papers. In these studies, two similar papers were found to cite many of the same papers (bibliographic coupling [5]), or were cited from many other papers (co-citation analysis [15]). Here, by regarding hypernym-hyponym relations as citation relations, we can apply citation analysis techniques to calculate the similarities between terms. Now, we explain how to align a Japanese hypernym-hyponym relation "半導体素子 > トランジスタ" and an English relation "semiconductor device > transistor" using Figure 2. When "半導体素子 > トランジスタ" and "semiconductor device > transistor" have many hypernyms and hyponyms in common, they are considered to be semantically similar.

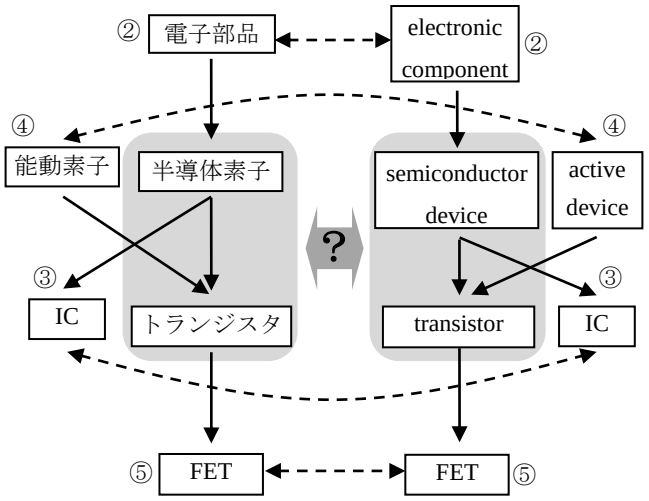

Figure 2. Alignment between English and Japanese hypernym-hyponym relations using citation analysis and a phrase table

We combine feature ① and one of the features ②, ③, ④, and ⑤. In the following, we describe the evaluation procedure.

(1) Calculation of the validity of each candidate

We calculate a value of "$a^{\beta} + \alpha \times b$", which indicates the validity of each candidate hypernym-hyponym relation. Here, a and b indicate the values of features ① and ②, respectively. Both α and β are parameters. We set β to the values 1/5, 1/10, 1/15, and 1/20, while α ranged from 0.1 to 0.9 at 0.1 intervals.

(2) Parameter tuning

We consider that a candidate is valid when its score, obtained in the previous step, exceeds a threshold value x, and calculate F-values. We examine values x from 0 to 2 at 0.01 intervals, and obtain the best threshold value when the F-value is highest.

(3) Evaluation

We evaluate our methods with the threshold value obtained in the previous step using a test data set.

4. EXPERIMENTS

To confirm the effectiveness of the above method, we conducted several experiments.

4.1 Extraction of Hypernym-Hyponym Relations from Patent Documents

Using the pattern that Hearst [4] proposed, we obtained 3,898,060 hypernym-hyponym relations from US patents over an eight-year period. Some examples of English hypernym/hyponym relations are shown in Table 1.

Similarly, we obtained 7,031,159 hypernym-hyponym relations from Japanese patent documents over a ten-year period, using the pattern "NP_0 (,|と|や) NP_2 (等の|などの)NP_n". Some examples of Japanese hypernym-hyponym relations are shown in Table 2.

Table 1. Examples of English hypernym-hyponym relations

Freq.	Hypernym	Hyponym
23	magnetic storage medium	magnetic disk
20		magnetic tape
11		magnetic disc
8		computer disk
6		floppy disk

Table 2. Examples of Japanese hypernym-hyponym relations

Freq.	Hypernym	Hyponym
101	磁気記憶装置 (magnetic storage medium)	ハードディスク (hard disc)
39		磁気ディスク装置 (magnetic disc system)
26		ハードディスク装置 (hard disc system)
25		ＨＤＤ
22		ハードディスクドライブ (hard disc drive)

4.2 Experimental Setting

- **Data**

We used 2,635 manually evaluated pairs of English-Japanese hypernym-hyponym relations. Among these, 982 pairs were identified as correct.

- **Alternatives**

We propose four methods[2], which combine two features of the five identified in Figure 1 (see Table 3). In addition to these four methods, we also examined a baseline method, which uses translation probabilities alone.

Table 3. Features used in the experiment

		Features				
		①	②	③	④	⑤
Our Methods	(a)	○	○			
	(b)	○		○		
	(c)	○			○	
	(d)	○				○
Baseline method (e)		○				

[2] We did not examine combinations of more than two features, because only our method (d) could outperform the baseline method (e) (see Table 4), and we could not expect combinations of more than two features to outperform the baseline method.

- **Parameter tuning**

We performed a four-fold cross validation test for parameter tuning.

- **Evaluation measures**

We used Recall, Precision, and F-measure for the evaluation.

4.3 Experimental Results

The experimental results are shown in Table 4. As can be seen, Recall values of our methods outperformed the baseline method. The F-value of our method (d) is also better than that of the baseline method.

Table 4. Experimental results

		α	β	Precision	Recall	F-measure
Our methods	(a)	0.1	1/10	76.4	**78.1**	77.1
	(b)	0.1	1/20	76.3	**79.5**	77.4
	(c)	0.1	1/15	75.8	**78.4**	76.9
	(d)	0.1	1/15	77.5	**79.4**	**78.3**
Baseline	(e)	0	1/15	78.5	77.8	78.0

5. DISCUSSION

5.1 Error Analysis

5.1.1 Cases that our method mistakenly aligned

Our method mistakenly aligned English-Japanese hypernym-hyponym pairs in 226 cases, and there are two typical errors: (1) alignment errors between terms having similar characteristics (73.9%); and (2) errors in extracting hypernym-hyponym relations from patents (22.4%). We describe these errors as follows.

(1) Alignment errors between terms having similar characteristics (73.9%)

Table 5 shows typical examples of this type of error. "亜鉛" (zinc) and "aluminum-zinc" in the first case, and "染料" (dye) and "organic dye" in the second case, were mistakenly aligned. Two terms that have similar characteristics tend to have many hypernyms or hyponyms in common, and as a result, our method using citation analysis mistakenly aligned them.

Table 5. Alignment errors between terms having similar characteristics

Japanese		English	
Hypernym	Hyponym	Hypernym	Hyponym
金属 (metal)	亜鉛 (zinc)	metal	aluminum-zinc
着色剤 (coloring agent)	染料 (dye)	coloring agent	organic dye

(2) Errors in extracting hypernym-hyponym relations from patents (22.4%)

This type of error is caused by problems in extracting hypenym-hyponym relations from patents, rather than from problems in our alignment method using citation analysis. Table 6 shows typical examples of this type of error. In the first case, "elastic body" should be extracted from texts as a hypernym of "rubber", instead of "elastic". In the second case, "solvent or" should be extracted, but the unnecessary word "or" is contained in the hypernym.

Table 6. Typical errors in extracting hypernym-hyponym relations from patents

Japanese		English	
Hypernym	Hyponym	Hypernym	Hyponym
弾性体 (elastic body)	ゴム (rubber)	elastic	rubber
溶媒 (solvent)	水 (water)	solvent or	water

5.1.2 Cases that our method could not align

Our method could not align English-Japanese hypernym-hyponym pairs in 201 cases, and there are three typical errors: (1) singular/plural forms (33.3%); (2) atomic symbols (22.9%); and (3) abbreviations (21.4%). We describe these errors as follows.

(1) Singular/plural forms (33.3%)

An example of this case is shown in the first line in Table 7. Our method could not align a plural form of the English term "recording media" with the Japanese term "記憶媒体" (recording medium). Although there is a pair "記憶媒体" – "recording media" in the phrase table, its translation probability is much lower than the pair "記憶媒体" – "recording medium", and as a result, our method did not align "記憶媒体" and "recording media".

(2) Atomic symbols (22.9%)

An example of this case is shown in the second line in Table 7. Our method could not align the Japanese term "銅" (copper) with the English term "Cu", which is expressed as an atomic symbol.

(3) Abbreviations (21.4%)

An example of this case is shown in the third line in Table 7. Our method could not align an English term "integrated circuit" with a Japanese term "ＩＣ" (IC), which is expressed as an abbreviation form.

Table 7. Typical errors that our method could not align

Japanese		English	
Hypernym	Hyponym	Hypernym	Hyponym
記憶媒体 (recording medium)	ＣＤ (CD)	recording media	CD
金属材料 (metallic material)	銅 (copper)	metallic material	Cu
電子部品 (electric part)	ＩＣ (IC)	electric part	integrated circuit

5.2 Improvement of Recall Values

As can be seen from Table 4, all of our methods improved on the baseline method. To confirm the effectiveness of our method more precisely, we counted the number of cases that our methods could align correctly that the baseline method could not. We found that there were 15 such cases, and there were no opposite cases.

The baseline method could not align "記憶媒体" (recording medium) and "record medium", because the translation probability between "記憶媒体" and "record medium" is low. On the other hand, our methods could align this pair correctly, and we can conclude that our method using citation analysis is useful for alignment, even when the translation probability is low.

6. CONCLUSION

We have proposed a method for constructing a bilingual thesaurus in two steps: (1) extraction of hypernym-hyponym relations from Japanese and US patents; and (2) alignment between them. In step 1, we applied Hearst's method to Japanese and US patents. In step 2, we used citation analysis techniques with a phrase table. To confirm the effectiveness of our method, we conducted some experiments. The results showed that our method (d) obtained Recall of 79.4%, Precision of 77.5%, and F-measure of 78.3%.

7. REFERENCES

[1] Fujii, A. and Ishikawa, T. 2000. Cross-Language Information Retrieval Based on Query Keyword Translation: An Internet Search Application. International Journal of Computer Processing of Oriental Languages, Vol.13, No.1, pp.1-13.

[2] Fujii, A., Iwayama, M., and Kando, N. 2007. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. Proceedings of the 6th NTCIR Workshop Meeting.

[3] Fujii, A., Utiyama, M. Yamamoto, M., Utsuro, T. Ehara, T., Echizen-ya, H., and Shimohata, S. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In Proceedings of the 8th NTCIR Workshop Meeting, pp.371-376.

[4] Hearst, M.A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545.

[5] Kessler, M.M. 1963. Bibliographic Coupling between Scientific Papers. American Documentation, Vol.14, No.1, pp.10-25.

[6] Lee., L. Measures of Distributional Similarity. 1999. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.25-32, 1999.

[7] Lin, D. 1998. Automatic Retrieval and Clustering of Similar Words. Proceedings of the 17th International Conference on Computational Linguistics, pp.768-774.

[8] Morishita, Y., Utsuro, T., and Yamamoto, M. 2008. Integrating a Phrase-based SMT Model and a Bilingual Lexicon for Human in Semi-Automatic Acquisition of Technical Term Translation Lexicon. Proceedings of the 8th Conference of the Association for Machine Translation in the Americas, pp.153-162.

[9] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. 2010. Overview of the Patent Mining Task at the NTCIR-8 Workshop. Proceedings of the 8th NTCIR Workshop Meeting, pp.293-302.

[10] Och, F.J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol.29, No.1, pp.19-51.

[11] Ohishi, Y., Itou, K., Takeda, K., and Fujii, A. 2006. Statistical Analysis for Thesaurus Construction using an Encyclopedic Corpus. Proceedings of the 5th International Conference on Language Resources and Evaluation, pp.1368-1371

[12] Roda, G., Tait, J., Piroi, F., and Zenz, V. 2010. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain, in Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Penas, A. and Roda, G. (eds) Multilingual Information Access Evaluation I. Text Retrieval Experiments 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Springer LNCS, Vol.6241, pp.385-409.

[13] Sato, S. and Sasaki, Y. 2003. Automatic Collection of Related Terms from the Web. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Vol. 2, pp.121-124.

[14] Shinzato, K. and Torisawa, K. 2004. Acquiring Hyponymy Relations from Web Documents. Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting, pp.73-80.

[15] Small, H. 1973. Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents. Journal of the American Society for Information Science, Vol.24, pp.265-269.

[16] Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T., and Sato, S. 2005. Translation Estimation for Technical Terms using Corpus collected from the Web. Proceedings of the Pacific Association for Computational Linguistics, pp.325-331.