Query Expansion using an Automatically Constructed Thesaurus

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

Abstract

Our group participated in the Japanese and English Retrieval Subtasks of NTCIR-6. Our goal was to evaluate the effectiveness of a thesaurus constructed from patents for invalidity search. To confirm the effectiveness of our thesaurus-based query expansion, we conducted experiments and found that our method can improve upon traditional document retrieval systems.

Keywords: *Invalidity Search, Thesaurus, Query Expansion*

1 Introduction

We are studying the automatic construction of a thesaurus from patent documents. Our goal in NTCIR-6 [Fujii et al. 2007] is to evaluate the effectiveness of the thesaurus for invalidity search.

Invalidity search aims to find patent that can invalidate the assertions in an existing claim. However, it is often difficult to retrieve such patents using traditional document retrieval systems, because the terms used in a claim are often abstract or creative, to expand the range of the claim. As a result, different patents tend to contain different terms, even though these terms refer to the same things. To solve this problem, we apply a thesaurus-based query expansion.

This paper is organized as follows. Section 2 describes some related works. Section 3 explains the construction of a thesaurus for query expansion. Section 4 contains a system description. To investigate the effectiveness of our method, we conducted experiments, as reported in Section 5. We present our conclusions in Section 6.

2 Related Works

Several methods for thesaurus construction have been proposed, and they are divided into two categories: (1) extraction of hypernym/hyponym relations [Hearst 1992, Morin and Jacquemin 2004, Oishi et al. 2006], and (2) collection of related terms [Lee 1999, Lin 1998, Mase et al. 2005, Sato and Sasaki 2003]. In the following, we summarize these methods, and describe their relationship to our work.

2.1 Hypernym/Hyponym Extraction

Hearst [Hearst 1992] 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 a 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 "such as" or "等の" (such as) expressions in patent applications, we use this pattern for hypernym/hyponym extraction.

2.2 Collection of Related Terms

Lin [1998] and Lee [1999] proposed methods for calculating the similarity between terms. They focused on the contexts in which terms are used, and define the similarity between two terms as the amount of information contained in the commonality between the terms, divided by the amount of information in the contexts of the terms. Although the effectiveness of their method was confirmed in their experiments, the method did also collect several dissimilar term pairs.

In patent applications, inventors may explicitly describe related terms by using parentheses, as in "floppy disc (magnetic recording medium)". The term preceding the parentheses and the term in parentheses have a broader/narrower relationship. Mase et al. [2005] extracted related terms from the text in a "Description of Symbols" of Japanese patents. Though there are fewer term pairs collected by Mase's method than by Lin's and Lee's methods, the results are more reliable.

When using related term collection for query expansion, the more reliable data is preferable despite being less extensive, because incorrect query expansion directly affects precision. We therefore use Mase's method for our task.

3 Construction of a Thesaurus for Query Expansion

To construct English-language and Japanese-language thesauri, we extracted four types of relation between terms: "Hypernym/Hyponym", "Abbreviation", "Synonym", and "Related Term".

3.1 Hypernym/Hyponym Extraction

Using the same pattern as Hearst [Hearst 1992] proposed, we obtained 3,898,060 hypernym/hyponym relations from United States Patent Office (USPTO) patents over an eight-year period. Some examples of English hypernym/hyponym relations are shown in Table 1.

In the same way, 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)

3.2 Abbreviation Extraction

We extracted abbreviations for terms when parentheses were used. In some sentences, which were used for extracting hypernym/hyponym relations in Section 3.1, terms are shown with their abbreviation in parentheses. The following is an example containing an abbreviation.

> この外管３１の内部には例えばＰＴＦＥ(ポリテトラフルオロエチレン)等の合成樹脂材料によって形成された可撓性内管２１を挿通し、
>
> (A flexible inner tube 21 containing synthetic resin, such as PTFE (poly tetra fluoro ethylene), is inserted through an outer tube.)

In this sentence, "ＰＴＦＥ" is an abbreviation of "ポリテトラフルオロエチレン" (poly tetra fluoro ethylene). We extracted 50,161 Japanese terms and their abbreviations (Table 3). In the same way, we extracted 17,792 English terms and their abbreviations (Table 4).

Table 3 Examples of Japanese Terms and their Abbreviations

Freq.	Term 1	Term 2
935	ポリエチレンテレフタレート(polyethylene terephthalate)	ＰＥＴ
658	コンパクトディスク (compact disc)	ＣＤ
586	アルミニウム(aluminium)	Ａｌ
545	ポリプロピレン (polypropylene)	ＰＰ
498	タングステン(tungsten)	Ｗ
490	銅(copper)	Ｃｕ
449	発光ダイオード (light emitting diode)	ＬＥＤ

Table 4 Examples of English Terms and their Abbreviations

Freq.	Term 1	Term 2
160	polyvinyl chloride	PVC
130	bis	cyclopentadienyl
124	polytetrafluoroethylene	PTFE
123	chemical vapour deposition	CVD
122	central processing unit	CPU
115	aluminium	AL
109	tungsten	W

3.3 Synonym Extraction

In Japanese patent documents, the same concept is often expressed in different terms. For example, "フロッピーディスク" (floppy disc) may be expressed as "磁気記録装置" (magnetic recording system) or "磁気記憶装置" (magnetic storage medium). To search patents exhaustively, the use of a synonym dictionary is inevitable. However, it is difficult to automatically construct such a dictionary by traditional term co-occurrence methods, because synonyms such as "magnetic recording system" and "magnetic storage medium" rarely appear in the same patent. To extract synonyms, we focus on citation relations between patents. Generally, two patents that have a citation relationship are topically related. Therefore, we collect synonyms for a given term X in the following three steps;

(Step 1) Search all patents related to the term X,

(Step 2) Collect patents that have citation relationships with patents collected in Step 1,

(Step 3) Extract terms from patents collected in Step 2.

In order to extract topic terms from patents in Step 3, we focus on patent claims. In most claims, noun phrases before "において" (concerning) and after "を特徴とする" (characterized by) indicate topic terms. Figure 1 is an example of a Japanese claim, together with its English translation. In this figure, the boldface terms "シフトレバー装置" (shift lever device),

“シフトレバー” (shift lever), and “シフトロック装置” (shift lock unit) are extracted as topic terms.

車体に固定する筐体内に前後揺動体を車体前後方向へ回動可能に軸支し、該前後揺動体に揺動基部を車体左右方向へ回動可能に軸支し、該揺動基部に植設したシフトレバーを車体前後及び左右方向へ揺動させることにより筐体上面に形成したゲート部を移動し所望のレンジを選択して自動変速機を切替操作する**シフトレバー装置**において、前記揺動基部に一対の上下に離間した突起部を設け、一方の突起部に当接可能なＰ係止部と、他方の突起部に当接可能なＮ係止部を有する回転ロック体を前記筐体に回動可能に軸支するとともに、シフトレバーがＰレンジ又はＮレンジに移動したとき該回転ロック体を回動させるアクチュエータを前記筐体に固定したことを特徴とする**シフトレバー**の**シフトロック装置**。

(A **shift lock unit** for a **shift lever device** for use in an automatic transmission having parking and neutral ranges, comprising: a casing; a base rotatably supported in said casing, said base including first and second protrusions separate d from each other in a longitudinal direction thereof; a lock member rotatably supported in said casing, said lock member including first and second engagements which can abut on said first and second protrusions, respectively; a shift lever arranged with said base, said **shift lever** serving to select a desired range of the transmission; and an actuator fixed to said casing, said actuator rotating said lock member when said shift lever is moved to one of the parking and neutral ranges.)

Figure 1 A Claim Extracted from a Japanese Patent (Publication Number = 10-184868)

We extracted 522,810 different topic terms from Japanese patents over a 10-year period. For each term, we applied Steps 1 to 3, and obtained Japanese synonyms (257,459 pairs of terms). Some examples are shown in Table 5.

3.4 Related Term Extraction

As described in Section 2, we use Mase et al.’s method [2005] for related term extraction. We applied this method to Japanese patents, and obtained 22,952 pairs of related terms. Some examples are shown in Table 6.

Table 5 Examples of Japanese Synonyms

Freq.	Term 1	Term 2
75	磁気記憶装置 (magnetic storage medium)	磁気記録媒体 (magnetic recording system)
37		磁気ディスク装置 (magnetic disc system)
18		磁気ヘッド (magnetic head)
13		磁気記録装置 (magnetic recording medium)
11		金属薄膜型磁気記録 (metallic thin film magnetic recording)

Table 6 Examples of Japanese Related Terms

Freq.	Term 1	Term 2
4420	ＣＰＵ	中央処理装置 (central processing unit)
2158	制御手段 (control method)	ＣＰＵ
1800	像担持体 (image carrier)	感光ドラム (photosensitive dram)
1573	制御手段 (control method)	制御装置(control device)
1544	制御手段 (control method)	制御部 (control unit)

4 System Description

Our system comprises the following four steps:

(Step 1) Morphological analysis

We introduce the Vector Space Model as a retrieval model and SMART [Salton 1971] for term weighting. We use GETA[1] as a retrieval engine. We use ChaSen[2] and TreeTagger[3] as Japanese and English morphological analysis tools, respectively.

(Step 2) Stopword deletion

Our system uses nouns, verbs, adjectives, and unknown words to retrieve relevant patents. In this method, unimportant words are stripped from those terms that are any of the above parts of speech and are extracted from the query.

(Step 3) Topic term extraction from a claim

Our system extracts topic terms from a claim using a claim analyser [Shinmori et al. 2002] that we described in Section 3.3. The extracted terms are expanded in the next step.

[1] GETA: http://geta.ex.nii.ac.jp/
[2] ChaSen: http://chasen.naist.jp/hiki/ChaSen/
[3] TreeTagger:
http://www.ims.uni-stuttgart.de/projekte
/corplex/TreeTagger/DecisionTreeTagger.html

(Step 4) Thesaurus-based query expansion

We use the following procedure for query expansion using the thesaurus shown in Table 7.

Table 7 Thesaurus for Query Expansion

	Japanese	English
Hyponyms	○	○
Abbreviations	○	○
Synonyms	○	—
Related terms	○	—

(1) Expand queries (topic terms) using "Hyponyms" (Section 3.1) with a term weight W_h,

(2) Expand queries using "Abbreviations" (Section 3.2) with a term weight W_a,

(3) Expand queries using "Synonyms" (Section 3.3) with a term weight W_s, (Japanese task only),

(4) Expand queries using "Related terms" (Section 3.4) with a term weight W_r (Japanese task only).

5 Evaluation

5.1 Data and Evaluation

We used 1,685 Japanese topics and 2,221 English topics to evaluate our patent retrieval method [Fujii et al. 2007]. All the systems were evaluated by mean average precision (MAP).

5.2 Systems

Japanese Retrieval Subtask

For the formal run of the Japanese Retrieval Subtask, we submitted the two results provided by systems "hcu1" and "hcu2". The difference between these systems is in the term weights used for the four query expansion methods. The weights for each system are shown in Table 8, and were determined using dry run data. Here, W_t and W_w indicate the weight for topic terms and for other terms in claims, respectively. For comparison with "hcu1" and "hcu2", we prepared a baseline system that omitted the query expansion.

Table 8 Term Weights for Each System (Japanese Retrieval Subtask)

	W_w	W_t	W_a	W_s	W_r	W_h
hcu1	2	5	2	2	2	0
hcu2	2	5	1	2	2	1
baseline	2	5	0	0	0	0

English Retrieval Subtask

For the formal run of the English Retrieval Subtask, we submitted the two results provided by systems "hcu1" and "hcu2". Table 9 gives the term weights for "hcu1" and "hcu2". These weights were determined using sample data that was distributed, instead of dry run data. In the same way as for the Japanese Retrieval Subtask, we prepared a baseline system that omitted the query expansion.

Table 9 Term Weights for Each System (English Retrieval Subtask)

	W_w	W_a	W_h
hcu1	1	1	0
hcu2	1	0	1
baseline	1	0	0

Table 10 Evaluation Results for Japanese Patent Retrieval (%)

def0	all		NTC4		NTC5		NTC6		SR	
	a	b	a	b	a	b	a	b	a	b
hcu1	**9.78**	**6.49**	**20.52**	**15.67**	**9.69**	**8.22**	-	**4.88**	7.39	**7.52**
hcu2	9.62	6.37	19.38	15.30	9.54	8.06	-	4.77	**7.47**	7.47
baseline	6.49	3.77	10.26	9.73	6.88	5.43	-	2.36	3.45	4.30
def1	all		NTC4		NTC5		NTC6		SR	
	a	b	a	b	a	b	a	b	a	b
hcu1	**5.33**	**6.49**	**10.77**	**15.67**	7.01	**8.22**	**4.16**	**4.88**	**6.58**	**7.52**
hcu2	5.26	6.37	10.35	15.30	**7.03**	8.06	4.06	4.77	6.44	7.47
baseline	2.86	3.77	6.52	9.73	4.23	5.43	2.05	2.36	3.06	4.30

Table 11 Evaluation Results for English Patent Retrieval (%)

	a	b
hcu1	**3.37**	**6.94**
hcu2	3.23	6.65
baseline	3.23	6.93

5.3 Results

Japanese Retrieval Subtask

The results are shown in Table 10. For comparison with “hcu1” and “hcu2”, we also show the results for the baseline system that omits the query expansion. As can be seen from Table 10, “hcu1” is superior to the baseline system.

This indicates that “Abbreviations”, “Synonyms”, and “Related Terms” can improve the mean average precision (MAP). “Hcu2” is superior to “hcu1” in SR.a and in NTC5.a, which means that “Hyponym” can also contribute to MAP improvement.

English Retrieval Subtask

The results are shown in Table 11. For comparison with “hcu1” and “hcu2”, we also show the results for the baseline system that omits the query expansion. As can be seen from Table 11, “hcu1” is superior to the baseline system.

From these results, “Abbreviations” is useful not only for Japanese but also for English. Although “hcu2” does not improve upon the baseline method, we cannot conclude that “hyponyms” is not useful, because there were fewer USPTO patents, which were used in constructing the thesaurus, than Japanese patents. We believe that using “hyponyms” still has the potential to improve upon the baseline systems.

5.4 Discussion

To confirm the effectiveness of our method more precisely, we calculated the MAP for each topic and counted the number of topics that our system was superior to the baseline system. The result for Japanese Retrieval Subtask is shown in Table 12. As can be seen from Table 12, both “hcu1” and “hcu2” improved 66 percent of cases in set a, and 73 percent in set b, while impaired in 13 percent in set a, and 17 percent in set b. We show an example topic that our system performed the best in appendix. From the result, we can conclude that our method can correctly expand queries in most cases.

Table 12 The Number of Topics that “hcu1” and “hcu2” could Improve the MAP (Japanese Retrieval Subtask)

		Improve	*Same*	*Impair*
a	hcu1 vs. baseline	1009 (0.667)	305 (0.202)	199 (0.132)
	hcu2 vs. baseline	997 (0.659)	304 (0.201)	212 (0.140)
b	hcu1 vs. baseline	1243 (0.738)	157 (0.093)	285 (0.169)
	hcu2 vs. baseline	1227 (0.728)	161 (0.096)	297 (0.176)

In the same way as for Japanese Retrieval Subtask, we also compared our systems with the baseline system. The result is shown in Table 13. From the result, we can find that performances of “hcu1” and the baseline system are almost the same. “Hcu2” improved 20 percent of cases in both sets a and b, but on the whole, the system impaired the baseline system.

Table 13 The Number of Topics that “hcu1” and “hcu2” could Improve the MAP (English Retrieval Subtask)

		Improve	*Same*	*Impair*
a	hcu1 vs. baseline	115 (0.056)	1826 (0.882)	128 (0.062)
	hcu2 vs. baseline	407 (0.197)	898 (0.434)	765 (0.370)
b	hcu1 vs. baseline	147 (0.066)	1877 (0.845)	199 (0.089)
	hcu2 vs. baseline	465 (0.209)	739 (0.333)	1017 (0.458)

6 Conclusions

We participated in the NTCIR-6 Patent Retrieval Task and evaluated our system for the Japanese and English Retrieval Subtasks. We constructed English-language and Japanese-language thesauri, and used them for query expansion. As a result, we confirmed that “abbreviations”, “synonyms”, and “related terms” are useful for improving the MAP in the Japanese and English Retrieval Subtasks. We also confirmed the effectiveness of “hyponyms” for the Japanese Subtask.

7 Acknowledgment

The author would like to express his gratitude to the organizers of the NTCIR-6 Patent Task for their suggestions to improve his paper. This work was supported by Industrial Technology Research Program in ’06 from New Energy and Industrial Technology Development Organization (NEDO) of Japan.

References

A. Fujii, M. Iwayama, and N. Kando. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. In Proceedings of the Sixth NTCIR Workshop Meeting, 2007.

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

L. Lee. Measures of Distributional Similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 25–32, 1999.

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

H. Mase, T. Matsubayashi, Y. Ogawa, T. Yayoi, Y. Sato, and M. Iwayama. NTCIR-5 Patent Retrieval Experiments at Hitachi. In Proceedings of the fifth NTCIR Workshop Meeting, pp. 318–323, 2005.

E. Morin and C. Jacquemin. Acquisition and Expansion of Hypernym Links. Computer and the Humanities, Vol. 38, No. 4, pp. 343–362, 2004.

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

G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, 1971.

S. Sato and Y. Sasaki. Automatic Collection of Related Terms from the Web. In ACL-03 Companion Volume to the Proceedings of the Conference, pp. 121–124, 2003.

A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases. In Proceedings of the Third NTCIR Workshop, 2002.

Appendix

Topic: 3792

非磁性支持体の少なくとも一方の面上に、電子線硬化性樹脂を含む非磁性層と、該非磁性層上に鉄（Ｆｅ）を主成分とする磁性粉末を含む磁性層を設けてなる**磁気記録媒体**であって、磁性層のガラス転移温度（Ｔｇ）が６５℃以上であり、非磁性層は、電子線硬化性樹脂を含む非磁性層用塗料を非磁性支持体上に塗布して形成した塗膜を電子線硬化してなるものであって、電子線硬化前の塗膜のガラス転移温度（Ｔｇ）が３０℃～５３℃であり、かつ、該電子線硬化性樹脂が、イオウ含有極性基を有する電子線硬化性の塩化ビニル系樹脂と、リン含有極性基を有する電子線硬化性のウレタン樹脂とを含有することを特徴とする**磁気記録媒体**。

(**A magnetic recording system**, which has a non-magnetic layer containing reinforced plastic by electronic beam and a magnetic layer containing magnetic powder, which primarily constituent is Iron (Fe). **A magnetic recording system**, whose glass transition temperature (Tg) in the magnetic layer is over 65°C and ...)

Figure 2 A Sample of Topic

Table 14 Results of Expansion for "磁気記録媒体" (magnetic recording system)

Expansion	Abbreviations	Synonyms	Related Terms	Hyponyms
Weight	W_a	W_s	W_r	W_h
hcu1	2	2	2	0
hcu2	1	2	2	1
1	FD 等 (FD etc.)	磁気記憶装置 (magnetic storage medium)	磁気ディスク (magnetic disc)	磁気シート (magnetic sheet)
2	基盤 (base)	カメラ (camera)	磁気テープ (magnetic tape)	ビデオ用 (for video)
3	磁気ディスク (magnetic disc)	磁気抵抗型ヘッド (magnetoresistive head)	磁気カード (magnetic card)	オーディオ用 (for audio)
4	MO (MO)	薄膜磁気ヘッド (thin film magnetic head)	ディスク (disc)	磁気記録カード (magnetic recording card)
5	APS (APS)	情報記録媒体 (information recording medium)	カード (card)	コンピュータテープ (computer tape)

Figure 2 is a sample of topic that our system improved the MAP in comparison with the baseline system. This topic is about "magnetic recording system", and our system expanded this term in four methods. The expanded terms for each method are shown in Table 14. We also show term weights for each method. As can be seen from Table 14, four methods (especially Synonyms and Related Terms) could expand valid terms. Using these terms in Table 14, both "hcu1" and "hcu2" retrieved two correct patents in ranks one and two (MAP: 0.667), while the baseline system retrieved in ranks 668 and 669 (MAP: 0.002).