Proceedings of the International Symposium on Matching and Meaning, Michael Chan and Fiona McNeill (Eds.),
at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Integrating Technical Terms into Thesaurus using Extracted Related Words

Yoshimi Suzuki and **Fumiyo Fukumoto** [1]

Abstract. Thesaurus expansion is important for most of NLP applications. Because word senses are useful information for them and new words which are not listed in a dictionary appear in documents: web pages, patent documents, and so on. This paper describes a method to integrate new technical terms which appear in patent documents into a thesaurus. To integrate new words into a thesaurus which has hierarchical semantic features, we used some words which are listed in the thesaurus and are extracted as the similar words of the new words. For validity test of the method, we performed some experiments using words which are listed in the thesaurus as test data. And also we performed some experiments using new words. The results suggest that the proposed method is effective for thesaurus expansion.

1 Introduction

There are a large collection of patent documents, and many people with different interests and with different specialized fields. When a person who is not the expert in the area retrieves patents, it is very difficult for him to find the relevant patent documents. Because he may not known the technical terms which are related to the keywords in the area. Also, researchers in same area may not remember keywords in a moment. For example, researchers of a company use 'stall' and researchers of another company use 'hang up'. Therefore, it is important to construct thesaurus of technical terms for the person who is not the expert in the area. For example, using thesaurus of the technical terms, user can use the words which he did not know but are listed in the thesaurus as synonyms of the keyword. Also using the thesaurus, the system can retrieve patent documents using query term expansion.

Currently there are a lot of electronic thesauri, e.g. Roget's Thesaurus, WordNet[5], BunruiGoiHyo [7] and EDR concept dictionary. However, they are thesauri of common words and they have few technical terms or their hierarchal semantic features are not for technical terms. For example, in EDR concept dictionary which has 410 thousands words and has about 8 levels of hierarchy, the hypernym of "collector (one of electrodes of transistor)" is "machine component", and "base" and "emitter" which are also electrodes of transistor are not listed in EDR concept dictionary. EDR concept dictionary which lists 410 thousands words can not be expanded by integrating technical terms, because its classification for technical terms is very coarse. Therefore, it is necessary to construct thesaurus of technical terms.

There are a lot of studies for thesaurus construction and thesaurus expansion. Tokunaga, et al. [11] and Uramoto [8] proposed methods for extending an existing thesaurus by classifying new words in terms of that thesaurus. However, their studies are for commonly used words and not for technical terms. Also there are many studies of ontology mapping, ontology alignment and ontology matching (e.g., [10], [2], [9]). These studies are able to use structure information of both ontology. However, for integrating unknown words into a thesaurus, we can not use semantic structure of unknown words.

[1] University of Yamanashi, Japan, email: ysuzuki@yamanashi.ac.jp

For thesaurus construction or thesaurus expansion, we have to extract similar word pairs. In order to extract similar word pairs, some methods ([3], [4] and [6]) based on dependency relationships are proposed. However, their methods are for commonly-used words and they did not mentioned whether their methods were effective for technical terms.

For extracting similar words of technical terms, we have to deal with the following difficulties.

- Some technical terms appear less frequency.
- Some technical terms are used in same context. Therefore it is difficult to obtain various dependency relationships.

In this paper, for deal with these difficulties, we present a method for detecting word senses of unknown technical terms using extracted coordinate terms of the technical terms and hierarchical semantic features of a concept dictionary. Also we propose a method to integrate new technical terms into a core thesaurus. We perform some experiments in order to confirm effectiveness of the proposed method.

2 System Overview

Figure 1 shows the overview of our system. Our system consists of 4 procedures.

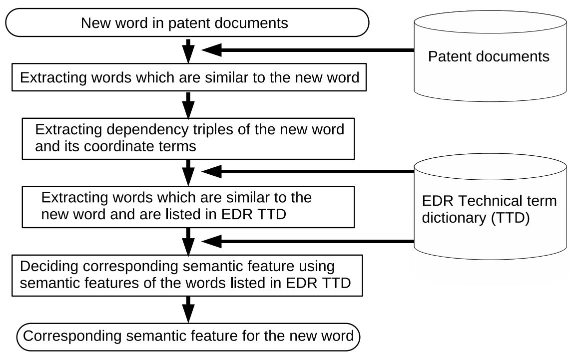

Figure 1. System Overview

We explain 4 procedures of our system using examples of the new technical terms: 'JBIG', 'corpus' and 'packet'.

3 Extracting words Which are Similar to the new word

Some technical terms do not frequently appear in even if large corpora. Table 1 shows frequencies of 5 technical terms in the patent documents. Although the technical terms in Table 1 are technical terms about data compression, 'JPEG' appears frequently, however 'JBIG' appears few times. Therefore, in order to deal with the technical terms like a 'JBIG', we have to extract technical terms which are the coordinate terms of the technical terms.

Table 1. Frequencies of 5 technical terms in the patent documents

Technical terms	code	frequency
JBIG	none	200
JPEG	none	1446
GIF	none	2
image compression	3cd3e6	3643
data compression	3c8276	7783

3.1 Dependency Relationship

Dependency information is used for extracting semantic similar pairs. For example, Lin proposed "dependency triple" [4]. A dependency triple consists 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.

We used 10 kinds of case particles as r. Table 2 illustrates the case particles which we used as r.

Table 2. 10 case particles

case particle	description
ga	nominative case
no	genitive case
wo	accusative case
ni	dative case
he	goal
to	comitative case
kara	elative case
yori	from, at
de	inessive case
ya	coordination

For example, a dependency triple (apply, object, JBIG) is obtained for the unknown word "JBIG". However, some technical terms do not appear frequently, then we have to use hierarchical semantic features for smoothing technique.

4 Extracting Dependency Triples of the new word and Its Coordinate Terms

In order to extract the corresponding semantic feature of the new word, we extract dependency triples of the new word and the extracted words. Table 3 illustrates examples of dependency triples of the extracted words. Using some extracted words, many types of dependency triples are extracted.

Table 3. Examples of dependency triples of the extracted words

w	r	w'
message	object (*wo*)	save
message	object (*wo*)	send
frame	object (*wo*)	discard
information	subject (*ga*)	change
signal	subject (*ga*)	contain

5 Extracting Words Which are Similar to the new word and are Listed in EDR Technical term dictionary

We extract words which are similar to the new word and are listed in the core thesaurus. For extracting the similar words from the core thesaurus, first, $I(w, r, w')$ is calculated using Formula (1).

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

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).

$$Sim(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)} \tag{2}$$

Table 4 shows examples of extracted similar words of "JBIG".

Table 4. Extracted similar words of "JBIG" (top 5)

rank	words in the core thesaurus	similarity
1	image compression	0.331
2	data compression	0.313
3	algorithm	0.182
4	program	0.155
5	command	0.153

6 Detecting Corresponding Semantic Features

Finally, candidates of corresponding semantic features of the new word are detected using the hierarchical semantic features of the core thesaurus. Table 5 illustrates the extracted technical terms which are similar to "packet" and are listed in the core thesaurus. In Table 5, most of top 10 words are technical terms about transmitting data.

Figure 2 illustrates part of hierarchical semantic features associated with "packet". For estimation, we use node distance between each node of the extracted word and node of the new word (packet). We defined node distance was number of arcs from the new word to the node which was the ancestor of the new word and the extracted word. For example, in Figure 2 node distance between "packet" and

Proceedings of the International Symposium on Matching and Meaning, Michael Chan and Fiona McNeill (Eds.),
at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Table 5. Extracted proper nouns for "3c8e70 packet" (top 10)

rank	noun	code	similarity
1	message	3c9112	0.542
2	frame	2e6a69	0.535
3	information	2dd8aa	0.532
4	signal	3c8966	0.523
5	bit	2dee6f	0.442
6	transmit data	2fb213	0.415
7	content	2decb4	0.410
8	image data	3c74fa	0.404
9	received data	2fe07a	0.398
10	block	2def72,3c8561	0.396

"bit" is 1. In Figure 2, average of node distance of top 10 words and "packet" is 2.36 ($= \frac{1+1+2+2+2+2+2+3+3+4+4}{11}$).

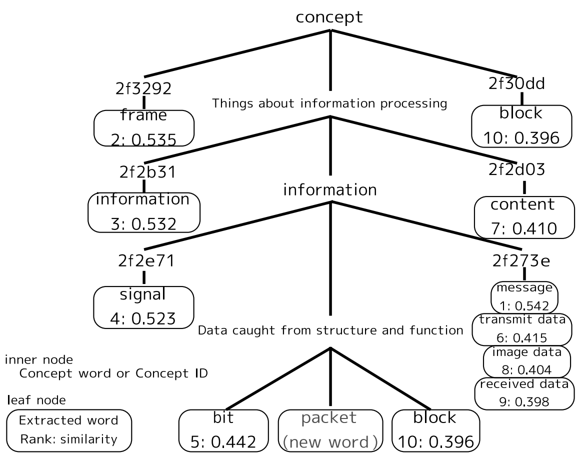

Figure 2. Hierarchical semantic features of "packet" and its similar nouns (top 10)

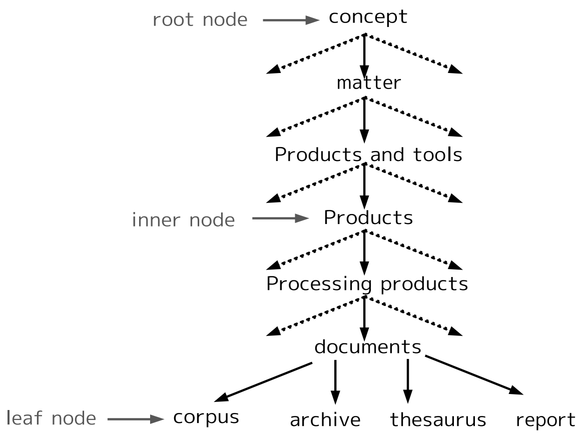

Figure 3. Hierarchical semantic feature of "corpus" in EDR Concept Classification Dictionary of Technical Terms

7 Experiments

7.1 Experimental setup

We used patent publication bulletins written in Japanese (1993) which were provided by patent retrieval task of NTCIR Workshop 4 [1]. The patent publication bulletins consist of about 210 thousands patent documents. The new unknown technical terms are selected from the patent documents.

We used technical term dictionary (domain: information processing) by EDR as a core dictionary. It has about 90 thousands concept classification records. Figure 3 illustrates hierarchic structure of semantic feature of "corpus" in the dictionary..

The depth of the hierarchical semantic features is 6. 'corpus' is one of child nodes of 'document'. The lower semantic features of 'document' are 'corpus', 'archive' and so on.

Figure 4 shows three types of related words and dictionary structure of the core thesaurus. "processing products" is the hypernym of 'documents'. 'corpus', 'archive' and 'thesaurus' are hyponyms of 'documents'. And also these words are coordinate terms of each other.

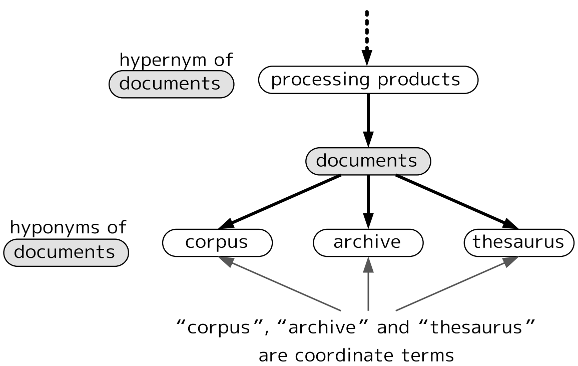

Figure 4. Dictionary structure and related words

7.2 Experimental results and Discussion

We performed experiments of integrating technical terms into a thesaurus using 100 test words. We confirmed that 100 technical terms are classified into the corresponding semantic features by using our method. The technical terms were selected from top 400 words which are frequently searched in "Dictionary of IT technical term 'e-Words'"[2].

Table 6 shows the results.

Table 6. 100 technical terms and their average of node distance

Technical terms	Average of node distance
CCD	4.12
LAN	3.56
LED	3.20
algorithm	3.43
bit	3.30
buffer	3.71
computer	3.22
counter	3.85
database	2.79
data bus	3.33
device	2.61
driver	3.21
flag	1.89
flip-flop	2.66
format	4.05
header	3.52
interface	2.23
low pass filter	4.00
microprocessor	3.87
memory	2.41
modem	1.00
module	3.91
network	3.89
node	2.97
offset	2.92
packet	2.36
parameter	3.11
password	4.83
pixel	3.33
processor	3.85
protocol	4.22
polling	3.54
port	2.34
register	4.75
scanning	4.00
segment	3.88
socket	2.26
software	3.71
subroutine	3.74
throughput	2.62
...	...
average of 100 words	3.54

In Table 6, "average of node distance" indicates average of node distance between the new word and its similar words (top 10 words and their similarities are higher than 0.5) which are listed in the core thesaurus. As the results, we obtained average of node distance was 3.54 using the hierarchical semantic features whose depth is 8.

The worst case was 'password', whose average of node distance was 4.83. Table 7 shows the extracted similar words of 'password'. Although the average of node distance of 'password' is not so good and node distance between 'passcode' and 'password' is 6, 'passcode' and 'password' are similar to each other. From these results, it appears that the core thesaurus is too small and some polysemous words are unconsidered.

Table 7. Similar words of 'password'

rank	similar word	similarity	node distance
1	passcode	0.548	6
2	identification number	0.538	6
3	number	0.534	5
4	ID	0.532	5
5	ID number	0.521	2
6	identification code	0.509	2

Table 8 shows the results of the words which are not listed in the core thesaurus. As the results a node is selected as the hypernym the words.

Table 8. Unknown words and the selected node as the hypernym of the words

num	word	extracted hypernym
1	filament	parts of hardware
2	solar battery	parts of hardware

Figure 5 shows how to integrate new word: 'filament' into core thesaurus. For deciding the hypernym of 'filament', we trace links from root node to a leaf node. If a node has more than one child node, we select a child node whose subtree has extracted words more than other subtrees. If the extracted words in the subtree are less than 2, the nodes are decided as the hypernym of the new words.

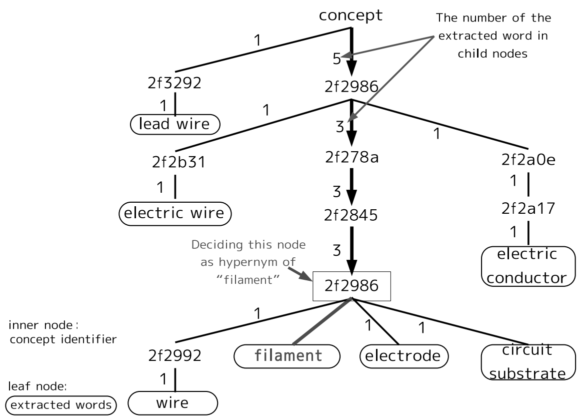

Figure 5. Experimental Results: "filament"

Although we use very few test data and use a core thesaurus which has coarse classification, we can integrate new technical terms into a thesaurus.

8 Conclusion

This paper proposed a method to integrate new technical terms into a thesaurus using extracted similar words of the technical terms. Using the extracted words from patent documents and hierarchical structure of a technical term dictionary, we showed the possibility of thesaurus expansion.

[2] http://e-words.jp/

9 Acknowledgments

This work was supported by the Grant-in-aid for the Japan Society for the Promotion of Science (JSPS).

REFERENCES

[1] A.Fujii, M.Iwayama, and N.Kando, 'Test collection for patent-to-patent retrieval and patent map generation in ntcir-4 workshop', in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1643–1646, (2004).
[2] Domenico Beneventano and Daniele Montanari, 'Ontological mappings of product catalogues', in *In Proceedings of The Third International Workshop on Ontology Matching*, (2008).
[3] D.Hindle, 'Noun classification from predicate-argument structures', in *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275, (1990).
[4] D.Lin, 'Automatic retrieval and clustering of similar words', in *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).
[5] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, Bradford Books, 1998.
[6] M. Hagiwara, Y. Ogawa, and K. Toyama, 'Selection of effective contextual information for automatic synonym acquisition', in *In Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 353–360, (2006).
[7] National Language Research Institute, *Bunruigoihyo*, Shuei publisher (In Japanese), 1964.
[8] N.Uramoto, 'Positioning unknown words in a thesaurus by using information extracted from a corpus', in *In proceedings of COLING'96*, pp. 956–961, (1996).
[9] Dominique Ritze, Christian Meilicke, Ondrej Svab-Zamazal, and Heiner Stuckenschmidt, 'A pattern-based ontology matching approach for detecting complex correspondences', in *In Proceedings of The 4th International Workshop on Ontology Matching*, (2008).
[10] Konstantin Todorov and Peter Geibel, 'Ontology mapping via structural and instance-based similarity measures', in *In Proceedings of The Third International Workshop on Ontology Matching*, (2008).
[11] T.Tokunaga, 'Extending a thesaurus by classifying words', in *In Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pp. 16–21, (1997).