

Detecting unknown word senses using concept dictionary

Yoshimi Suzuki¹ and Fumiyo Fukumoto

Abstract. In this paper, we present a method for detecting unknown word senses using a concept dictionary and newspaper articles. It is very important to detecting unknown word senses for document classification, information retrieval, information extraction, etc. Although for extracting similar word pairs the methods which use similarity of case structure between two words are used, comparison between similarity of two words suffers word sparseness problem. Especially, it is necessary to solve this problem for detecting word senses of proper nouns which are not listed in the dictionary. The proposed method used hierarchical semantic features of a concept dictionary in order to deal with this problem. We performed some experiments in order to confirm effectiveness of the method.

1 Introduction

In newspaper, web pages, patent documents, etc., many different proper nouns frequently appear, and furthermore, new proper nouns are generated in these documents day after day. Most of proper nouns did not given in dictionaries, therefore we have to detect their senses for some applications of natural language processing: thesaurus construction and the rest. For recognizing proper nouns, named entity extraction and named entity recognition are studied frequently. For named entity extraction and recognition, machine learning techniques are used in order to categorize each named entity into some types [1]: ORGANIZATION, PERSON, LOCATION ARTIFACT, DATE, TIME, MONEY and PERCENT. Sekine proposed extended 200 named entity categories [2]. The extended categories are effective for QA application. But the types may be coarse for detecting unknown word senses.

Thesaurus construction is also studied actively. Many studies use context information for similar noun pairs [3], and many similar noun pairs of common nouns can be extracted accurately. But it is difficult to extract pairs of proper noun and common noun which are semantically similar. Because some proper nouns do not frequently appear and we can extract very few information of dependency relationship for the proper nouns.

In this paper, we present a method for detecting unknown word senses using concept dictionary and newspaper articles. Although the methods which use similarity of case structure between two words are used for extracting similar word pairs, calculating similarity between two words suffers word sparseness problem. Especially, it is necessary to solve this problem for detecting word senses of proper nouns which are not listed in the dictionary.

The proposed method used hierarchical semantic features of a concept dictionary in order to deal with this problem. We performed some experiments in order to confirm effectiveness of the method.

2 System Overview

Figure 1 illustrates the overview of our system. Our system consists of 5 steps which are described in the following list.

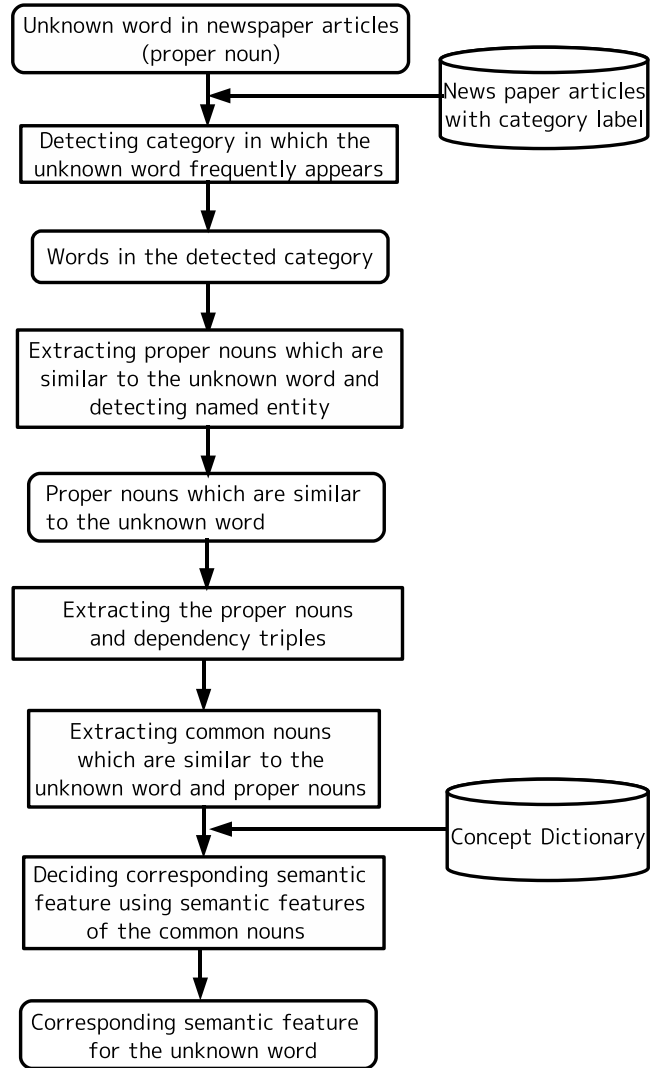


Figure 1. System Overview

1. Detecting the category in which the target unknown words appear, and extracting dependency pattern which include target unknown word.

¹ University of Yamanashi, Kofu 400-8511, Japan, email: ysuzuki@yamanashi.ac.jp

2. Extracting similar proper nouns of the target word from the detected category. Then extracting named entity information of the target unknown word using a CaboCha [4], and extracting words whose dependency pattern of the words which are categorized into same named entity category.
3. Extracting dependency triples of the extracted proper nouns.
4. Extracting common nouns which are similar to the unknown word and its similar proper nouns.
5. Deciding corresponding word senses using semantic features of the extracted common nouns.

We explain our system step by step with an example: “Tigers”. “Tigers” is not listed the dictionary, but must be “baseball team”.

3 Detecting Category

The aim of our study is to detect the appropriate semantic feature of target unknown words. Some proper nouns have different meanings in different categories. For example, “Tigers” illustrates names of baseball teams and also a name of musical band. Therefore “Tigers” has different dependency pattern in different category.

Table 1 shows frequency of appearance of “Tigers” and similar proper nouns in each category of the newspaper articles. In Table 1, most of “Tigers” appear in Sports category. Therefore we used newspaper articles in Sports category for deciding semantic feature of “Tigers” appeared in Sports.

Table 1. Frequency of appearance in each category

Category	Tigers	Giants	Twins	Mariners
1st page	70	65	20	304
2nd page	26	13	0	25
3rd page	28	2	7	30
Sports	1207	2163	1269	4942
Home	5	2	2	3
Science	0	0	1	1
Commentary	26	9	0	38
Economy	36	3	0	17
Show biz	20	2	0	7
Int'l	35	31	34	193
Local	406	381	70	760
Editorial	3	9	0	10
General	96	100	12	151
Special	45	123	28	191
Reading	6	7	0	4
Culture	0	0	0	2

It is necessary to extract word candidates with category information, because semantic features of concept dictionary are classified without their categories.

4 Extracting proper nouns which are similar to the unknown word

Some proper nouns do not frequently appear in even if large corpora. Table 2 shows frequencies of the team names of Japan Professional Baseball in the newspaper articles of The Mainichi Newspapers (2000). Although the 6 teams in Table 2 are members of the same league, “Tigers” appears frequently, however “Swallows” appears 3 times. Therefore in order to deal with the proper nouns like a “Swallows”, we have to extract proper nouns which are same members of coordinate terms.

Table 3 illustrates the extracted proper nouns which are similar to “Tigers”. In table 3 All of top 5 are baseball teams.

Table 2. Frequencies of 6 team names of Japan Professional Baseball in the articles of The Mainichi Newspaper (2000)

Team Names	frequency
Giants	182
Tigers	398
Dragons	22
Carp	34
Swallows	3
Baystars	49

Table 3. Extracted proper nouns for “Tigers” (top 5)

rank	proper noun	similarity
1	Orix (baseball team)	0.448
2	Seibu (baseball team)	0.439
3	Yokohama (basebal team)	0.417
4	Chunichi (baseball team)	0.416
5	Kyojin (baseball team)	0.400

4.1 Dependency relationship

Dependency information is used for extracting semantic similar pairs. For example, Lin proposed “dependency triple” [5]. 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.

When it applies for the unknown word “Tigers”, (quit, object, Tigers) is obtained.

But most of the unknown words do not appear frequently, then we have to use hierarchical semantic feature for smoothing technique.

5 Extracting dependency triples of the proper nouns

In order to extract corresponding common nouns, we extract dependency triples of the extracted proper nouns. Table 4 illustrates examples of dependency triples of the extracted proper nouns. Using some extracted proper nouns, many types of dependency triples are extracted.

Table 4. Examples of dependency triples of the extracted proper nouns

w	r	w'
Chunichi	subject (<i>ga</i>)	escape (the cellar)
Hanshin	subject (<i>ga</i>)	rise (to second place)
Seibu	subject (<i>ga</i>)	run away
Taiyo	object (<i>wo</i>)	steamroller
Orix	to (<i>he</i>)	transfer

6 Extracting common nouns

We extract common nouns which are similar to the unknown word and the extracted proper nouns. Table 5 shows examples of extracted common nouns for “Tigers”.

7 Detecting corresponding semantic features

Finally, candidates of corresponding semantic features of the unknown word using the concept dictionary.

Table 5. Extracted common nouns for “Tigers” (top 5)

rank	common noun	similarity
1	team	0.382
2	delegate	0.321
3	player	0.313
4	excellent team	0.185
5	horse	0.157

7.1 Hierarchical semantic features

We used hierarchical semantic features made by EDR [6]. Table 6 shows hierarchical semantic features of “baseball team”.

Table 6. Hierarchical semantic structure (baseball team)

depth level	code	label
6	3c1e0f	team
5	444999	colleague, team mate
4	30f6b1	associate
3	3cfacc	group
2	3aa912	active object
1	3aa911	human being
0	3aa966	concept

8 Experiments

We confirmed that an unknown word “Tigers” correspond to baseball team by using our method.

8.1 Experimental setup

We used newspaper articles of The Mainichi Newspapers (from 1991 to 2004, written in Japanese). There are about 500 thousands articles. Unknown words are selected from newspaper articles of The Mainichi Newspapers (from 2000 to 2004, written in Japanese). We used concept dictionary by EDR. It has about 410 thousands concept classification records.

8.2 Experimental results and Discussion

Although we use very few test data, we can detect an unknown word sense using our method. At each step, we evaluate the results. All test data frequently appear in one category, by contrast, in other categories they do not frequently appear. When the unknown word appeared frequently in some categories, the unknown word has some different meanings. For example, “Tigers” meant it any place other than a baseball team. Therefore, we found that it was important to use the articles which are classified into categories.

9 Conclusion

This paper proposed a method to detect unknown word senses using concept dictionary. We used concept dictionary by EDR, categorized newspaper articles, and we showed the possibility of detecting unknown word senses using the proposed method. In the future, we plan to the following.

- Using test data to confirm the method is effective.
- Using machine learning technique.
- Using Web documents for collecting .
- Detecting unknown word senses in patent document

REFERENCES

- [1] Satoshi, Sekine., Hitoshi, Isahara., “IREX: IR and IE Evaluation project in Japanese”, In Proceedings of the Forth International conference on Language Resources and Evaluation, 2000.
- [2] Satoshi, Sekine., Chikashi Nobata., “Definition, dictionaries and tagger for Extended Namd Entity Hierarchy”, In Proceedings of the Forth International conference on Language Resources and Evaluation, pp.1977-1980, 2004.
- [3] Hagiwara, M., Ogawa, Y., Toyama, K.: Selection of Effective Contextual Information for Automatic Synonym Acquisition. In Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 353-360 (2006)
- [4] 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.
- [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] EDR ELECTRONIC DICTIONARY VERSION 2.0 TECHNICAL GUIDE, National Institute of Information and Communications Technology (1996)