Construction of Domain Dictionary for Fundamental Vocabulary

Chikara Hashimoto
Faculty of Engineering,
Yamagata University
4-3-16 Jonan, Yonezawa-shi, Yamagata,
992-8510 Japan

Sadao Kurohashi
Graduate School of Informatics,
Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto,
606-8501 Japan

Abstract

For natural language understanding, it is essential to reveal semantic relations between words. To date, only the IS-A relation has been publicly available. Toward deeper natural language understanding, we semi-automatically constructed the domain dictionary that represents the domain relation between Japanese fundamental words. This is the first Japanese domain resource that is fully available. Besides, our method does not require a document collection, which is indispensable for keyword extraction techniques but is hard to obtain. As a task-based evaluation, we performed blog categorization. Also, we developed a technique for estimating the domain of unknown words.

1 Introduction

We constructed a lexical resource that represents the domain relation among Japanese fundamental words (JFWs), and we call it the **domain dictionary**.[1] It associates JFWs with domains in which they are typically used. For example, ホームラン *home run* is associated with the domain SPORTS[2]. That is, we aim to make explicit the horizontal relation between words, the domain relation, while thesauri indicate the vertical relation called IS-A.[3]

[1]In fact, there have been a few domain resources in Japanese like Yoshimoto et al. (1997). But they are not publicly available.

[2]Domains are CAPITALIZED in this paper.

[3]The lack of the horizontal relationship is also known as the "tennis problem" (Fellbaum, 1998, p.10).

2 Two Issues

You have to address two issues. One is what domains to assume, and the other is how to associate words with domains without document collections.

The former is paraphrased as how people categorize the real world, which is really a hard problem. In this study, we avoid being too involved in the problem and adopt a simple domain system that most people can agree on, which is as follows:

CULTURE	LIVING	SCIENCE
RECREATION	DIET	BUSINESS
SPORTS	TRANSPORTATION	MEDIA
HEALTH	EDUCATION	GOVERNMENT

It has been created based on web directories such as Open Directory Project with some adjustments. In addition, NODOMAIN was prepared for those words that do not belong to any particular domain.

As for the latter issue, you might use keyword extraction techniques; identifying words that represent a domain from the document collection using statistical measures like TF*IDF and matching between extracted words and JFWs. However, you will find that document collections of common domains such as those assumed here are hard to obtain.[4] Hence, we had to develop a method that does not require document collections. The next section details it.

[4]Initially, we tried collecting web pages in Yahoo! JAPAN. However, we found that most of them were index pages with a few text contents, from which you cannot extract reliable keywords. Though we further tried following links in those index pages to acquire enough texts, extracted words turned out to be site-specific rather than domain-specific since many pages were collected from a particular web site.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 137–140,
Prague, June 2007.

Table 1: Examples of Keywords for each Domain

Domain	Examples of Keywords
CULTURE	映画 *movie,* 音楽 *music*
RECREATION	観光 *tourism,* 花火 *firework*
SPORTS	選手 *player,* 野球 *baseball*
HEALTH	手術 *surgery,* 診断 *diagnosis*
LIVING	育児 *childcare,* 家具 *furniture*
DIET	箸 *chopsticks,* 昼食 *lunch*
TRANSPORTATION	駅 *station,* 道路 *road*
EDUCATION	先生 *teacher,* 算数 *arithmetic*
SCIENCE	研究 *research,* 理論 *theory*
BUSINESS	輸入 *import,* 市場 *market*
MEDIA	放送 *broadcast,* 記者 *reporter*
GOVERNMENT	司法 *judicatory,* 税 *tax*

3 Domain Dictionary Construction

To identify which domain a JFW is associated with, we use manually-prepared keywords for each domain rather than document collections. The construction process is as follows: ① Preparing keywords for each domain (§3.1). ② Associating JFWs with domains (§3.2). ③ Reassociating JFWs with NODOMAIN (§3.3). ④ Manual correction (§3.5).

3.1 Preparing Keywords for each Domain

About 20 keywords for each domain were collected manually from words that appear most frequently in the Web. Table 1 shows examples of the keywords.

3.2 Associating JFWs with Domains

A JFW is associated with a domain of the highest A_d score. An A_d score of domain is calculated by summing up the top five A_k scores of the domain. Then, an A_k score, which is defined between a JFW and a keyword of a domain, is a measure that shows how strongly the JFW and the keyword are related (Figure 1). Assuming that two words are related if they cooccur more often than chance in a corpus, we adopt the χ^2 statistics to calculate an A_k score and use web pages as a corpus. The number of co-occurrences is approximated by the number of search engine hits when the two words are used as queries. Among various alternatives, the combination of the χ^2 statistics and web pages is adopted following Sasaki et al. (2006).

Based on Sasaki et al. (2006), A_k score between

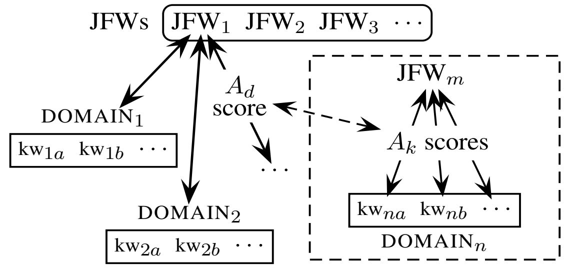

Figure 1: Associating JFWs with Domains

a JFW (jw) and a keyword (kw) is given as below.

$$A_k(jw, kw) = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where n is the total number of Japanese web pages,

$$a = hits(jw \,\&\, kw), \quad b = hits(jw) - a,$$
$$c = hits(kw) - a, \quad d = n - (a + b + c).$$

Note that $hits(q)$ represents the number of search engine hits when q is used as a query.

3.3 Reassociating JFWs with NODOMAIN

JFWs that do not belong to any particular domain, i.e. whose highest A_d score is low should be reassociated with NODOMAIN. Thus, a threshold for determining if a JFW's highest A_d score is low is required. The threshold for a JFW (jw) needs to be changed according to $hits(jw)$; the greater $hits(jw)$ is, the higher the threshold should be.

To establish a function that takes jw and returns the appropriate threshold for it, the following semi-automatic process is required after all JFWs are associated with domains: **(i)** Sort all tuples of the form $< jw$, $hits(jw)$, the highest A_d of the $jw >$ by $hits(jw)$.[5] **(ii)** Segment the tuples. **(iii)** For each segment, extract manually tuples whose jw should be associated with one of the 12 domains and those whose jw should be deemed as NODOMAIN. Note that the former tuples usually have higher A_d scores than the latter tuples. **(iv)** For each segment, identify a threshold that distinguishes between the former tuples and the latter tuples by their A_d scores. At this point, pairs of the number of hits (represented by each segment) and the appropriate threshold for it are obtained. **(v)** Approximate the relation between

[5]Note that we acquire the number of search engine hits and the A_d score for each jw in the process ②.

the number of hits and its threshold by a linear function using least-square method. Finally, this function indicates the appropriate threshold for each jw.

3.4 Performance of the Proposed Method

We applied the method to JFWs installed on JUMAN (Kurohashi et al., 1994), which are 26,658 words consisting of commonly used nouns and verbs. As an evaluation, we sampled 380 pairs of a JFW and its domain, and measured accuracy.[6] As a result, the proposed method attained the accuracy of 81.3% (309/380).

3.5 Manual Correction

Our policy is that simpler is better. Thus, as one of our guidelines for manual correction, we avoid associating a JFW with multiple domains as far as possible. JFWs to associate with multiple domains are restricted to those that are EQUALLY relevant to more than one domain.

4 Blog Categorization

As a task-based evaluation, we categorized blog articles into the domains assumed here.

4.1 Categorization Method

(i) Extract JFWs from the article. **(ii)** Classify the extracted JFWs into the domains using the domain dictionary. **(iii)** Sort the domains by the number of JFWs classified in descending order. **(iv)** Categorize the article as the top domain. If the top domain is NODOMAIN, the article is categorized as the second domain under the condition below.

$$|W(\text{2ND DOMAIN})| \div |W(\text{NODOMAIN})| > 0.03$$

where $|W(\text{D})|$ is the number of JFWs classified into the domain D.

4.2 Data

We prepared two blog collections; $B_{controlled}$ and B_{random}. As $B_{controlled}$, 39 blog articles were collected (3 articles for each domain including NODOMAIN) by the following procedure: **(i)** Query the Web using a keyword of the domain.[7] **(ii)** From the top of the search result, collect 3 articles that meet the following conditions; there are enough text contents in it, and people can confidently make a judgment about which domain it is categorized as. As B_{random}, 30 articles were randomly sampled from the Web. Table 2 shows its breakdown.

Table 2: Breakdown of B_{random}

Domain	#	Domain	#
CULTURE	4	DIET	4
RECREATION	1	BUSINESS	12
SPORTS	3	NODOMAIN	5
HEALTH	1		

Note that we manually removed peripheral contents like author profiles or banner advertisements from the articles in both $B_{controlled}$ and B_{random}.

4.3 Result

We measured the accuracy of blog categorization. As a result, the accuracy of 89.7% (35/39) was attained in categorizing $B_{controlled}$, while B_{random} was categorized with 76.6% (23/30) accuracy.

5 Domain Estimation for Unknown Words

We developed an automatic way of estimating the domain of unknown word (uw) using the dictionary.

5.1 Estimation Method

(i) Search the Web by using uw as a query. **(ii)** Retrieve the top 30 documents of the search result. **(iii)** Categorize the documents as one of the domains by the method described in §4.1. **(iv)** Sort the domains by the number of documents in descending order. **(v)** Associate uw with the top domain.

5.2 Experimental Condition

(i) Select 10 words from the domain dictionary for each domain. **(ii)** For each word, estimate its domain by the method in §5.1 after removing the word from the dictionary so that the word is unknown.

5.3 Result

Table 3 shows the number of correctly domain-estimated words (out of 10) for each domain. Accordingly, the total accuracy is 67.5% (81/120).

[6] In the evaluation, one of the authors judged the correctness of each pair.

[7] To collect articles that are categorized as NODOMAIN, we used 日記 *diary* as a query.

Table 3: # of Correctly Domain-estimated Words

Domain	#	Domain	#
CULTURE	7	TRANSPORTATION	7
RECREATION	4	EDUCATION	9
SPORTS	9	SCIENCE	6
HEALTH	9	BUSINESS	9
LIVING	3	MEDIA	2
DIET	7	GOVERNMENT	9

As for the poor accuracy for RECREATION, LIVING, and MEDIA, we found that it was due to either the ambiguous nature of the words of domain or a characteristic of the estimation method. The former brought about the poor accuracy for MEDIA. That is, some words of MEDIA are often used in other contexts. For example, 中継 *live coverage* is often used in the SPORTS context. On the other hand, the method worked poorly for RECREATION and LIVING for the latter reason; the method exploits the Web. Namely, some words of the domains, such as 観光 *tourism* and シャンプー *shampoo*, are often used in the web sites of companies (BUSINESS) that provide services or goods related to RECREATION or LIVING. As a result, the method tends to wrongly associate those words with BUSINESS.

6 Related Work

HowNet (Dong and Dong, 2006) and WordNet provide domain information for Chinese and English, but there has been no domain resource for Japanese that are publicly available.[8]

Domain dictionary construction methods that have been developed so far are all based on highly structured lexical resources like LDOCE or WordNet (Guthrie et al., 1991; Agirre et al., 2001) and hence not applicable to languages for which such highly structured lexical resources are not available.

Accordingly, contributions of this study are twofold: **(i)** We constructed the first Japanese domain dictionary that is fully available. **(ii)** We developed the domain dictionary construction method that requires neither document collections nor highly structured lexical resources.

[8]Some human-oriented dictionaries provide domain information. However, domains they cover are all technical ones rather than common domains such as those assumed here.

7 Conclusion

Toward deeper natural language understanding, we constructed the first Japanese domain dictionary that contains 26,658 JFWs. Our method requires neither document collections nor structured lexical resources. The domain dictionary can satisfactorily classify blog articles into the 12 domains assumed in this study. Also, the dictionary can reliably estimate the domain of unknown words except for words that are ambiguous in terms of domains and those that appear frequently in web sites of companies.

Among our future work is to deal with domain information of multiword expressions. For example, 源泉 *fount* and 徴収 *collection* constitute 源泉徴収 *tax deduction at source*. Note that while 源泉 itself belongs to NODOMAIN, 源泉徴収 should be associated with GOVERNMENT.

Also, we will install the domain dictionary on JUMAN (Kurohashi et al., 1994) to make the domain information fully and easily available.

References

Eneko Agirre, Olatz Ansa, David Martinez, and Ed Hovy. 2001. Enriching wordnet concepts with topic signatures. In *Proceedings of the SIGLEX Workshop on "WordNet and Other Lexical Resources: Applications, Extensions, and Customizations" in conjunction with NAACL*.

Zhendong Dong and Qiang Dong. 2006. *HowNet And the Computation of Meaning*. World Scientific Pub Co Inc.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad. 1991. Subject-Dependent Co-Occurence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 146–152.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese Mophological Analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 22–28.

Yasuhiro Sasaki, Satoshi Sato, and Takehito Utsuro. 2006. Related Term Collection. *Journal of Natural Language Processing*, 13(3):151–176. (in Japanese).

Yumiko Yoshimoto, Satoshi Kinoshita, and Miwako Shimazu. 1997. Processing of proper nouns and use of estimated subject area for web page translation. In *tmi97*, pages 10–18, Santa Fe.