Construction of a Domain Dictionary for Fundamental Vocabulary and its Application to Automatic Blog Categorization Using Dynamically Estimated Domains of Unknown Words

Chikara Hashimoto† and Sadao Kurohashi‡

The semantic relations between words are essential for natural language understanding. Toward deeper natural language understanding, we semi-automatically constructed a domain dictionary that represents the domain relations between fundamental Japanese words. Our method does not require a document collection. As a task-based evaluation of the domain dictionary, we categorized blogs by assigning a domain for each word in a blog article and categorizing it as the most dominant domain. Thus, we dynamically estimated the domains of unknown words, (i.e., those not listed in the domain dictionary), resulting in our blog categorization achieving an accuracy of 94.0% (564/600). Moreover, the domain estimation technique for unknown words achieved an accuracy of 76.6% (383/500).

Key Words: domain, lexicon, blog, text categorization, unknown words’ domain

1 Introduction

For deep semantic processing of language, a thesaurus is an indispensable resource. Thesauri represent is-a relations between words. An is-a relation is, so to speak, a “vertical” relation, and we believe that “horizontal” semantic relationships between words are also required to fully capture the meaning of words. Accordingly, we propose the domain relation in this paper. For example, textbook and teacher belong to the education domain, whereas kitchen knife and surgical knife belong to the diet health domains, respectively. We have constructed a domain dictionary in which we have assigned domains to approximately 30,000 fundamental Japanese words are given appropriate domains.

Domain information enables a more natural semantic classification of words. For example, a thesaurus would regard textbook as a book and teacher as a profession; however both words are

† National Institute of Information and Communications Technology
‡ Graduate School of Informatics, Kyoto University
* This is an augmented edition of (Hashimoto and Kurohashi 2007, 2008); the authors augmented the material with additional experiments and detailed descriptions.
** This article has been partially revised for better understanding of overseas readers.
part of the same domain, i.e., the education domain. In contrast, a thesaurus would regard both kitchen knife and surgical knife as knives; however, they should be distinguished by their domains, i.e., the diet domain for the former and the health domain for the latter.

Domain information has been used for various natural language processing (NLP) tasks. We use it for blog categorization, but it has also been used for document filtering (Liddy and Paik 1993), word-sense disambiguation (Rigau, Atserias, and Agirre 1997; Tanaka, Bond, Baldwin, Fujita, and Hashimoto 2007), and machine translation (Yoshimoto, Kinoshita, and Shimazu 1997; Lange and Yang 1999).

Our method for constructing the domain dictionary is semi-automatic. First, we automatically construct a domain dictionary by measuring the degree to which a target fundamental word and a domain keyword are semantically related. Then, we manually correct the automatically constructed results. For measuring semantic relatedness, we exploit hit counts returned by a web search engine. We adopt a semi-automatic process because semantic information about fundamental words is crucial for many NLP tasks and must therefore be highly accurate. However, it is difficult to achieve such high accuracy using a fully automatic process. But, it is also undesirable to manually define all semantic relations because of the high cost, low consistency, and low maintainability of manual work. Hence, we decided to manually correct reasonably accurate and automatically constructed results.¹

Our domain dictionary is the first publicly available Japanese domain dictionary. Our domain dictionary construction method only requires an access to a web search engine; no training data or any other language resource is required.

We also propose a method for dynamically estimating a domain for unknown words on-the-fly using a web search engine.

As a task-based evaluation of our domain dictionary construction method and our domain estimation method for unknown words, we conduct automatic blog categorization. We categorize blog articles into one of the domains assumed using the domain information of words in blog articles. As a result, we achieve 94.0% accuracy for blog categorization and 76.6% for unknown words' domain estimation.

In this study, our domain dictionary contains only fundamental words used in daily life and does not contain technical terms.²

¹ For a similar reason, Kyoto University Text Corpus (http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Kyoto%20University%20Text%20Corpus) has been constructed in a semi-automatic way using a precise Japanese dependency parser called the Kurohashi-Nagao Parser (KNP) (http://nlp.ist.i.kyoto-u.ac.jp/EN/).
² To be more precise, about 30,000 content words from the JUMAN dictionary (Kurohashi, Nakamura, Matsumoto, and Nagao 1994) (version 5.1) were used for constructing our domain dictionary.
The rest of the paper is organized as follows. In Section 2, we describe issues related to the construction of the domain dictionary. Section 3 presents our domain dictionary construction method. Section 4 reports the details of the domain dictionary we constructed. In Section 5, we describe our blog categorization method, and in Section 6, we describe our domain estimation method for unknown words. In Section 7, we report evaluation results of our blog categorization method and domain estimation method for unknown words. In Section 8, we compare our study with previous ones, and in Section 9, we conclude the paper.

2 Two Issues

During creation of semantic domains, two issues must be addressed: which domains to create and how to associate words with domains without document collections.

Ideally, the choice of semantic relations should be based on how people understand and categorize the real world, understanding of which is really a challenging problem. In this study, without getting too involved in the problem, we adopt simple domain categories (listed in Table 1) that most people can agree upon. It has been created based on web directories, such as the Open Directory Project (http://www.dmoz.org) with some adjustments. In addition, NODOMAIN was prepared for words not belonging to any particular domain.

The association of words with domains can be done using standard keyword extraction techniques; identifying words that represent a domain from the document collection of the domain using statistical measures like TF*IDF and matching between those extracted words and the target fundamental words. However, document collections of common domains, such as those assumed in this study, are harder to obtain than those of technical or specialized domains. Web directories like the Open Directory Project or Yahoo! JAPAN (http://dir.yahoo.co.jp) might seem like good sources of such document collections. However, when we tried collecting web pages registered in Yahoo! JAPAN, we found that most of them were index pages with little text content from which reliable keywords could not be extracted. Though we further tried following the links on the index pages to acquire enough text content and extracting words from them,

<table>
<thead>
<tr>
<th>CULTURE</th>
<th>LIVING</th>
<th>SCIENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECREATION</td>
<td>DIET</td>
<td>BUSINESS</td>
</tr>
<tr>
<td>SPORTS</td>
<td>TRANSPORTATION</td>
<td>MEDIA</td>
</tr>
<tr>
<td>HEALTH</td>
<td>EDUCATION</td>
<td>GOVERNMENT</td>
</tr>
</tbody>
</table>

Table 1 Domains in this study
some of those words turned out to be site-specific rather than domain-specific because many pages were collected from a particular web site. Therefore, we had to develop a method not requiring document collections of domains. The next section details our method.

3 Domain Dictionary Construction

To identify which domain a fundamental word is associated with, we use manually prepared keywords for each domain rather than document collections.

First, each domain is represented by 20 to 30 keywords, as described in Section 3.1. Then, an association score between each fundamental word and a domain ($A_d$ score) is calculated by summing up the association scores between the fundamental word and the keywords of the domain ($A_k$ scores). After calculating $A_d$ scores between the fundamental word and all domains, the fundamental word is associated with the domain with the highest $A_d$ score. This process is repeated for each fundamental word. Then, some of the fundamental words can be reassigned with NODOMAIN if their $A_d$ scores are low. This association process is described in Section 3.2. Finally, as described in Section 3.6, we corrected the association results for the association dictionary. Figure 1 illustrates the complete construction process.

Fig. 1 Whole Construction Process (“JFW” and “kw” stand for Japanese fundamental word and keyword respectively)

---

3 One more obstacles faced while using web directories as document collections is the fact that on web pages, there are sometimes banner advertisement links having little to do with a target domain.
Note that the method for constructing the domain dictionary is independent from the 12 domains specified in Table 1; you can use the same method with different domains.

### 3.1 Preparing Keywords for each Domain

Keywords for each domain are collected manually from the list of words appearing most frequently on the web. To be precise, starting from the top of the list, words are chosen as keywords of a domain if they represent the domain. This process is repeated until 20 to 30 words are collected per domain. If we are uncertain about which domain a word should belong to, we ignore the word.\(^4\) Table 2 shows examples of the keywords.

If you adopt different domains, you must collect appropriate keywords for each domain yourself. However, after that, the same procedure can be applied to them.

### 3.2 Associating Fundamental Words with Domains

A fundamental word is associated with the domain with the highest \(A_d\) score. The \(A_d\) score of the domain is calculated by summing up the top five \(A_k\) scores for the domain.\(^5\) An \(A_k\) score, which is defined between a fundamental word and a keyword of a domain, is a measure that shows how strongly the fundamental word and the keyword are related. Assuming that two words are related if they co-occur more often than chance in a corpus, we adopt \(\chi^2\) statistics to calculate an

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

\(^4\) This has been conducted by one of the authors. In our future work, we plan to examine how (in)consistent this judgment is between annotators.

\(^5\) An experiment we conducted showed that this yielded a better results than summing up all \(A_k\) scores.
$A_k$ 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.\footnote{We used Yahoo! JAPAN (www.yahoo.co.jp).} We chose to combine $\chi^2$ statistics and web pages because Sasaki, Sato, and Utsuro (2006) reports that this combination provides good results during estimation of term relatedness.

Following Sasaki et al. (2006), the $A_k$ score between a fundamental word ($jw$) and a keyword ($kw$) is

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

where $n$ is the total number of Japanese web pages\footnote{We used 10,000,000,000 as $n$.} and

\[
\begin{align*}
    a &= \text{hits}(jw \& kw), \\
    b &= \text{hits}(jw) - a, \\
    c &= \text{hits}(kw) - a, \\
    d &= n - (a + b + c).
\end{align*}
\]

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

### 3.3 Reassociating Fundamental Words with NODOMAIN

Fundamental words not belonging to any particular domain, (i.e., fundamental words with low $A_d$ scores) should be reassociated with NODOMAIN. Accordingly, a low $A_d$ score threshold must be established, below which fundamental words should be reassociated. A pilot study we conducted showed that the threshold for a fundamental word ($jw$) needs to be changed according to $\text{hits}(jw)$; the higher the $\text{hits}(jw)$, 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 fundamental words are associated with domains:

(i) Sort all tuples of \{ $jw$, $\text{hits}(jw)$, the highest $A_d$ of the $jw$ \} by $\text{hits}(jw)$.\footnote{Note that we acquired the number of search engine hits and the $A_d$ score for each $jw$ in the process \(\ddagger\) in Figure 1}

(ii) Segment the tuples. We segmented them into 130.

(iii) For each segment, manually extract tuples for which $jw$ should be associated with one of the 12 domains and those for which $jw$ should be deemed as NODOMAIN.\footnote{We extracted five tuples for those associated with one of the domains and five for those associated with NODOMAIN, respectively.} 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
3.4 Performance of the Domain Association Method

This subsection reports the performance of the proposed method. We applied the method to fundamental words installed in the Japanese morphological analyzer JUMAN (Kurohashi et al. 1994). To be more precise, we used 26,658 words comprising commonly used nouns and verbs as fundamental words.

For our evaluation, we sampled 380 pairs of a fundamental words and domains, and measured the accuracy of our method, i.e., we counted the number of pairs for which the association was correct.

For comparison, a baseline was defined, which was the accuracy when all fundamental words were associated with nomain. This is because a pilot study we conducted showed that more than half of fundamental words tended to be associated with nomain.

As a result, the proposed method attained an accuracy of 81.3% (309/380), whereas the baseline provided an accuracy of 69.5% (264/380). This result shows that our method works very well.

3.5 Allowing Multiple Domains for Fundamental Word

This subsection shows the extended version of the proposed method, in which a fundamental word can be associated with more than one domain. In fact, some fundamental words should be

Figure 2 illustrates the process from (i) to (iv).

by each segment) and the appropriate threshold for the number of hits should be obtained.

(v) Approximate the relation between the number of hits and its threshold by a linear function using the least square method. This linear function provides the appropriate threshold for each jw.
associated with multiple domains. For example, "大学院" (graduate school) seems to be associated with both EDUCATION and SCIENCE, and "登山" (climbing) belongs to both RECREATION and SPORTS. However, the proposed method is designed to associate a fundamental word with only one domain.

The extended method associates a fundamental word with any domain that meets the following two conditions:

(i) The domain’s $A_d$ score is above the threshold described in Section 3.3.
(ii) The domain’s $A_d$ score is close to the highest $A_d$ score, as formalized as below.

$$\frac{\text{the highest } A_d - \text{the domain’s } A_d}{\text{the highest } A_d} < 0.01$$

With the extended method, an additional 814 pairs of fundamental words and domains were identified, but the accuracy dropped to 78.6% (308/392).

### 3.6 Manual Correction

Generally speaking, manual annotation of linguistic data can be arbitrary and thus should be subject to guidelines. Among the guidelines that we established, this subsection describes the criteria for which words to associate with multiple domains, how to deal with polysemous words, and which words to associate with.NODEomain.

**Words Belonging to Multiple Domains:** Our principal policy is that simpler is better, and hence we avoid associating a fundamental word with multiple domains as much as possible. Fundamental words to associate with multiple domains are restricted to those that are EQUALLY relevant to more than one domain. As mentioned in Section 3.5, “大学院” (graduate school) (EDUCATION and SCIENCE) and “登山” (climbing) (RECREATION and SPORTS) are such examples. In contrast, “ゴルフ” (golf), which some people might consider RECREATION but is more directly related to SPORTS, so we associate it with only SPORTS. Similarly, “微分” (mathematical derivation) is associated with only EDUCATION and not SCIENCE.

**Dealing with Polysemous Words:** We associated each polysemous word with multiple domains, one for each meaning of the word. For example, “ボール” can be either ball or bowl, which we associated with SPORTS and DIET, respectively.

---

10 To improve the domain dictionary using manual correction, we adopted a method of associating a fundamental word with only one domain for greater accuracy. Fundamental words that should be associated with multiple domains were then manually associated, as described in Section 3.6.
Criteria for NODOMAIN: We tried to be as conservative as possible. We associated a word with NODOMAIN if either (1) people are likely to disagree about which domain is appropriate, or (2) more than four domains could be associated with it, rather than forcing them into a domain. For example, “委員” (committee member) is so vague that it can be associated with GOVERNMENT, BUSINESS, EDUCATION, and more. We associated words like this with NODOMAIN.

4 Resulting Domain Dictionary

Table 3 shows the breakdown of the words in the resulting domain dictionary. The most common domain is NODOMAIN due to the manual annotation criteria described in the previous section. The number of words associated with multiple domains is 787 (26.2%).

We incorporated the resulting domain dictionary into the Japanese morphological analyzer JUMAN, which is available at http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN. JUMAN uses the domain dictionary when performing the morphological analysis, including tokenization, POS-tagging, and domain-tagging.

5 Blog Categorization

We categorized blog articles into one of our domains (Table 1) as a task-based evaluation of our method of constructing a domain dictionary. Our blog categorization method is quite simple; it annotates words in an article with domains and categorizes the article as the most dominant domain. The procedure is as follows (Figure 3):

<table>
<thead>
<tr>
<th>Domain</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CULTURE</td>
<td>4.7</td>
</tr>
<tr>
<td>RECREATION</td>
<td>1.0</td>
</tr>
<tr>
<td>SPORTS</td>
<td>2.5</td>
</tr>
<tr>
<td>HEALTH</td>
<td>3.4</td>
</tr>
<tr>
<td>LIVING</td>
<td>5.4</td>
</tr>
<tr>
<td>DIET</td>
<td>3.9</td>
</tr>
<tr>
<td>TRANSPORTATION</td>
<td>1.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUCATION</td>
<td>2.3</td>
</tr>
<tr>
<td>SCIENCE</td>
<td>1.4</td>
</tr>
<tr>
<td>BUSINESS</td>
<td>3.6</td>
</tr>
<tr>
<td>MEDIA</td>
<td>0.7</td>
</tr>
<tr>
<td>GOVERNMENT</td>
<td>6.2</td>
</tr>
<tr>
<td>NODOMAIN</td>
<td>63.4</td>
</tr>
</tbody>
</table>

11 These numbers are drawn from our original Japanese journal paper published in 2008. The breakdown of the domain information in the current JUMAN’s dictionary is different.
1. Extract words from an article.
2. Assign domains and IDFs to the words.
3. Sum up IDFs for each domain.
4. Categorize the article in the domain with the highest IDF.\textsuperscript{12}

We used three kinds of words in our experiment (described in Section 7):
- Fundamental words only, such as "業者" (trader)
- Fundamental and unknown words, such as "コンプライアンス" (compliance)
- Fundamental, unknown, and compound words, such as "贈収賄・容疑" (bribery-allegation)

Compound words are not in the domain dictionary like unknown words.

In extracting fundamental words only, compound words are split and fundamental words that constitute the compounds are extracted separately. For example, "贈収賄容疑" is split into "贈収賄" (bribery) and "容疑" (suspicion), and they are used for categorization independently.

\textsuperscript{12} If the domain of the highest IDF is \textit{nодомен}, the article is categorized as the second highest domain.

\textsuperscript{13} The hyphen is a word boundary.
As for \[2\], in which domains and IDFs are assigned, the IDF of word \((w)\) is calculated as follows:\footnote{We used 10,000,000,000 as the total number.}

\[
IDF(w) = \log \frac{\text{Total } \# \text{ of Japanese web pages}}{\# \text{ of hits of } w}
\]

(1)

Also, fundamental words are assigned their domains and IDFs by the domain dictionary,\footnote{We assigned each word of the domain dictionary an IDF and domain in advance.} whereas those for unknown and compound words are dynamically estimated by the method described in Section 6.

In what follows, we refer to both unknown and compound words as “unknown words.” When we need to distinguish between them, we call the former as “simplex unknown words” and the latter as “complex unknown words.”

6 Domain Estimation for Unknown Words

We dynamically estimate the domain (and IDF) of an unknown word using the web. The intuition behind this is that the web shows how unknown words are used and interpreted in the world, which provides important clues for identifying the domains with which they should be associated with. More specifically, we use Wikipedia articles and snippets in web search results in addition to the domain dictionary.

The estimation process is as follows (Figure 4):

1. Search the web for an unknown word, acquire the top 100 records, and calculate the IDF of the word.\footnote{Note that calculating the IDF of a word requires only the number of hits for it, which the search results provide.}

2. Get the Wikipedia article about the word from the search results if one exists, estimate the domain of the word using the Wikipedia-strict module (Section 6.1), and exit the process.

3. If no Wikipedia article about the word is found, then obtain any Wikipedia article among the top 30 of the search results, estimate the domain of the word using the Wikipedia-loose module (Section 6.1), and exit the process.

4. If no Wikipedia article is found among the top 30 of the search results, then remove all advertisement snippets from the search results (Section 6.2).

5. If a snippet remains in the search results, identify the best domain for the word using the Snippets module (Section 6.2) and then exit the process.
If no snippet is left but the unknown word is a compound word containing fundamental words, then identify the best domain for the word using the Components module (Section 6.3), and then exit the process.

If no snippet is left and the word is not a compound word containing fundamental words, then the process is a failure.

The next three subsections describe the Wikipedia-strict, Wikipedia-loose, Snippets, and Components modules in detail. All of them share the idea that the domain to which an unknown word should belong is the most dominant domain in a text they deal with like a Wikipedia article, snippets in search results, or text related to component words of the unknown word. Also, all of them use only fundamental words in the text to estimate domains.
6.1 Wikipedia-strict and Wikipedia-loose Modules

The two Wikipedia modules follow this procedure:

1. Extract only fundamental words from the Wikipedia article.
2. Assign domains and IDFs to the fundamental words using the domain dictionary.
3. Sum up IDFs for each domain.
4. Assign the domain with the highest IDF to the unknown word. If the domain with the highest IDF is NODOMAIN, the second highest domain is chosen for the unknown word using the condition below:

   \[
   \frac{\text{The second highest IDF}}{\text{NODOMAIN's IDF}} > 0.15
   \]  

   (2)

Figure 5 shows the procedure of the domain estimation for an unknown word using the Wikipedia-strict or Wikipedia-loose module.

Each of the two Wikipedia modules assigns a domain in about 10 seconds (from 1 to 2 or 3).\(^\text{17}\)

6.2 Snippets Module

The Snippets module takes the snippets left in the search results after advertisement websites (5) are removed as input. We remove snippets in which the keywords in Table 4 appear more than once. These words were collected from the error analysis of our preliminary experimental result. Removing advertisement snippets is indispensable because they bias estimation towards BUSINESS.

The Snippets module is the same as the two Wikipedia modules, except that it extracts fundamental words from residual snippets in the search result.

The Snippets module assigns a domain in about 6 seconds (from 1 to 5).\(^\text{18}\)

6.3 Components Module

This module is basically the same as the others except that it extracts fundamental words from the unknown word itself. For example, the domain of “金融市場” (finance market) is assigned based on the domain of “金融” (finance) and “市場” (market).

The Component module assigns a domain in about 4 seconds (from 1 to 6).\(^\text{19}\)

\(^{17}\) We used Dell PowerEdge 830 (Pentium D 3.00GHz).
\(^{18}\) Though the Wikipedia modules are executed before this module, the latter takes a shorter time. This is because it uses already acquired snippets, while the Wikipedia modules have to retrieve an article from the web.
\(^{19}\) This is the fastest because the text from which fundamental words are extracted (the unknown word itself) is rather small, and unlike the Wikipedia modules, it does not require additional web access.
7 Evaluation

We conducted blog categorization using the above-mentioned method and measured the accuracy with which the method assigned domains to blogs. Also, we measured the accuracy of the results of estimating unknown words’ domains that were obtained during the Blog categorization.
Table 4  Keywords of Advertisement Snippets

| “会社” (company) | “株式” (stock) | “商品” (goods) |
| “企業” (enterprise) | “製品” (product) | “価格” (price) |
| “ショップ” (shop) | “無料” (charge-free) | “市場” (market) |
| “サービス” (service) | “通販” (mail-order) | “事業” (project) |
| “購入” (purchase) | “発売” (release) | “店舗” (store) |
| “採用” (acceptance) | “業務” (operation) | “販売” (sales) |
| “営業” (business) | “工業” (industry) | “広告” (ads) |
| “出荷” (shipping) | “料金” (charge) | “仕事” (job) |
| “会員” (membership) | “法人” (fictitious person) |
| “当社” (our company) | “ビジネス” (business) |

7.1 Experimental Condition

Data: We categorized 600 blog articles into 12 domains. 50 articles for each domain were collected from Yahoo! Blog (blogs.yahoo.co.jp). In Yahoo! Blog, the authors of articles manually classify those articles into categories, which correspond to what we call domains. In this experiment, we selected appropriate categories for each domain and collected 50 articles from the categories, as shown in Table 5. Note that some articles (about 30%) either do not fit in the categories into which they are classified\(^{20}\) or contain only photos and no textual content. We replaced those articles with more appropriate ones in advance.

Evaluation Method for Blog Categorization: We measured the accuracy of blog categorization and the domain estimation for unknown words. In blog categorization, we extracted three kinds of words from articles: fundamental words (F only in Table 6), fundamental and simplex unknown words (F + SU), and fundamental and all unknown words (both simplex and complex, F + AU). We also measured the accuracy of not only the domain of the highest IDF but also the top N domains (Top N in Table 6). Furthermore, we evaluated the performance when we used the domain dictionary without manual correction and the performance when we used the number of words for each domain as the score instead of the IDF value.

Evaluation Method for Domain Estimation for Unknown Words: During categorization, about 12,000 unknown words were found in 600 articles, and the domain estimation for them was conducted on-the-fly. We then sampled 500 estimation results from them and measured the accuracy of domain estimation. Unknown words assigned to more than one domain were judged as correct if they were assigned to one of the correct domains. We also examined the number of times each estimation module was used and how accurate they were.

\(^{20}\) For example, an article on a leisure trip is categorized as the science category.
Table 5  Correspondence between Domains and Yahoo! Blog’s Categories

<table>
<thead>
<tr>
<th>Domain</th>
<th>Yahoo! Blog category</th>
</tr>
</thead>
<tbody>
<tr>
<td>CULTURE</td>
<td>Entertainment &gt; Movie&lt;br&gt;Entertainment &gt; Music&lt;br&gt;Entertainment &gt; Entertainer&lt;br&gt;Arts and humanities &gt; Arts&lt;br&gt;Arts and humanities &gt; Play&lt;br&gt;Living and culture &gt; Holiday, anniversary, and annual events</td>
</tr>
<tr>
<td>RECREATION</td>
<td>Entertainment &gt; Theme park&lt;br&gt;Hobby and sports &gt; Leisure&lt;br&gt;Hobby and sports &gt; Hobby</td>
</tr>
<tr>
<td>SPORTS</td>
<td>Hobby and sports &gt; Sports</td>
</tr>
<tr>
<td>HEALTH</td>
<td>Health and medicine</td>
</tr>
<tr>
<td>LIVING</td>
<td>Home and house &gt; Home</td>
</tr>
<tr>
<td>DIET</td>
<td>Living and culture &gt; Gourmet and drinks</td>
</tr>
<tr>
<td>TRANSPORTATION</td>
<td>Hobby and sports &gt; Vehicle</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>School and education &gt; School &gt; High school&lt;br&gt;School and education &gt; School &gt; Elementary school&lt;br&gt;School and education &gt; School &gt; Junior high school&lt;br&gt;School and education &gt; Education</td>
</tr>
<tr>
<td>SCIENCE</td>
<td>Computer and the internet&lt;br&gt;Science</td>
</tr>
<tr>
<td>BUSINESS</td>
<td>Business and economics</td>
</tr>
<tr>
<td>MEDIA</td>
<td>Entertainment &gt; Television &gt; Announcer</td>
</tr>
<tr>
<td>GOVERNMENT</td>
<td>Government&lt;br&gt;Living and culture &gt; Incidents and accidents &gt; Incidents&lt;br&gt;Living and culture &gt; Incidents and accidents &gt; Accidents</td>
</tr>
</tbody>
</table>

Table 6  Accuracy of Blog Categorization

<table>
<thead>
<tr>
<th>Top N</th>
<th>F only</th>
<th>F+SU</th>
<th>F+AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.89</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>2.</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>3.</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>4.</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>5.</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>

7.2 Result of Blog Categorization

Table 6 shows the accuracy of categorization. The F only column indicates that a rather simple method like the one in Section 5 works well if fundamental words are given good clues for categorization: the domain in our case. Furthermore, F + SU slightly outperformed F only,
and $F + AU$ outperformed the others. This shows that the domain estimation for unknown words moderately improves blog categorization.

Errors are mostly due to the system’s incorrect focus on topics of secondary importance. For example, in an article on a sightseeing trip, which should have been assigned the RECREATION domain, the author frequently mentioned the means of transportation. As a result, the article was wrongly categorized as TRANSPORTATION. Another example is an article that discussed software engineers working on business management systems. It should have been assigned the SCIENCE domain,\(^{21}\) but because it included numerous business keywords, it was assigned the BUSINESS domain.

Table 7 shows the results when we used the domain dictionary without manual correction. The accuracy was over 80% even without manual correction, which indicates that the method for constructing the domain dictionary provides high performance. In contrast, the fact that the accuracy was over 10% lesser than when we manually corrected the dictionary (0.82, compared to 0.94) indicates that our manual correction of the domain dictionary was conducted accurately and that our resulting domain dictionary is of high quality.

We used the sum of the IDF values for each domain as the score for blog categorization. A simpler method would be to use the number of words for each domain as the score, in which case the results are as described in Table 8. The performance was lower than that of our method, which used the sum of the IDF values. This shows that our method of using the sum of the IDF values is superior to the simpler method that uses the number of words.

### 7.3 Result of Domain Estimation for Unknown Words

The accuracy of the domain estimation for unknown words was 77.2% (386/500).

\(^{21}\) Note that the SCIENCE domain includes computer science.
Table 9 shows the frequency in use and the accuracy for each domain estimation module. The Snippets module was used most frequently and achieved a reasonably good accuracy of 78%. Though the Wikipedia-strict module showed the best performance, it was used infrequently. However, we expect that as the number of Wikipedia articles increases, the best performing module will be used more frequently, thus improving the accuracy of domain estimation as a whole.

An example of newly coined words for which domains were estimated correctly is “デイトレ,” which is an abbreviation of “デイトレード” (day-trade). It was correctly assigned the business domain by the Wikipedia-loose module. Names of famous people also appear in blog articles one after another and provide clues for blog categorization. The Wikipedia-strict module correctly judged “レオナルド・ディカプリオ” (Leonardo DiCaprio) as CULTURE, for instance.

Compound words for which domains were estimated correctly include “支持-率” (approval rate) and “運動-野” (the motor area of the brain). The former consists of “支持” (support) and “率” (rate), both of which belong to NODOMAIN, but the compound word as a whole was correctly assigned the GOVERNMENT domain by the Wikipedia-strict module. Similarly, the latter term consists of “運動” (movement) and “野” (field), both of which belong to NODOMAIN, but the Snippets module correctly assigned the compound word to the HEALTH domain.

Errors made in the experiment were mostly due to the subtle boundaries between NODOMAIN and the other particular domains. For example, names of townships and persons that are common and popular were assigned wrong domains. Both of them should be assigned NODOMAIN. However, the names of townships were often assigned the GOVERNMENT domain because most townships have their own local government websites, which mislead the Wikipedia or Snippets modules. Common and popular personal names are often sources of error because virtually any person’s name is linked to a particular domain on the web. In other words, you will always find web sites of a particular domain by searching the web for a common and popular name.
8 Related Work

8.1 Related Work on Domain Dictionary

First of all, much lesser research has been carried out on the domain relations than on is-a relations; fewer resources have been constructed and fewer proposals on construction methods have been made.

The domain information of a word is available in only a few lexical resources such as HowNet (Dong and Dong 2006) and WordNet. HowNet includes 32 domains such as economy, industry, agriculture, education. In WordNet (2.0), domain relations hold between synsets. For example, forehand, rally, and match point are associated with tennis. Some (human-oriented) dictionaries like LDOCE (Proctor 1987) describe which domain (subject) a word belongs to. However, such resources are available only for a few languages, such as English and Chinese.\(^{22}\)

Thus, towards deeper natural language understanding for languages other than English or Chinese, an efficient way for constructing a domain dictionary is required. However, most proposals on domain dictionary construction that have been made so far rely on existing resources, like LDOCE or WordNet. Guthrie, Guthrie, Wilks, and Aidinejad (1991) exploited the domain (subject) information of LDOCE to establish domain links between words. To enrich WordNet with domain information, Magnini and Cavaglià (2000) manually annotated upper synsets with domain information and then automatically extended the manual assignments to all the reachable synsets by exploiting WordNet relations. Agirre, Ansa, Martinez, and Hovy (2001) extracted from documents topically related words for each word sense in WordNet. The documents were collected from the web by querying a search engine. To construct an effective query, they used semantic information from WordNet. Chang, Huang, Ker, and Yang (2002) assigned domain tags to WordNet by exploiting WordNet relations. They defined their domain tags on the basis of a small amount of existing domain information in WordNet and also information from the Far East Dictionary. As is seen, these methods are not applicable to languages like Japanese, for which lexical resources corresponding to LDOCE or WordNet are not available.\(^{23}\)

To construct a domain dictionary without relying on existing resources like WordNet, you might use keyword extraction techniques that have been developed in the context of information retrieval (IR) or term recognition (Frantzi, Ananiadou, and Tsujii 1998; Hisamitsu and Tsujii 2003; Nakagawa, Mori, and Yumoto 2003, and so on). However, as we mentioned in Section 2,

---

22 Some Japanese dictionaries occasionally describe domain information, but they cover only a few words.
23 The Japanese WordNet http://nlpwww.nict.go.jp/wn-ja/index.en.html has been available only recently and was not available when the original journal paper of this article was published.
keyword extraction techniques would not work well, because document collections are hard to obtain for the common domains assumed in this study.

Ultimately, our domain dictionary provides two innovations: the first fully-available Japanese domain dictionary and a method for constructing a domain dictionary that requires neither highly structured existing lexical resources like WordNet nor document collections from which to extract keywords.

Our domain dictionary is not tailored to any particular NLP application. But among other things, domain information has been used for document filtering (Liddy and Paik 1993), word-sense disambiguation (Rigau et al. 1997), and machine translation (Yoshimoto et al. 1997; Lange and Yang 1999).

### 8.2 Related Work on Text Categorization

Text-categorization methods that have been developed so far are mostly based on machine learning, such as $k$-nearest neighbor (Yang 1999), decision tree (Lewis and Ringuette 1994), naive Bayes (Lewis 1998), decision list (Li and Yamanishi 1999), support vector machines (Joachims 1999), boosting (Schapire and Singer 2000).

Methods based on machine learning require huge quantities of text to be used as training data, which constitutes the bottleneck for those methods. Though there has been a growing interest in a technique using a small number of annotated text and a large number of unannotated text for machine learning (Abney 2007), methods based on this technique are still in the early phase of development.

In contrast, our method, which achieved an accuracy of 94%, requires no training data. All you need is a manageable number of fundamental words with good clues for categorization (into domains). The construction of our domain dictionary also requires no training data. All you need is access to the web, and manual correction (Section 3.6) is not labor-intensive.

Also note that our categorization method is NOT tailored to the 12 domains in Table 1, though we used them in this study. If you want to categorize texts based on your own domains, you must only construct your own domain dictionary using the method described in Section 3, which is neither time-consuming nor painstaking. Moreover, it is unlikely that the domains into which texts are categorized will need to be changed frequently in practice. Thus, you will likely have to construct your own domain dictionary only once.

Another important feature of our method is the utilization of the on-the-fly estimation of unknown words to domains, which achieved 77% accuracy. This feature is useful for categorizing texts like blog articles, which are updated on a daily basis and are filled with newly coined words.
or neologisms. In contrast, machine learning approaches need to collect huge quantities of texts at short intervals in order to update classifiers for frequently updated texts and the neologisms in them.

Consequently, as a result of our blog categorization research, we have developed a simple but high-performance text categorization method that i) uses no machine learning, and thus requires no text collection; and ii) can process unknown words dynamically.

9 Conclusion

Towards deeper natural language understanding, we constructed a domain dictionary, where about 30,000 fundamental words are grouped by domain. There are two issues in domain dictionary construction; choosing domains and designing a method by which to associate words with domains. We avoided being too involved in the former issue by adopting 12 domains that were based on web directories like the Open Directory Project. As for the latter issue, based on the work by Sasaki et al. (Sasaki et al. 2006), we developed a semi-automatic method that does not require document collections or lexical resources like WordNet. As a result, 81.3% of the target fundamental words were associated with the correct domains. We further improved the domain dictionary by manually correcting the association results.

We also presented a blog categorization method that exploits our domain dictionary and the dynamic domain estimation for unknown words. Our method categorized blogs with 94% accuracy, and estimated domains for unknown words with 77% accuracy.

The contribution of this study is as follows. First, we constructed the world’s first publicly-available Japanese domain dictionary. Second, we developed the domain dictionary construction method for fundamental words that requires neither training data nor highly-structured language resources like WordNet. Finally, we developed a blog categorization method that requires no training data and flexibly assigns unknown words to domains.

In our future work, we plan to apply our domain dictionary to word-sense disambiguation and translation word selection, among other things. Tanaka et al. used the domain information of words that appear in the same sentence as a target word of word-sense disambiguation (Tanaka et al. 2007). We will use domain information from broader contexts, such as the domain information of words that appear in the neighboring sentences as well. Furthermore, we will exploit the domain information of unknown words for the task.
Acknowledgement

We would like to thank all the members of the collaborative research group of Kyoto University and NTT Communication Science Laboratories for their stimulating discussion.

Reference


**Chikara Hashimoto**: Chikara Hashimoto received Ph.D. in Linguistic Sciences from Kobe Shoin Graduate School in 2004 and Ph.D in Informatics from Kyoto University in 2011. He is currently a research manager of Information Analysis Laboratory at National Institute of Information and Communications Technology (NICT). He is interested in natural language processing. He received the Young Scientists' Prize (The Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology) in 2014, IPSJ Best Paper Award 2011, and Best Paper Award 2008 given by the Association for Natural Language Processing.

**Sadao Kurohashi**: Sadao Kurohashi received the B.S., M.S., and PhD in Electrical Engineering from Kyoto University in 1989, 1991 and 1994, respectively. He has been a visiting researcher of IRCS, University of Pennsylvania in 1994. He is currently a professor of the Graduate School of Informatics at Kyoto University. His research interests include natural language processing, knowledge acquisition/representation, and information retrieval. He received the 10th and 20th anniversary best paper awards from Journal of Natural Language Processing in 2004 and 2014 respectively, 2009 Funai IT promotion award, and 2009 IBM faculty award.

(Received December 26, 2007)
(Revised April 11, 2008)
(Accepted May 11, 2008)