Identification of Domain-Specific Senses in a Machine-Readable Dictionary

Fumiyo Fukumoto
Interdisciplinary Graduate School of Medicine and Engineering,
Univ. of Yamanashi
fukumoto@yamanashi.ac.jp

Yoshimi Suzuki
Interdisciplinary Graduate School of Medicine and Engineering,
Univ. of Yamanashi
ysuzuki@yamanashi.ac.jp

Abstract

This paper focuses on domain-specific senses and presents a method for assigning category/domain label to each sense of words in a dictionary. The method first identifies each sense of a word in the dictionary to its corresponding category. We used a text classification technique to select appropriate senses for each domain. Then, senses were scored by computing the rank scores. We used Markov Random Walk (MRW) model. The method was tested on English and Japanese resources, WordNet 3.0 and EDR Japanese dictionary. For evaluation of the method, we compared English results with the Subject Field Codes (SFC) resources. We also compared each English and Japanese results to the first sense heuristics in the WSD task. These results suggest that identification of domain-specific senses (IDSS) may actually be of benefit.

1 Introduction

Domain-specific sense of a word is crucial information for many NLP tasks and their applications, such as Word Sense Disambiguation (WSD) and Information Retrieval (IR). For example, in the WSD task, McCarthy et al. presented a method to find predominant noun senses automatically using a thesaurus acquired from raw textual corpora and the WordNet similarity package (McCarthy et al., 2004; McCarthy et al., 2007). They used parsed data to find words with a similar distribution to the target word. Unlike Buitelaar et al. approach (Buitelaar and Sacaleanu, 2001), they evaluated their method using publically available resources, namely SemCor (Miller et al., 1998) and the SENSEVAL-2 English all-words task. The major motivation for their work was similar to ours, i.e., to try to capture changes in ranking of senses for documents from different domains.

Domain adaptation is also an approach for focusing on domain-specific senses and used in the WSD task (Chand and Ng, 2007; Zhong et al., 2008; Agirre and Lacalle, 2009). Chan et al. proposed a supervised domain adaptation on a manually selected subset of 21 nouns from the DSO corpus having examples from the Brown corpus and Wall Street Journal corpus. They used active learning, count-merging, and predominant sense estimation in order to save target annotation effort. They showed that for the set of nouns which have different predominant senses between the training and target domains, the annotation effort was reduced up to 29%. Agirre et al. presented a method of supervised domain adaptation (Agirre and Lacalle, 2009). They made use of unlabeled data with SVM (Vapnik, 1995), a combination of kernels and SVM, and showed that domain adaptation is an important technique for WSD systems. The major motivation for domain adaptation is that the sense distribution depends on the domain in which a word is used. Most of them adapted textual corpus which is used for training on WSD.

In the context of dictionary-based approach, the first sense heuristic applied to WordNet is often used as a baseline for supervised WSD systems (Cotton et al., 1998), as the senses in WordNet are ordered according to the frequency data in the manually tagged resource SemCor (Miller et al., 1998). The usual
drawback in the first sense heuristic applied to the WordNet is the small size of the SemCor corpus. Therefore, senses that do not occur in SemCor are often ordered arbitrarily. More seriously, the decision is not based on the domain but on the frequency of SemCor data. Magnini et al. presented a lexical resource where WordNet 2.0 synsets were annotated with Subject Field Codes (SFC) by a procedure that exploits the WordNet structure (Magnini and Cavaglia, 2000; Bentivogli et al., 2004). The results showed that 96% of the WordNet synsets of the noun hierarchy could have been annotated using 115 different SFC, while identification of the domain labels for word senses was required a considerable amount of hand-labeling.

In this paper, we focus on domain-specific senses and propose a method for assigning category/domain label to each sense of words in a dictionary. Our approach is automated, and requires only documents assigned to domains/categories, such as Reuters corpus, and a dictionary with gloss text, such as WordNet. Therefore, it can be applied easily to a new domain, sense inventory or different languages, given sufficient documents.

2 Identification of Domain-Specific Senses

Our approach, IDSS consists of two steps: selection of senses and computation of rank scores.

2.1 Selection of senses

The first step to find domain-specific senses is to select appropriate senses for each domain. We used a corpus where each document is classified into domains. The selection is done by using a text classification technique. We divided documents into two sets, i.e., training and test sets. The training set is used to train SVM classifiers, and the test set is to test SVM classifiers. For each domain, we collected noun words. Let \( D \) be a domain set, and \( S \) be a set of senses that the word \( w \in W \) has. Here, \( W \) is a set of noun words. The senses are obtained as follows:

1. For each sense \( s \in S \), and for each \( d \in D \), we applied word replacement, i.e., we replaced \( w \) in the training documents assigning to the domain \( d \) with its gloss text in a dictionary.

2. All the training and test documents are tagged by a part-of-speech tagger, and represented as term vectors with frequency.

3. The SVM was applied to the two types of training documents, i.e., with and without word replacement, and classifiers for each category are generated.

4. SVM classifiers are applied to the test data. If the classification accuracy of the domain \( d \) is equal or higher than that without word replacement, the sense \( s \) of a word \( w \) is judged to be a candidate sense in the domain \( d \).

The procedure is applied to all \( w \in W \).

2.2 Computation of rank scores

We note that text classification accuracy used in selection of senses depends on the number of words consisting gloss in a dictionary. However, it is not so large. As a result, many of the classification accuracy with word replacement were equal to those without word replacement\(^1\). Then in the second procedure, we scored senses by using MRW model.

Given a set of senses \( S_d \) in the domain \( d \), \( G_d = (S_d, E) \) is a graph reflecting the relationships between senses in the set. Each sense \( s_i \) in \( S_d \) is a gloss text assigned from a dictionary. \( E \) is a set of edges, which is a subset of \( S_d \times S_d \). Each edge \( e_{ij} \) in \( E \) is associated with an affinity weight \( f(i \rightarrow j) \) between senses \( s_i \) and \( s_j \) \( (i \neq j) \). The weight is computed using the standard cosine measure between two senses. The transition probability from \( s_i \) to \( s_j \) is then defined by normalizing the corresponding affinity weight \( p(i \rightarrow j) = \frac{f(i \rightarrow j)}{\sum_{k=1}^{S_d} f(i \rightarrow k)} \), if \( \Sigma f \neq 0 \), otherwise, 0.

We used the row-normalized matrix \( U_{ij} = (U_{ij})_{|S_d| \times |S_d|} \) to describe \( G \) with each entry corresponding to the transition probability, where \( U_{ij} = p(i \rightarrow j) \). To make \( U \) a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to \( \frac{1}{|S_d|} \). The matrix form of the saliency score \( \text{Score}(s_i) \) can be formulated in a recursive form as in the MRW model: \( \hat{\lambda} = \mu U^T \hat{\lambda} + \frac{1-\mu}{|S_d|} \vec{c} \), where \( \hat{\lambda} = [\text{Score}(s_i)]_{|S_d| \times 1} \) is a column vector with all elements equal to 1. \( \mu \) is a

\(^1\)In the experiment, the classification accuracy of more than 50% of words has not changed.
damping factor. We set \( \mu \) to 0.85, as in the PageRank (Brin and Page, 1998). The final transition matrix is given by the formula (1), and each score of the sense in a specific domain is obtained by the principal eigenvector of the new transition matrix \( M \).

\[
M = \mu U^T + \frac{(1-\mu)}{|S_d|} \overline{e} \overline{e}^T
\]  

We applied the algorithm for each domain. We note that the matrix \( M \) is a high-dimensional space. Therefore, we used a ScaLAPACK, a library of high-performance linear algebra routines for distributed memory MIMD parallel computing (Netlib, 2007). We selected the topmost \( K \% \) senses according to rank score for each domain and make a sense-domain list. For each word \( w \) in a document, find the sense \( s \) that has the highest score within the list. If a domain with the highest score of the sense \( s \) and a domain in a document appearing \( w \) match, \( s \) is regarded as a domain-specific sense of the word \( w \).

3 Experiments

3.1 WordNet 3.0

We assigned Reuters categories to each sense of words in WordNet 3.0. The Reuters documents are organized into 126 categories (Rose et al., 2002). We selected 20 categories consisting a variety of genres. We used one month of documents, from 20th Aug to 19th Sept 1996 to train the SVM model. Similarly, we classified the following one month of documents into these 20 categories. All documents were tagged by Tree Tagger (Schmid, 1995).

Table 1 shows 20 categories, the number of training and test documents, and F-score (Baseline) by SVM. For each category, we collected noun words with more than five frequencies from one-year Reuters corpus. We randomly divided these into two: 10% for training and the remaining 90% for test data. The training data is used to estimate \( K \) according to rank score, and test data is used to test the method using the estimated value \( K \). We manually evaluated a sense-domain list. As a result, we set \( K \) to 50%. Table 2 shows the result using the test data, i.e., the total number of words and senses, and the number of selected senses (Select_S) that the classification accuracy of each domain was equal or higher than the result without word replacement. We used these senses as an input of MRW.

There are no existing sense-tagged data for these 20 categories that could be used for evaluation. Therefore, we selected a limited number of words and evaluated these words qualitatively. To do this, we used SFC resources (Magnini and Cavaglia, 2000), which annotate WordNet 2.0 synsets with domain labels. We manually corresponded Reuters and SFC categories. Table 3 shows the results of 12 Reuters categories that could be corresponded to SFC labels. In Table 3, “Reuters” shows categories, and “IDSS” shows the number of senses assigned by our approach. “SFC” refers to the number of senses appearing in the SFC resource. “S & R” denotes the number of senses appearing in both SFC and Reuters corpus. “Prec” is a ratio of correct assignments by “IDSS” divided by the total number of “IDSS” assignments. We manually evaluated senses not appearing in SFC resource. We note that the corpus used in our approach is different from SFC. Therefore, recall denotes a ratio of the number of senses matched in our approach and SFC divided by the total number of senses appearing in both SFC and Reuters.

As shown in Table 3, the best performance was “weather” and recall was 0.986, while the result for “war” was only 0.149. Examining the result of text classification by word replacement, the former was 0.07 F-score improvement by word replacement, while that of the latter was only 0.02. One reason is related to the length of the gloss in WordNet: the average number of words consisting the gloss assigned to “weather” was 8.62, while that for “war” was 5.75. IDSS depends on the size of gloss text in WordNet. Efficacy can be improved if we can assign gloss sentences to WordNet based on corpus statistics. This is a rich space for further exploration.

In the WSD task, a first sense heuristic is often applied because of its powerful and needless of expensive hand-annotated data sets. We thus compared the results obtained by our method to those obtained by the first sense heuristic. For each of the 12 categories, we randomly picked up 10 words from the senses assigned by our approach. For each word, we
Table 1: Classification performance (Baseline)

| Cat           | Train | Test | F-score | Cat           | Train | Test | F-score |
|---------------|-------|------|---------|---------------|-------|------|---------|
| Legal/judicial| 897   | 808  | .499    | Funding       | 3,245 | 3,588| .709    |
| Production    | 2,179 | 2,267| .463    | Research      | 204   | 180  | .345    |
| Advertising   | 113   | 170  | .477    | Management    | 923   | 812  | .753    |
| Employment    | 1,224 | 1,305| .703    | Disasters     | 757   | 522  | .726    |
| Arts/entertainments | 526  | 295  | .536    | Environment   | 532   | 420  | .476    |
| Fashion       | 13    | 50   | .333    | Health        | 524   | 447  | .513    |
| Labour issues | 1,278 | 1,343| .741    | Religion      | 257   | 251  | .665    |
| Science       | 158   | 128  | .528    | Sports        | 2,311 | 2,682| .967    |
| Travel        | 47    | 64   | .517    | War           | 3,126 | 2,674| .678    |
| Elections     | 1,107 | 1,208| .689    | Weather       | 409   | 247  | .688    |

Table 2: The # of candidate senses (WordNet)

| Cat           | Words  | Senses | Senses | Senses | Cat           | Words  | Senses | Senses | Senses |
|---------------|--------|--------|--------|--------|---------------|--------|--------|--------|--------|
| Legal/judicial| 10,920 | 62,008 | 25,891 |        | Funding       | 11,383 | 28,299 | 26,209 |        |
| Production    | 13,967 | 31,398 | 30,541 |        | Research      | 7,047  | 19,423 | 18,600 |        |
| Advertising   | 7,960  | 23,154 | 20,414 |        | Management    | 9,386  | 24,374 | 22,961 |        |
| Employment    | 11,056 | 28,413 | 25,915 |        | Disasters     | 10,176 | 28,420 | 24,266 |        |
| Arts/entertainments | 5,366 | 16,216 | 15,032 |        | Environment   | 10,176 | 28,420 | 24,266 |        |
| Fashion       | 4,039  | 15,001 | 12,319 |        | Health        | 8,547  | 21,845 | 21,468 |        |
| Labour issues | 11,043 | 28,410 | 25,845 |        | Religion      | 8,643  | 23,121 | 21,861 |        |
| Science       | 8,643  | 29,303 | 28,410 |        | Sports        | 12,946 | 31,209 | 29,049 |        |
| Travel        | 5,366  | 16,216 | 15,032 |        | War           | 13,864 | 32,476 | 30,476 |        |
| Elections     | 11,602 | 29,310 | 26,978 |        | Weather       | 6,059  | 18,239 | 16,402 |        |

Table 3: The results against SFC resource

| Cat           | Reuters | IDSS | SFC | S&R | Rec | Prec |
|---------------|---------|------|-----|-----|-----|------|
| Legal/judicial| 25,715  | 2,254| 3,187| 809 | .904| .893 |
| Funding       | 935     | 3,978| 3,482| 576 | .791| .812 |
| Arts          | 3,725   | 7    | .875| .763|     |     |
| Environment   | 12,108  | 2,112| 241 | .892| .793|     |
| Fashion       | 935     | 1,394| 338 | .800| .820|     |
| Health        | 10,347  | 79   | 79  | .329| .302|     |
| Science       | 21,635  | 941  | 62,513| 2,736| .810| .783|
| Religion      | 1,766   | 506  | 213 | .359| .365|     |
| Travel        | 14,925  | 506  | 213 | .359| .365|     |
| War           | 2,999   | 1,668| 301 | .149| .102|     |
| Weather       | 16,244  | 253  | 72  | .986| .970|     |

Table 3: The results against SFC resource

4 also shows that overall performance obtained by our method was better than that with the first sense heuristic in all categories.

3.2 EDR dictionary

We assigned categories from Japanese Mainichi newspapers to each sense of words in EDR Japanese dictionary. The Mainichi documents are organized into 15 categories. We selected 4 categories, each of which has sufficient number of documents. All documents were tagged by a morphological analyzer Chasen (Matsumoto et al., 2000), and nouns are extracted. We used 10,000 documents for each category from 1991 to 2000 year to train SVM model. We classified other 600 documents from the same period into one of these four categories. Table 5 shows categories and F-score (Baseline) by SVM. We used the same ratio used in English data to estimate $K$. As a result, we set $K$ to 30%. Table 6 shows the result of IDSS. “Prec” refers to the precision of IDSS, i.e., we randomly selected 300 senses
Table 4: IDSS against the first sense heuristic (WordNet)

| Cat            | First sense | IDSS | Sense |
|----------------|-------------|------|-------|
|                | Correct     | Wrong| Prec  |
|                | Correct     | Wrong| Prec  |
| Legal/judicial | 69          | 31   | .69   |
| Funding        | 60          | 40   | .60   |
| Arts/entertainments | 62          | 38   | .62   |
| Environment    | 72          | 28   | .72   |
| Fashion        | 74          | 26   | .74   |
| Sports         | 72          | 28   | .72   |
| Health         | 68          | 32   | .68   |
| Science        | 69          | 31   | .69   |
| Religion       | 54          | 46   | .54   |
| Travel         | 75          | 25   | .75   |
| War            | 53          | 47   | .53   |
| Weather        | 60          | 40   | .60   |
| Average        | 4.95        | 35.1 | 0.648 |

Table 5: Text classification performance (Baseline)

| Cat            | Words | Senses | Senses | Prec |
|----------------|-------|--------|--------|------|
| International  | 3,607 | 11,292 | 10,647 | .642 |
| Economy       | 3,180 | 9,921  | 9,537  | .571 |
| Science       | 4,759 | 17,061 | 13,711 | .673 |
| Sport         | 3,724 | 12,568 | 11,074 | .681 |
| Average       | 3,818 | 12,711 | 11,242 | .642 |

Table 7: IDSS against the first sense heuristic (EDR)

| Cat            | Sense | IDSS | First sense |
|----------------|-------|------|-------------|
|                | Correct | Wrong | Prec | Correct | Wrong | Prec |
| International  | 2.873   | .630  | .587 |
| Economy       | 2.793   | .677  | .637 |
| Science       | 4.223   | .723  | .610 |
| Sports        | 2.873   | .620  | .477 |
| Average       | 3.191   | .662  | .593 |

4 Conclusion

We presented a method for assigning categories to each sense of words in a machine-readable dictionary. For evaluation of the method using WordNet 3.0, the average precision was 0.661, and recall against the SFC was 0.686. Moreover, the result of WSD obtained by our method outperformed against the first sense heuristic in both English and Japanese. Future work will include: (i) applying the method to other part-of-speech words, (ii) comparing the method with existing other automated method, and (iii) extending the method to find domain-specific senses with unknown words.
References

E. Agirre and O. L. Lacalle. 2009. Supervised domain adaption for wsd. In Proc. of the 12th Conference of the European Chapter of the ACL, pages 42–50.

L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. 2004. Revising the WORDNET DOMAINS Hierarchy: Semantics, Coverage and Balancing. In In Proc. of COL-ING 2004 Workshop on Multilingual Linguistic Resources, pages 101–108.

S. Brin and L. Page. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. In Computer Networks and ISDN Systems, volume 30, pages 1–7.

P. Buitelaar and B. Sacaleanu. 2001. Ranking and Selecting Synsets by Domain Relevance. In Proc. of WordNet and Other Lexical Resources: Applications, Extensions and Customization, pages 119–124.

Y. S. Chand and H. T. Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics, pages 49–56.

S. Cotton, P. Edmonds, A. Kilgarriff, and M. Palmer. 1998. SENSEVAL-2, http://www.sle.sharp.co.uk/senseval2/.

B. Magnini and G. Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In In Proc. of LREC-2000.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, Y. Matsuda, K. Takaoka, and M. Asahara. 2000. Japanese Morphological Analysis System ChaSen Version 2.2.1. In NAIST Technical Report NAIST.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding Predominant Senses in Untagged Text. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics, pages 280–287.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. Computational Linguistics, 33(4):553–590.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1998. A Semantic Concordance. In Proc. of the ARPA Workshop on Human Language Technology, pages 303–308.

Netlib. 2007. http://www.netlib.org/scalapack/index.html. In Netlib Repository at UTK and ORNL.

T. G. Rose, M. Stevenson, and M. Whitehead. 2002. The Reuters Corpus Volume 1 - from yesterday’s news to tomorrow’s language resources. In Proc. of Third International Conference on Language Resources and Evaluation.

H. Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In Proc. of the EACL SIGDAT Workshop.

V. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer.

Z. Zhong, H. T. Ng, and Y. S. Chan. 2008. Word sense disambiguation using ontonotes: An empirical study. In Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 1002–1010.