IDENTIFYING DOMAIN-SPECIFIC SENSES AND ITS APPLICATION TO TEXT CLASSIFICATION

Fumiyo Fukumoto and Yoshimi Suzuki
Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Japan

Keywords: Domain-specific senses, WordNet, Text classification, Markov random walk model.

Abstract: This paper focuses on domain-specific senses and presents a method for identifying predominant sense depending on each domain. The method consists of two steps: selecting senses by text classification and scoring senses by link analysis. Sense selection is to identify each sense of a word to the corresponding domain. We used a text classification technique. Senses were scored by computing the rank scores using the Markov Random Walk (MRW) model. The method was tested on WordNet 3.0 and the Reuters corpus. For evaluation of the method, we compared the results with the Subject Field Codes resources, which annotate WordNet 2.0 synsets with domain labels. Moreover, we applied the results to text classification. The results demonstrated the effectiveness of the method.

1 INTRODUCTION

Domain-specific sense of a word is crucial information for many NLP tasks, such as Word Sense Disambiguation (WSD) and Information Retrieval. For example, the first sense heuristic applied to WordNet is often used as a baseline for supervised WSD systems, as the senses in WordNet are ordered according to the frequency data in the manually tagged resource SemCor (Miller et al., 1993). Cotton *et al.* reported that the heuristic outperformed many of the WSD systems that use surrounding context as the heuristics (Cotton et al., 1998). The usual drawback in the first sense heuristic applied to WordNet is the small size of SemCor corpus, consisting of 250,000 words. Therefore, senses that do not occur in SemCor are often ordered arbitrarily. More seriously, the decision of the first sense is not based on the domain but on the frequency of SemCor data. Consider the noun word, “court”. There are eleven noun senses of “court” in WordNet. The first sense of “court” is “an assembly (including one or more judges) to conduct judicial business”, and it is often used in the “judge/law” domain rather than the “sports” domain. On the other hand, the fourth sense of “court”, *i.e.*, “a specially marked horizontal area within which a game is played,” is more likely to be used in the “sports” domain.

In this paper, we focus on domain-specific senses and propose a method for identifying the predominant sense in WordNet depending on each domain/category defined in the Reuters corpus (Rose et al., 2002). To assign categories of the Reuters corpus to each sense of the word in WordNet, we first selected each sense of a word to the corresponding domain. We used a text classification technique. For each sense *s* of a word *w*, we replace *w* in the training stories assigning to the domain *d* with its gloss text in WordNet. (hereafter, referred to as word replacement). If the classification accuracy of the domain *d* is equal or higher than that without word replacement, the sense *s* is regarded to be a domain-specific sense. However, the sense selection is not enough for identifying domain-specific senses (IDSS). Because the number of words consisting gloss in WordNet is not so large. As a result, the classification accuracy with word replacement was equal to that without word replacement[1]. Then we scored senses by computing the rank scores. We used the Markov Random Walk (MRW) model (Bremaud, 1999). For each category, the senses with large rank scores were chosen as the domain-specific senses.

[1] In the experiment, the classification accuracy of more than 50% of words has not changed.

Fukumoto F. and Suzuki Y..
IDENTIFYING DOMAIN-SPECIFIC SENSES AND ITS APPLICATION TO TEXT CLASSIFICATION.
DOI: 10.5220/0003094102630268
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 263-268
ISBN: 978-989-8425-29-4

2 SYSTEM DESIGN

2.1 Selecting Senses

The first step to IDSS is to select appropriate senses for each domain. We used the 1996 Reuters corpus and WordNet 3.0 thesaurus. The documents are organized into 126 categories (domains). We used the documents from 20 August to 19 September 1996 to train the SVM model, and documents for the following one month to test the SVM model. For each category, we collected noun words from the Reuters one-year corpus (20 August 1996 to 19 August 1997). Let D be a domain set, and S be a set of the 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 apply word replacement, *i.e.*, we replace w in the training documents assigning to the domain d with its gloss text in WordNet. For example, the first sense of the word “court” is “tribunal” or “judicature” in WordNet, and its gloss is “an assembly (including one or more judges) to conduct judicial business.” We replaced “court” in the training documents with the gloss.
2. All the documents of training and test data are tagged by a part-of-speech tagger, stop words are removed, and represented as term vectors with frequency.
3. The SVM is applied to the two types of the training documents, *i.e.*, with and without word replacement, and classifiers for each domain 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 the word w is judged to be the candidate sense in the domain d.

The procedure is applied to all $w \in W$.

2.2 Scoring Senses by Link Analysis

The next procedure for IDSS is to score each sense for each domain. We used the MRW model, which is a ranking algorithm that has been successfully used in Web-link analysis, and more recently in text processing applications. This approach decides the importance of a vertex within a graph based on global information drawn recursively from the entire graph (Bremaud, 1999). The essential idea is that of “voting” between the vertices. A link between two vertices is considered a vote cast from one vertex to the other. The score associated with a vertex is determined by the votes that are cast for it, and the score of the vertices casting these votes. We applied the algorithm to detect the domain-specific sense of words.

Given a set of senses S_d in the domain d, G_d = (V, E) is a graph reflecting the relationships between senses in the set. V is the set of vertices, and each vertex v_i in V is the gloss text assigned from WordNet. E is a set of edges, which is a subset of $V \times V$. Each edge e_{ij} in E is associated with an affinity weight $f(i \rightarrow j)$ between senses v_i and v_j ($i \neq j$). The weight is computed using the standard cosine measure between the two senses.

$$f(i \rightarrow j) \quad = \quad \frac{\vec{v}_i \cdot \vec{v}_j}{\mid \vec{v}_i \mid \times \mid \vec{v}_j \mid}. \qquad (1)$$

where $\vec{v}_i$ and $\vec{v}_j$ are the corresponding term vectors of v_i and v_j. Two vertices are connected if their affinity weight is larger than 0 and we let $f(i \rightarrow i)$= 0 to avoid self transition. The transition probability from v_i to v_j is then defined as follows:

$$p(i \rightarrow j) \; = \; \begin{cases} \dfrac{f(i \rightarrow j)}{\sum\limits_{k=1}^{|V|} f(i \rightarrow k)}, & \text{if } \Sigma f \; \neq \; 0 \\ 0 \,, & \text{otherwise.} \end{cases} \qquad (2)$$

We used the row-normalized matrix U_{ij} = $(U_{ij})_{|V| \times |V|}$ 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 $\frac{1}{|V|}$. The matrix form of the saliency score $Score(v_i)$ can be formulated in a recursive form as in the MRW model.

$$\vec{\lambda} \quad = \quad \mu U^T \vec{\lambda} + \frac{(1-\mu)}{\mid V \mid} \vec{e}. \qquad (3)$$

where $\vec{\lambda}$ = $[Score(v_i)]_{|V| \times 1}$ is the vector of saliency scores for the senses. $\vec{e}$ is a column vector with all elements equal to 1. μ is the damping factor. We set μ to 0.85, as in the PageRank (Brin and Page, 1998). The final transition matrix is given by formula (4), and each score of the sense in a specific domain is obtained by the principal eigenvector of the matrix M.

$$M \quad = \quad \mu U^T + \frac{(1-\mu)}{\mid V \mid} \vec{e}\vec{e}^T. \qquad (4)$$

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), which includes routines for solving systems of linear equations, least squares, eigenvalue problems. For implementation, we used a supercomputer, SPARC Enterprise M9000, 64CPU, 1TB memory. We selected the topmost *K*% words (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 appeared in the word *w* match, *s* is regarded as a domain-specific sense of the word *w*.

3 EXPERIMENTS

We evaluated our method using the 1996 Reuters corpus and WordNet 3.0 thesaurus. Moreover, we applied the results of IDSS to text classification to examine how well the automatically acquired senses contribute to classification accuracy.

3.1 Assigning Domain-specific Senses

The Reuters documents are organized into 126 categories. Many of these are related to economy, but there are also several other categories, such as "sports" and "fashion." We selected 20 of the 126 categories including categories other than economy, and used one month of documents, from 20 August to 19 September 1996 to train the SVM model. Similarly, we classified the following one month of documents from 20 September to 19 October into these 20 categories. All documents were tagged by Tree Tagger (Schmid, 1995), and stop words were removed.

For each domain, we collected the topmost 500 noun words. We randomly divided these nouns into two: training and test data. The training data is used to estimate *K*% words (senses) 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%. We used WordNet 3.0 to assign senses. Table 1 shows the result using the test data, *i.e.*, the total number of 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 the MRW model.

Table 1: The # of selected senses.

Cat	Total	Select_S
legal/judicial	62,008	25,891
funding	28,299	26,209
production	31,398	30,541
research	19,423	18,600
advertising	23,154	20,414
management	24,374	22,961
arts/entertainments	29,303	28,410
environment	26,226	25,413
fashion	15,001	12,319
health	25,065	24,630
labour issues	28,410	25,845
religion	21,845	21,468
science	23,121	21,861
sports	31,209	29,049
travel	16,216	15,032
war	32,476	30,476
elections	29,310	26,978
weather	18,239	16,402

Table 2: The correspondence of Reuters and SFC categories.

Reuters	SFC
legal/judicial	law
funding	economy
arts/entertainments	drawing
environment	environment
fashion	fashion
sports	sports
health	medicine, body_care
science	applied_science, pure_science
religion	religion
travel	tourism
war	military
weather	meteorology

There are no existing sense-tagged data for these 20 domains that could be used for evaluation. Therefore, we selected a limited number of words and evaluated these words qualitatively. To do this, we used the Subject Field Codes (SFC) resource (Magnini and Cavaglia, 2000), which annotates WordNet 2.0 synsets with domain labels. The SFC consists of 115,424 words assigning 168 domain labels with hierarchy. It contains some Reuters categories, such as "sports" and "fashion", while other Reuters categories and SFC labels do not match completely. Table 2 shows the Reuters categories that correspond to SFC labels[2]. We tested these 12 categories, and the results are shown in Table 3.

In Table 3, "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 the Reuters corpus. "Prec" is the ratio of correct assignments by IDSS divided by the total number of the

[2] We used these labels and its children labels in a hierarchy.

Table 3: The results against SFC resource.

Cat	IDSS	SFC	S&R	Rec	Prec	Cat	IDSS	SFC	S&R	Rec	Prec
legal/judicial	25,715	3,187	809	.904	.893	funding	2,254	2,944	747	.632	.650
arts/entertainments	3,978	3,482	576	.791	.812	environment	3,725	56	7	.857	.763
fashion	12,108	2,112	241	.892	.793	sports	935	1,394	338	.800	.820
health	10,347	79	79	.329	.302	science	21,635	62,513	2,736	.810	.783
religion	1,766	3,408	213	.359	.365	travel	14,925	506	86	.662	.673
war	2,999	1,668	301	.149	.102	weather	16,244	253	72	.986	.970
						Average	9,719	6,800	517	.686	.661

IDSS assignments[3]. We note that the corpus used in our approach is different from SFC, *i.e.*, our approach assigns domain-specific senses to the words collected from the Reuters corpus, while SFC assigns domain labels to the words appearing in WordNet. Therefore, recall denotes the 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 later 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 to WordNet based on corpus statistics. This is a rich space for further exploration.

It is interesting to note that some senses of words that were obtained correctly by our approach did not appear in the SFC resource because of the difference in WordNet version, *i.e.*, we used WordNet 3.0, while SFC was based on WordNet 2.0. Table 4 illustrates some examples obtained by our approach but that did not appear in the SFC. Because of space limitations, Table 4 gives one example for each domain. These observations clearly support the usefulness of our automated method.

3.2 Application to Text Classification

We evaluated a limited number of words by using SFC resource, *i.e.*, 6,205 senses from 12 domains. Therefore, we applied all the results of domain-specific senses to text classification to examine how the results obtained by our approach affect text classification performance. We used the documents collected from 20 October to 19 November as training data to train SVM. We used the following one month of documents from 20 November to 19 December as test data for text classification. Similar to the sense selection procedure, we used 20 categories. All documents were tagged by Tree Tagger (Schmid, 1995), and stop words were removed. All training documents are represented as term vectors with frequency where each term is replaced with the domain-specific sense with ranked scores obtained by using the training data from 20 August to 19 September and test data from 20 September to 19 October.

The classification using SVM is as follows. For the target category, we replace each word in the test document with its gloss. SVM classifiers are applied to the test document. If the category assigned to the test document by SVM classifier and the target category match, the test document is judged to classify into the target category. The procedure is applied to each test document and the target category. The results are shown in Table 5.

Table 5 shows categories, the number of training and test documents, and classification performance (F-score) with and without IDSS. Overall, the results showed that IDSS improved text classification performance. The best improvement was "fashion" (+0.330), and the poorest was "sports" (+0.012), as the F-score without replacement of the category "fashion" was very low(0) and that of "sports" had very high accuracy(0.971). It should be noted that our approach was not effective for the small size of training data, as the number of training data in the category "fashion" was only fifteen and the F-score was 0.330. Moreover, words assigned domain-specific sense by our approach did not have a "fashion" sense, as the topmost F-score was not larger than 0.330, regardless of how many words (senses) were used. This is not surprising because the F-score without word replacement was 0. The text classification used here is very simple, *i.e.*, SVM with vector representation of term frequency. There are many text classification techniques applicable to the small number of training documents (Nigam et al., 2000), and it will be worthwhile examining these with our approach.

[3] We manually evaluated senses that did not appear in the SFC resource.

Table 4: Some examples obtained by our approach.

Cat	Example of words and their senses	
legal/judicial	break:	(an escape from jail) “the breakout was carefully planned”
funding	depositary:	(a facility where things can be deposited for storage or safekeeping)
arts/entertainments	shopping	(the commodities purchased from stores)“she loaded her shopping into the car”;”woman carrying home shopping didn’t give me a second glance”
environment	pit	(a workplace consisting of a coal mine plus all the buildings and equipment connected with it)
fashion	fashion	(consumer goods (especially clothing) in the current mode)
sports	era	(baseball) a measure of a pitcher’s effectiveness
health	sickness	(the state that precedes vomiting)
science	float	(an air-filled sac near the spinal column in many fishes that helps maintain buoyancy)
religion	retreat	(the act of withdrawing or going backward (especially to escape something hazardous or unpleasant))
travel	runway	(a strip of level paved surface where planes can take off and land)
war	sturmabteilung	(Nazi militia created by Hitler in 1921 that helped him to power but was eclipsed by the SS after 1943)
weather	climatologist	(someone who is expert in climatology)

Table 5: Classification performance.

Cat	Train	Test	F-score without	F-score with	Cat	Train	Test	F-score without	F-score with
legal/judicial	909	1,003	.659	.769(+.110)	funding	3,936	3,770	.835	.883(+.048)
production	2,370	2,288	.589	.767(+.178)	research	258	245	.509	.707(+.198)
advertising	169	134	.504	.564(+.060)	management	929	964	.820	.874(+.054)
employment	1,384	1,671	.715	.783(+.068)	disasters	830	698	.748	.838(+.090)
arts/entertainment	326	327	.598	.701(+.103)	environment	497	489	.601	.734(+.133)
fashion	15	5	0	.330(+.330)	health	582	531	.611	.829(+.218)
labour issues	1,436	1,723	.772	.879(+.107)	religion	215	226	.602	.626(+.024)
science	215	233	.664	.723(+.059)	sports	2,882	2,955	.971	.983(+.012)
travel	71	81	.662	.791(+.129)	war	3,126	2,554	.784	.878(+.094)
elections	1,786	677	.679	.787(+.108)	weather	249	241	.715	.735(+.020)

4 RELATED WORK

Many real semantic-oriented applications, such as Question Answering, Paraphrasing, and Machine Translation systems, must have not only fine-grained and large-scale semantic knowledge, *e.g.*, WordNet, and COMLEX dictionary, but also tune the sense of the word heuristic depending on the domain in which the word is used. Magnini *et al.* presented a lexical resource where WordNet 2.0 synsets were annotated with Subject Field Codes (SFC) by a procedure that exploits WordNet structure (Magnini and Cavaglia, 2000). The results showed that 96% of WordNet synsets of the noun hierarchy could have been annotated, while identification of the domain labels for word senses was semi-automated and required a considerable amount of hand-labeling. Our approach is automated, and requires only documents from the given domain/category, such as the Reuters corpus, and dictionaries with gloss, such as WordNet. Therefore, it can be applied easily to a new domain or sense inventory, given sufficient documents.

In the context of WSD or semantic classification of words, Agirre *et al.* and Chan *et al.* presented a method of topic domain adaptation (Agirre and Lacalle, 2009; Chan and Ng, 2007). McCarthy *et al.* presented a method to find predominant noun senses automatically using a thesaurus acquired from raw textual corpora and WordNet similarity package (McCarthy et al., 2004). They used parsed data to find words with a similar distribution to the target word. Unlike (Buitelaar and Sacaleanu, 2001) *et al.* approach, they evaluated their method using publically available resources, *i.e.*, the method was evaluated against the hand-tagged resources SemCor and the SENSEVAL-2 English all-words task, and obtained precision of 64% on an all-nouns 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. They tested 38 words containing two domains of Sports and Finance from the Reuters corpus (Rose et al., 2002), while we tested 12 domains with 6,205 senses in all. Moreover, we applied the results to text classification to evaluate the results quantitatively.

In the context of link analysis, the graph-based ranking method has been widely and successfully used in NLP and its applications, such as WSD (Agirre and Soroa, 2009; Navigli and Lapata, 2010), text semantic similarity (Ramage et al., 2009), query expansion in IR (Lafferty and Zhai, 2001), and document summarization (Mihalcea and Tarau, 2005). The basic idea is that of "voting" or "recommendations" between nodes. The model first constructs a directed or undirected graph to reflect the relationships between the nodes and then applies the graph-based ranking algorithm to compute the rank scores for the nodes. The nodes with large rank scores are chosen as important nodes. Our methodology is the first aimed at applying a graph-based ranking method to identify the predominant sense depending on each domain/category.

5 CONCLUSIONS

We presented a method for identifying predominant sense of WordNet depending on each domain/category defined in the Reuters corpus. The average precision was 0.661, and recall against the Subject Field Code resources was 0.686. Moreover, the results applying text classification significantly improved classification accuracy. 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 our approach to find domain-specific senses with unknown words (Ciaramita and Johnson, 2003).

REFERENCES

Agirre, E. and Lacalle, O. L. (2009). Supervised Domain Adaption for WSD. In *Proc. of the EACL*, pages 42–50.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proc. of the EACL*, pages 33–41.

Bremaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag.

Brin, S. and Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117.

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

Chan, Y. S. and Ng, H. T. (2007). Domain Adaptation with Active Learning for Word Sense Disambiguation. In *Proc. of the ACL*, pages 49–56.

Ciaramita, M. and Johnson, M. (2003). Supersense Tagging of Unknown Nouns in WordNet. In *Proc. of the EMNLP2003*, pages 168–175.

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

Lafferty, J. and Zhai, C. (2001). Document Language Modeling, Query Models, and Risk Minimization for Information Retrieval. In *Proc. of the SIGIR2001*, pages 111–119.

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

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding Predominant Senses in Untagged Text. In *Proc. of the ACL*, pages 280–287.

Mihalcea, R. and Tarau, P. (2005). Language Independent Extractive Summarization. In *In Proc. of the ACL*, pages 49–52.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Proc. of the ARPA Workshop on HLT*, pages 303–308.

Navigli, R. and Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(4):678–692.

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

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134.

Ramage, D., Rafferty, A. N., and Manning, C. D. (2009). Random Walks for Text Semantic Similarity. In *Proc. of the 4th TextGraphs Workshop on Graph-based Algorithms in NLP*, pages 23–31.

Rose, T. G., Stevenson, M., and Whitehead, M. (2002). The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *Proc. of LREC-2002*.

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