

Word Sense Disambiguation

Using Multiple Contextual Features

Liang-Chih Yu*, Chung-Hsien Wu⁺, and Jui-Feng Yeh[#]

Abstract

Word sense disambiguation (WSD) is a technique used to identify the correct sense of polysemous words, and it is useful for many applications, such as machine translation (MT), lexical substitution, information retrieval (IR), and biomedical applications. In this paper, we propose the use of multiple contextual features, including the predicate-argument structure and named entities, to train two commonly used classifiers, Naïve Bayes (NB) and Maximum Entropy (ME), for word sense disambiguation. Experiments are conducted to evaluate the classifiers' performance on the OntoNotes corpus and are compared with classifiers trained using a set of baseline features, such as the bag-of-words, n-grams, and part-of-speech (POS) tags. Experimental results show that incorporating both predicate-argument structure and named entities yields higher classification accuracy for both classifiers than does the use of the baseline features, resulting in accuracy as high as 81.6% and 87.4%, respectively, for NB and ME.

Keywords: Word Sense Disambiguation, Predicate-Argument Structure, Named Entity, Natural Language Processing.

1. Introduction

A given word may have multiple meanings, and incorrect word sense recognition may reduce system effectiveness in semantic-oriented applications. Word sense disambiguation (WSD) identifies the correct sense of polysemous words, and it has emerged as a useful technique for

* Department of Information Management, Yuan-Ze University, Chung-Li, Taiwan, R.O.C.

E-mail: lcyu@saturn.yzu.edu.tw

⁺ Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.

E-mail: chwu@csie.ncku.edu.tw

[#] Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, R.O.C.

E-mail: ralph@mail.ncyu.edu.tw

many applications, such as machine translation (MT) (Carpuat & Wu, 2007; Chan *et al.*, 2007), lexical substitution (McCarthy, 2002; Dagan *et al.*, 2006), information retrieval (IR) (Agirre *et al.*, 2010), and biomedical applications (Schuemie *et al.*, 2005; Stevenson *et al.*, 2012). For example, in machine translation, WSD can be used to determine the correct translation for an ambiguous word. In lexical substitution, it is used to determine whether or not a target word can be replaced by another word (*e.g.*, a near synonym) by determining whether both words share a common sense. Currently, WSD has been a critical component in the SemEval workshop¹ series (Kilgarriff & Palmer, 2000; Edmonds & Kilgarriff, 2002; Agirre *et al.*, 2009).

Navigli (2009) provides an extensive survey of WSD approaches, investigating various features and machine learning algorithms to address specific tasks. For example, bag-of-words, n-grams, part-of-speech (POS) tags, and syntactic and semantic information have been used to build WSD systems with machine learning algorithms (Lee & Ng, 2002; Ando, 2006; Tratz *et al.*, 2007; Cai *et al.*, 2007; Agirre & Lopez de Lacalle, 2007; Specia *et al.*, 2007). Word sense annotated corpora, such as SemCor (Miller *et al.*, 1993), LDC-DSO (Ng & Lee, 1996), Hinoki (Kasahara *et al.*, 2004), and sense annotated corpora constructed with the help of Web users (Chklovski & Mihalcea, 2002) are also useful resources for building WSD systems. This paper proposes the use of multiple contextual features, including the predicate-argument structure and named entities, to train two commonly used classifiers: Naïve Bayes (NB) and Maximum Entropy (ME) from the OntoNotes corpus, a multilingual corpus of large-scale semantic annotations, including word senses, predicate-argument structure, ontology linking, and coreference (Hovy *et al.*, 2006; Pradhan *et al.*, 2007a). We then examine whether the two proposed features can improve WSD performance.

The rest of this work is organized as follows. Section 2 gives a brief description for the OntoNotes Corpus. Section 3 presents the features used to train classifiers for WSD. Section 4 summarizes the experimental results. Conclusions are drawn in Section 5.

2. Word Sense Annotation in OntoNotes Corpus

The OntoNotes corpus contains a set of sentences with word senses annotated. In the word sense inventory, the sense definitions of words are created by manually grouping fine-grained sense distinctions obtained from WordNet (Fellbaum, 1998) and dictionaries into more coarse-grained senses. There are two reasons for this grouping instead of using WordNet senses directly. First, people have trouble distinguishing many of the WordNet-level distinctions in real text and make inconsistent choices; thus, the use of coarse-grained senses can improve inter-annotator agreement (ITA) (Palmer *et al.*, 2004; 2006). Second, improved

¹ <http://www.senseval.org>

ITA enables machines to more accurately learn how to perform sense tagging automatically. Sense grouping in OntoNotes has been calibrated to ensure that ITA averages at least 90%. Table 1 shows the OntoNotes sense tags and definitions for the word *arm* (noun sense). Once the sense definitions are created, the sense of words in the sentences can be annotated. To accomplish this goal, the sentences containing the words in the inventory are retrieved first. For each target word (*i.e.*, a word in the inventory) in the sentences, its sense is annotated by two annotators, according to its sense definitions in the inventory. If the two annotators agree on the same sense, then their selection is stored in the corpus. Otherwise, the sense annotation is double-checked by an adjudicator for final decision. Recently, the OntoNotes corpus has been used for many applications, including the SemEval-2007 evaluation (Pradhan *et al.*, 2007b), sense merging (Snow *et al.*, 2007), class imbalance problems (Zhu & Hovy, 2007), sense pool verification (Yu *et al.*, 2007; 2010), parsing and named entity recognition (Finkel & Manning, 2009), semantic role labeling (Che *et al.*, 2010), and coreference resolution (Pradhan *et al.*, 2011).

Table 1. OntoNotes sense tags and definitions. The WordNet version is 2.1.

Sense Tag	Sense Definition	WordNet sense
arm.01	The forelimb of an animal	WN.1
arm.02	A weapon	WN.2
arm.03	A subdivision or branch of an organization	WN.3
arm.04	A projection, a narrow extension of a structure	WN.4 WN.5

3. The WSD System

The features used to build the WSD system include POS tags, local collocations, bag-of-words, named entities, and predicate-argument structure. These features are extracted from the OntoNotes corpus as follows.

Part-of-Speech (POS) tags: This feature includes the POS tags in the positions ($P_{-3}, P_{-2}, P_{-1}, P_0, P_1, P_2, P_3$), relative to the POS tag of the target word. For instance, the POS sequence of the constituent "...mediator in an attempt to break the..." is "NN NN IN DT TO VB DT".

Local Collocations: This feature includes single words and multi-word n-grams. The single words include ($W_{-3}, W_{-2}, W_{-1}, W_0, W_1, W_2, W_3$), relative to the target word W_0 . Similarly, the multi-word n-grams include ($W_{-2,-1}, W_{-1,1}, W_{1,2}, W_{-3,-2,-1}, W_{-2,-1,1}, W_{-1,1,2}, W_{1,2,3}$). For instance, the multi-word n-grams of the above example constituent include {in_an, an_to, to_break, mediator_in_an, in_an_to, an_to_break, to_break_the}.

Bag-of-Words: This feature can be considered a global feature, consisting of 5 words prior to and after the target word, without regard to position.

Named Entity: OntoNotes Release 1.0² provides 18 types of named entities, such as PERSON, ORGANIZATION, GPE, LOCATION, and PRODUCT.

Predicate-Argument Structure: The predicate-argument structure captures the semantic relations between the predicates and their arguments within a sentence. Consider the following example sentence.

[Arg₀ The New York arm of the London-based firm] auctioned off [Arg₁ the estate of John T. Dorrance Jr., the Campbell's Soup Co. heir,] [Arg_{M-TMP} last week].

The argument label Arg₀ is usually assigned to the *agent*, *causer*, and *experiencer*, while Arg₁ is usually assigned to the *patient*. The Arg_{M-TMP} represents a temporal modifier (Babko-Malaya, 2006; Palmer *et al.*, 2005). The predicate-argument structure of the above sentence is illustrated in Figure 1. The semantic relations can be either *direct* or *indirect*. A direct relation is used to model a verb-noun (VN), whereas an indirect relation is used to model a noun-noun (NN) relation. Additionally, an NN-relation can be built from the combination of two VN-relations with the same predicate. Table 2 presents some examples. For instance, NN1 can be built by combining VN1 and VN2. Therefore, the two features, VN1 and NN3, can be used to disambiguate the noun *arm*³.

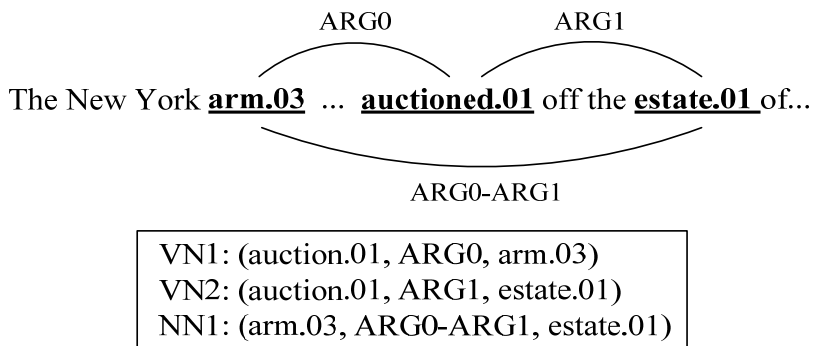


Figure 1. Example of predicate-argument structure.

Table 2. Examples of VN and NN-relations.

Relation Type	Example
VN relation V — ARG1 — N	VN1: (auction.01, Arg0, arm.03) VN2: (auction.01, Arg1, estate.01) VN3: (auction.01, ArgM-TMP, <DATE>)

² <http://www ldc.upenn.edu/Catalog/docs/LDC2007T21/ontonotes-1.0-documentation.pdf>.

³ Our WSD system does not include the sense identifier (except for the target word) for word-level training and testing.

<p>NN relation:</p> <pre> graph TD V[V] --- ARG0[ARG0] V --- ARG1[ARG1] ARG0 --- N1[N] ARG1 --- N2[N] </pre>	<p>NN1: (arm.03, Arg0-Arg1, estate.01) NN2: (estate.01, Arg1-ArgM-TMP, <DATE>) NN3: (arm.03, Arg0-ArgM-TMP, <DATE>)</p>
--	---

4. Experimental Results

4.1 Experimental Setup

OntoNotes Release 1.0 was used as the experimental corpus, with a total of 992 words in the sense inventory. Not all words, however, were polysemous, and some had a small number of sense annotated sentences. Therefore, we selected 477 polysemous words (247 nouns and 230 verbs) with at least 30 annotated sentences as the test data for the WSD task (see Table 3). The annotated sentences then were used to train two classifiers, Naïve Bayes (NB) and Maximum Entropy (ME), using the features presented in the previous section. We first trained the two classifiers using the baseline features, including the POS tag, local collocations, and bag-of-words. The named entities and predicate-argument structure then were added into both classifiers to determine whether these two features could improve WSD performance. The baseline classifier used for comparison was implemented using the principle of *most frequent sense* (MFS), with each word sense distribution retrieved from the OntoNotes corpus. The evaluation metric was accuracy, defined as the number of correctly identified senses (sentences) divided by the total number of test sentences.

Table 3. Statistics of the experimental data

		Nouns	Verbs
Num. of words (> 30 sentences)		247	230
Senses per word	Min.	2	2
	Avg.	3.26	3.58
	Max.	10	20
Sentences per sense	Min.	30	30
	Avg.	206	197
	Max.	3,053 (share.02)	2,551 (have.03)

4.2 Comparative Results

Table 4 shows the experimental results with 10-fold cross validation. The symbols B, PA, and NE in Table 4 represent the baseline features, predicate-argument structure, and named entities, respectively. For comparison of the classifiers, ME outperformed NB for all feature sets. For comparison of the feature sets, both B+PA and B+PA+NE outperformed B for both NB and ME, indicating that using both predicate-argument structure and named entities can improve performance over using the baseline features alone. Another observation is that the predicate-argument structure was more sensitive to ME than to NB because the improvement of B+PA over B in ME was greater than that in NB. Conversely, the named entity was more sensitive to NB.

Table 4. Comparative results of WSD accuracy for different features and classifiers.

Feature Types	Nouns			Verbs			ALL		
	MFS	NB	ME	MFS	NB	ME	MFS	NB	ME
B	0.820	0.810	0.865	0.764	0.797	0.856	0.793	0.805	0.862
B + PA		0.814	0.873		0.807	0.869		0.811	0.872
B + PA + NE		0.819	0.875		0.812	0.871		0.816	0.874

For more detailed analysis, Tables 5 and 6 list the WSD accuracy for parts of the nouns and verbs in the OntoNotes inventory. These words were also included in the SemEval-2007 English Lexical Sample Task (Pradhan *et al.*, 2007b).

Table 5. WSD accuracy for parts of the nouns.

Noun	# sense	MFS	NB			ME		
			B	B+PA	B+PA+NE	B	B+PA	B+PA+NE
authority	5	0.474	0.904	0.935	0.926	0.904	0.939	0.917
base	6	0.353	0.696	0.758	0.725	0.717	0.754	0.758
bill	4	0.668	0.872	0.881	0.887	0.895	0.901	0.916
carrier	8	0.765	0.704	0.808	0.815	0.758	0.792	0.819
chance	4	0.486	0.750	0.773	0.809	0.714	0.736	0.759
condition	3	0.713	0.800	0.823	0.839	0.806	0.803	0.842
defense	7	0.282	0.493	0.603	0.597	0.533	0.537	0.543
development	3	0.760	0.877	0.895	0.881	0.886	0.926	0.898
drug	2	0.684	0.783	0.811	0.845	0.791	0.789	0.800
effect	4	0.719	0.823	0.850	0.858	0.866	0.850	0.896
exchange	5	0.731	0.887	0.921	0.920	0.914	0.921	0.934

future	3	0.797	0.965	0.965	0.952	0.969	0.970	0.962
hour	4	0.854	0.847	0.880	0.882	0.863	0.888	0.873
job	3	0.780	0.738	0.757	0.768	0.809	0.849	0.845
management	2	0.618	0.837	0.853	0.866	0.821	0.840	0.806
network	3	0.605	0.750	0.788	0.824	0.705	0.750	0.736
order	8	0.722	0.871	0.877	0.892	0.876	0.869	0.883
part	5	0.702	0.915	0.900	0.907	0.944	0.931	0.940
people	4	0.912	0.917	0.915	0.918	0.925	0.933	0.937
point	9	0.737	0.853	0.851	0.875	0.885	0.877	0.870
policy	2	0.806	0.829	0.841	0.837	0.858	0.876	0.851
position	7	0.304	0.639	0.656	0.670	0.645	0.659	0.656
power	3	0.508	0.774	0.790	0.828	0.782	0.777	0.769
president	3	0.843	0.945	0.955	0.959	0.942	0.953	0.954
rate	2	0.924	0.944	0.933	0.940	0.946	0.943	0.955
source	5	0.368	0.803	0.841	0.833	0.797	0.844	0.830
space	5	0.565	0.741	0.782	0.794	0.829	0.788	0.806
state	4	0.830	0.840	0.848	0.855	0.858	0.858	0.857
system	6	0.544	0.728	0.749	0.751	0.722	0.717	0.705
Average	4.45	0.657	0.811	0.836	0.843	0.826	0.837	0.839

Table 6. WSD accuracy for parts of the verbs.

Verb	# sense	MSF	NB			ME		
			B	B+PA	B+PA+NE	B	B+PA	B+PA+NE
build	4	0.805	0.830	0.827	0.825	0.837	0.821	0.809
call	11	0.661	0.736	0.775	0.756	0.784	0.793	0.792
close	7	0.743	0.919	0.934	0.930	0.898	0.936	0.920
come	20	0.580	0.657	0.701	0.728	0.732	0.753	0.767
consider	2	0.788	0.840	0.836	0.852	0.875	0.891	0.905
cut	9	0.680	0.750	0.798	0.780	0.792	0.795	0.778
end	3	0.839	0.795	0.790	0.778	0.899	0.902	0.890
follow	7	0.666	0.766	0.795	0.804	0.756	0.825	0.825
get	14	0.447	0.656	0.682	0.676	0.721	0.748	0.748
go	18	0.275	0.545	0.599	0.585	0.617	0.672	0.664

grow	4	0.836	0.857	0.864	0.869	0.864	0.879	0.880
hold	11	0.667	0.737	0.756	0.759	0.754	0.764	0.749
keep	6	0.477	0.575	0.574	0.599	0.612	0.625	0.634
lead	3	0.417	0.859	0.882	0.886	0.870	0.895	0.891
leave	3	0.602	0.704	0.710	0.745	0.739	0.723	0.783
look	5	0.667	0.871	0.894	0.880	0.891	0.923	0.915
lose	6	0.709	0.806	0.829	0.867	0.835	0.835	0.861
make	13	0.336	0.557	0.616	0.613	0.604	0.664	0.670
put	12	0.677	0.757	0.756	0.755	0.789	0.808	0.806
raise	2	0.784	0.728	0.736	0.721	0.747	0.752	0.747
set	8	0.382	0.591	0.619	0.610	0.628	0.639	0.641
spend	2	0.700	0.878	0.972	0.969	0.885	0.987	0.991
take	20	0.663	0.611	0.619	0.616	0.670	0.684	0.683
tell	2	0.960	0.974	0.970	0.978	0.973	0.974	0.982
turn	16	0.285	0.591	0.635	0.623	0.705	0.726	0.723
Average	8.32	0.626	0.744	0.767	0.768	0.779	0.801	0.802

The “# sense” column lists the number of sense distinctions of a word, and the column “MFS” presents the sense distribution among all senses of the word. Both the number of sense distinctions and the sense distribution of words may affect WSD performance. Generally, a large number of sense distinctions with an even distribution may lead to confusion among the classifiers, hence, lower performance. For example, the noun *defense* in Table 5 has seven senses, and the proportion of the major sense is 0.282, indicating an even distribution (the distribution of the 7 senses is { .14, .18, .19, .08, .04, .28, .09 } in the OntoNotes corpus), thus yielding low accuracy. The verbs *go* and *make* in Table 6 also have similar results. Conversely, a small number of sense distinctions with a skewed distribution may have better performance. For example, in Table 5, the noun *rate* with a dominant sense of 0.924 yielded high accuracy, as did the verb *tell* in Table 6.

To further analyze the effect of the sense distribution of words in the whole corpus, we ranked the 247 nouns and 230 verbs in OntoNotes in descending order based on the proportion of their major senses. Nouns and verbs with major sense proportions within a given range then were grouped together (e.g., ≥ 0.95 , $0.90\sim 0.95$, $0.85\sim 0.90$, ..., $0.35\sim 0.4$, and < 0.35), and their average accuracy was calculated for comparison. Figures 2~4 present the results of nouns, verbs, and all words, respectively, with accuracy gradually decreasing as the sense becomes more evenly distributed. Another interesting observation is that, although ME outperformed NB, ME and NB achieved similar performance when the sense distribution became more

evenly distributed (the proportion of the major sense <0.65).

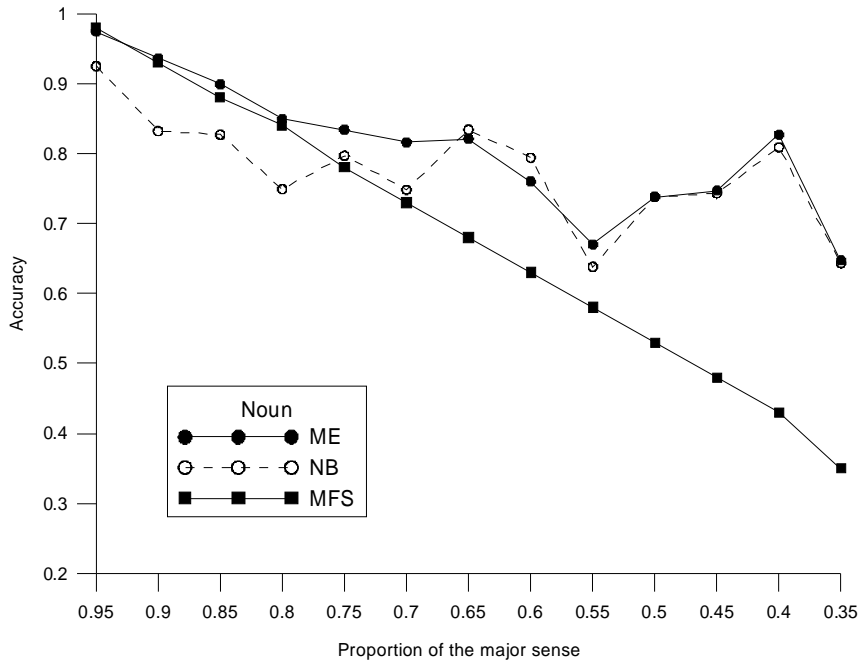


Figure 2. WSD performance against sense distribution for nouns.

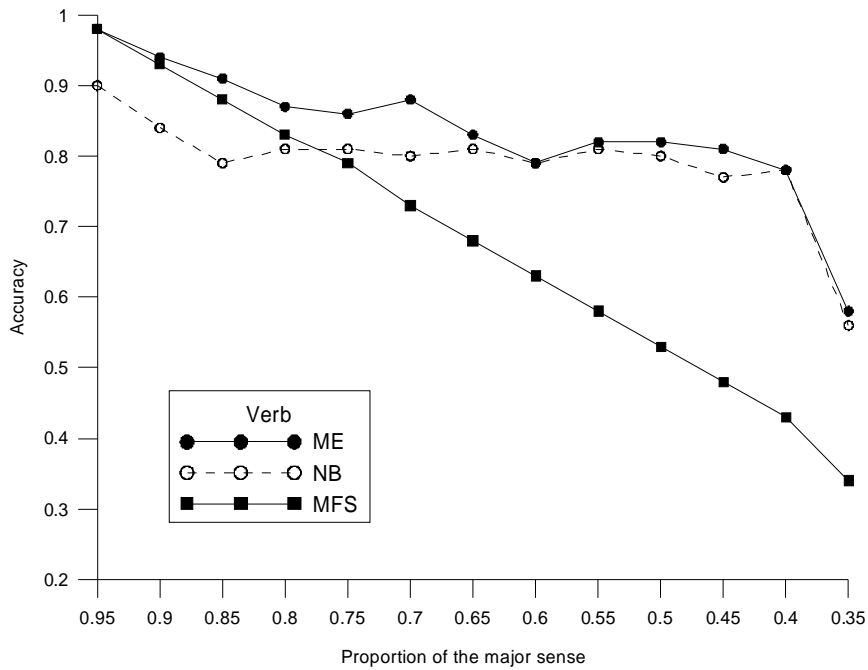


Figure 3. WSD performance against sense distribution for verbs.

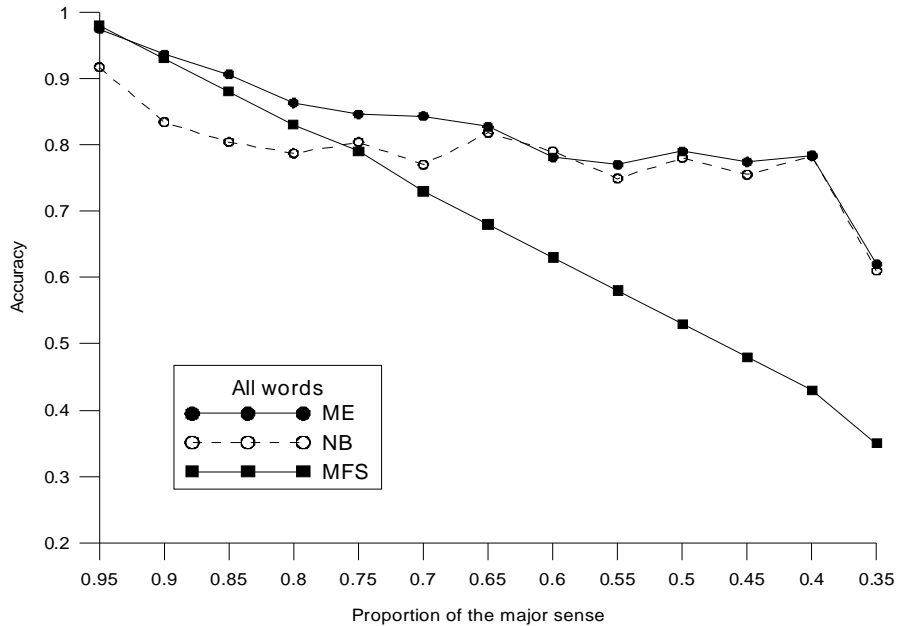


Figure 4. WSD performance against sense distribution for all words.

5. Conclusion

A WSD system was built from the OntoNotes corpus using multiple contextual features to analyze the effect of sense distribution on WSD performance. Experimental results show that both the predicate-argument structure and named entities improved WSD performance. In addition, there was a tendency for a skewed sense distribution to yield higher performance than evenly distributed word senses. Future work will focus on improving WSD performance by investigating more significant features and more effective machine learning algorithms.

Acknowledgments

This work was partially supported by the National Science Council, Taiwan, R.O.C., under Grant No. NSC 99-2221-E-155-036-MY3. The authors would also like to thank Dr. Eduard Hovy for his support.

References

- Agirre, E., & Lopez de Lacalle, O. (2007). UBC-ALM: Combining k-NN with SVD for WSD. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 342-345.
- Agirre, E., Mañriquez, L., & Wicentowski, R. (2009). Computational Semantic Analysis of Language: SemEval-2007 and Beyond. *Lang Resources and Evaluation*, 43(2), 97-104.

- Agirre, E., Otegi, A., & Zaragoza, H. (2010). Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR. *Lecture Notes in Computer Science*, 6241, 166-173.
- Ando, R.K. (2006). Applying Alternating Structure Optimization to Word Sense Disambiguation. In *Proc. of CoNLL-06*, 77-84.
- Babko-Malaya, O. (2006). PropBank Annotation Guidelines.
- Cai, J.F., Lee, W.S., & The, Y.W. (2007). Improving Word Sense Disambiguation Using Topic Features. In *Proc. of EMNLP/CoNLL-07*, 1015-1023.
- Carpuat, M., & Wu, D. (2007). Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proc. of EMNLP/CoNLL-07*, 61-72.
- Chan, Y.S., Ng, H.T., & Chiang, D. (2008). Word Sense Disambiguation Improves Statistical Machine Translation. In *Proc. of ACL-07*, 33-40.
- Che, W., Liu, T., & Li, Y. (2010). Improving Semantic Role Labeling with Word Sense. In *Proc. of HLT/NAACL-10*, 246-249.
- Chklovski, T., & Mihalcea, R. (2002). Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proc. of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions at ACL-02*, 116-122.
- Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E., & Strapparava, C. (2006). Direct Word Sense Matching for Lexical Substitution. In *Proc. of COLING/ACL-06*, 449-456.
- Edmonds, P., & Kilgarriff, A. (2002). Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. *Natural Language Engineering*, 8(4), 279-291.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Finkel, J. R., & Manning, C. D. (2009). Joint Parsing and Named Entity Recognition. In *Proc. of HLT/NAACL-09*, 326-334.
- Hovy, E.H., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proc. of HLT/NAACL-06*, 57-60.
- Kasahara, K., Sato, H., Bond, F., Tanaka, T., Fujita, S., Kanasugi, T., & Amano, S. (2004). Construction of a Japanese Semantic Lexicon: Lexeed. In *IPSG SIG: 2004-NLC-159*, Tokyo, 75-82.
- Kilgarriff, A., & Palmer, M. (2000). Introduction, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computer and the Humanities*, 34(1-2), 1-13.
- Lee, Y.K., & Ng, H.T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proc. of EMNLP*, 41-48.
- McCarthy, D. (2002). Lexical Substitution as a Task for WSD Evaluation. In *Proc. of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation at ACL-02*, 109-115.
- Miller, G., Leacock, C., Teng, R., & Bunker, R. (1993). A Semantic Concordance. In *Proc. of the 3rd DARPA Workshop on Human Language Technology*, 303-308.
- Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2), Article 10.

- Ng, H.T., & Lee, H.B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach. In *Proc. of ACL-96*, 40-47.
- Palmer, M., Babko-Malaya, O., & Dang, H.T. (2004). Different Sense Granularities for Different Applications. In *Proc. of the 2nd International Workshop on Scalable Natural Language Understanding at HLT/NAACL-04*.
- Palmer, M., Dang, H.T., & Fellbaum, C. (2006). Making Fine-grained and Coarse-grained Sense Distinctions, Both Manually and Automatically. *Journal of Natural Language Engineering*, 13, 137-163.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71-106.
- Pradhan, S., Hovy, E.H., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007a). OntoNotes: A Unified Relational Semantic Representation. In *Proc. of the First IEEE International Conference on Semantic Computing (ICSC-07)*, 517-524.
- Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007b). SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 87-92.
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proc. of CoNLL-11*, 1-27.
- Schuemie, M.J., Kors, J.A., & Mons, B. (2005). Word Sense Disambiguation in the Biomedical Domain: An Overview. *Journal of Computational Biology*, 12(5), 554-565.
- Snow, R., Prakash, S., Jurafsky, D., & Ng, A.Y. (2007). Learning to Merge Word Senses. In *Proc. of EMNLP/CoNLL-07*, 1005-1014.
- Specia, L., Stevenson, M., & das Gracas V. Nunes, M. (2007). Learning Expressive Models for Word Sense Disambiguation. In *Proc. of ACL-07*, 41-48.
- Stevenson, M., Agirre, E., & Soroa, A. (2012). Exploiting Domain Information for Word Sense Disambiguation of Medical Documents. *Journal of the American Medical Informatics Association*, 19(2), 235-240.
- Tratz, S., Sanfilippo, A., Gregory, M., Chappell, A., Posse, C., & Whitney, P. (2007). PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 264-267.
- Yu, L.C., Wu, C.H., Philpot, A., & Hovy, E.H. (2007). OntoNotes: Sense Pool Verification Using Google N-gram and Statistical Tests. In *Proc. of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*.
- Yu, L.C., Wu, C.H., Chang, R.Y., Liu, C.H., & Hovy, E.H. (2010). Annotation and Verification of Sense Pools in OntoNotes. *Information Processing and Management*, 46(4), 436-447.
- Zhu, J., & Hovy, E.H. (2007). Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In *Proc. of EMNLP/CoNLL-07*, 783-790.