One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction

Kaveh Taghipour
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
kaveh@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nght@comp.nus.edu.sg

Abstract

Supervised word sense disambiguation (WSD) systems are usually the best performing systems when evaluated on standard benchmarks. However, these systems need annotated training data to function properly. While there are some publicly available open source WSD systems, very few large annotated datasets are available to the research community. The two main goals of this paper are to extract and annotate a large number of samples and release them for public use, and also to evaluate this dataset against some word sense disambiguation and induction tasks. We show that the open source IMS WSD system trained on our dataset achieves state-of-the-art results in standard disambiguation tasks and a recent word sense induction task, outperforming several task submissions and strong baselines.

1 Introduction

Identifying the meaning of a word automatically has been an interesting research topic for a few decades. The approaches used to solve this problem can be roughly categorized into two main classes: Word Sense Disambiguation (WSD) and Word Sense Induction (WSI) (Navigli, 2009). For word sense disambiguation, some systems are based on supervised machine learning algorithms (Lee et al., 2004; Zhong and Ng, 2010), while others use ontologies and other structured knowledge sources (Ponzetto and Navigli, 2010; Agirre et al., 2014; Moro et al., 2014).

There are several sense-annotated datasets for WSD (Miller et al., 1993; Ng and Lee, 1996; Passonneau et al., 2012). However, these datasets either include few samples per word sense or only cover a small set of polysemous words. To overcome these limitations, automatic methods have been developed for annotating training samples. For example, Ng et al. (2003), Chan and Ng (2005), and Zhong and Ng (2009) used Chinese-English parallel corpora to extract samples for training their supervised WSD system. Diab (2004) proposed an unsupervised bootstrapping method to automatically generate a sense-annotated dataset. Another example of automatically created datasets is the semi-supervised method used in (Kübler and Zhekova, 2009), which employed a supervised classifier to label instances.

The two main contributions of this paper are as follows. First, we employ the same method used in (Ng et al., 2003; Chan and Ng, 2005) to *semi-automatically* annotate one million training samples based on the WordNet sense inventory (Miller, 1995) and release the annotated corpus for public use. To our knowledge, this annotated set of sense-tagged samples is the largest publicly available dataset for word sense disambiguation. Second, we train an open source supervised WSD system, IMS (Zhong and Ng, 2010), using our data and evaluate it against standard WSD and WSI benchmarks. We show that our system outperforms other state-of-the-art systems in most cases.

As any WSD system is also a WSI system when we treat the pre-defined sense inventory of the WSD system as the induced word senses, a WSD system can also be evaluated and used for WSI. Some researchers believe that, in some cases, WSI methods may perform better than WSD systems (Jurgens and Klapaftis, 2013; Wang et al., 2015). However, we argue that WSI systems have few advantages compared to WSD methods and according to our results, disambiguation systems consistently outperform induction systems. Although there are some cases where WSI systems can be useful (e.g., for resource-poor languages), in most cases WSD systems are preferable because

Proceedings of the 19th Conference on Computational Language Learning, pages 338–344,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics

of higher accuracy and better interpretability of output.

The rest of this paper is composed of the following sections. Section 2 explains our methodology for creating the training data. We evaluate the extracted data in Section 3 and finally, we conclude the paper in Section 4.

2 Training Data

In order to train a supervised word sense disambiguation system, we extract and sense-tag data from a freely available parallel corpus, in a *semi-automatic* manner. To increase the coverage and therefore the ultimate performance of our WSD system, we also make use of existing sense-tagged datasets. This section explains each step in detail.

Since the main purpose of this paper is to build and release a publicly available training set for word sense disambiguation systems, we selected the MultiUN corpus (MUN) (Eisele and Chen, 2010) produced in the EuroMatrixPlus project[1]. This corpus is freely available via the project website and includes seven languages. An automatically sentence-aligned version of this dataset can be downloaded from the OPUS website[2] and therefore we decided to extract samples from this sentence-aligned version.

To extract training data from the MultiUN parallel corpus, we follow the approach described in (Chan and Ng, 2005) and select the Chinese-English part of the MultiUN corpus. The extraction method has the following steps:

1. Tokenization and word segmentation: The English side of the corpus is tokenized using the Penn TreeBank tokenizer[3], while the Chinese side of the corpus is segmented using the Chinese word segmenter of (Low et al., 2005).

2. Word alignment: After tokenizing the texts, GIZA++ (Och and Ney, 2000) is used to align English and Chinese words.

3. Part-of-speech (POS) tagging and lemmatization: After running GIZA++, we use the OpenNLP POS tagger[4] and then the WordNet lemmatizer to obtain POS tags and word lemmas of the English sentence.

4. Annotation: In order to assign a WordNet sense tag to an English word w_e in a sentence, we make use of the aligned Chinese translation w_c of w_e, based on the automatic word alignment formed by GIZA++. For each sense i of w_e in the WordNet sense inventory (WordNet 1.7.1), a list of Chinese translations of sense i of w_e has been manually created. If w_c matches one of these Chinese translations of sense i, then w_e is tagged with sense i.

The average time needed to manually assign Chinese translations to the word senses of one word type for noun, adjective, and verb is 20, 25, and 40 minutes respectively (Chan, 2008). The above procedure annotates the top 60% most frequent word types (nouns, verbs, and adjectives) in English, selected based on their frequency in the Brown corpus. This set of selected word types includes 649 nouns, 190 verbs, and 319 adjectives.

Since automatic sentence and word alignment can be noisy, and a Chinese word w_c can occasionally be a valid translation of more than one sense of an English word w_e, the senses tagged using the above procedure may be erroneous. To get an idea of the accuracy of the senses tagged with this procedure, we manually evaluated a subset of 1,000 randomly selected sense-tagged instances. Although the sense inventory is fine-grained (WordNet 1.7.1), the sense-tag accuracy achieved is 83.7%. We also performed an error analysis to identify the sources of errors. We found that only 4% of errors are caused by wrong sentence or word alignment. However, 69% of erroneous sense-tagged instances are the result of a Chinese word associated with multiple senses of a target English word. In such cases, the Chinese word is linked to multiple sense tags and therefore, errors in sense-tagged data are introduced. Our results are similar to those reported in (Chan, 2008).

To speed up the training process, we perform random sampling on the sense tags with more than 500 samples and limit the number of samples *per sense* to 500. However, all samples of senses with fewer than 500 samples are included in the training data. This sampling method ensures that rare sense tags also have training samples during the selection process.

In order to improve the coverage of the training set, we augment it by adding samples from SEMCOR (SC) (Miller et al., 1993) and the DSO cor-

[1] http://www.euromatrixplus.eu/multi-un
[2] http://opus.lingfil.uu.se/MultiUN.php
[3] http://www.cis.upenn.edu/∼treebank/tokenization.html
[4] http://opennlp.apache.org

	Avg. # samples per word type
MUN (before sampling)	19,837.6
MUN	852.5
MUN+SC	55.4
MUN+SC+DSO	63.7

Table 3: Average number of samples per word type (WordNet 1.7.1)

pus (Ng and Lee, 1996). We only add the 28 most frequent adverbs from SEMCOR because we observe almost no improvement when adding all adverbs. We notice that the DSO corpus generally improves the performance of our system. However, since the annotated DSO corpus is copyrighted, we are unable to release a dataset including the DSO corpus. Therefore, we experiment with two different configurations, one with the DSO corpus and one without, although the released dataset will not include the DSO corpus.

Since some shared tasks use newer WordNet versions, we convert the training set sense labels using the sense mapping files provided by WordNet[5]. As replicating our results requires WordNet versions 1.7.1, 2.1, and 3.0, we release our sense-tagged dataset in all three versions. Some statistics about the sense-tagged training set can be found in Table 1 to Table 3.

3 Evaluation

For the WSD system, we use IMS (Zhong and Ng, 2010) in our experiments. IMS is a supervised WSD system based on support vector machines (SVM). This WSD system comes with out-of-the-box pre-trained models. However, since the original training set is not released, we use our own training set (see Section 2) to train IMS and then evaluate it on standard WSD and WSI benchmarks. This section presents the results obtained on four WSD and one WSI shared tasks. The four all-words WSD shared tasks are SensEval-2 (Edmonds and Cotton, 2001), SensEval-3 task 1 (Snyder and Palmer, 2004), and both the fine-grained task 17 and coarse-grained task 7 of SemEval-2007 (Pradhan et al., 2007; Navigli et al., 2007). These all-words WSD shared tasks provide no training data to the participants. The selected word sense induction task in our experiments is SemEval-2013 task 13 (Jurgens and Klapaftis, 2013).

3.1 WSD All-Words Tasks

The results of our experiments on WSD tasks are presented in Table 4. For the SensEval-2 and SensEval-3 test sets, we use the training set with the WordNet 1.7.1 sense inventory and for the SemEval-2007 test sets, we use training data with the WordNet 2.1 sense inventory.

In Table 4, IMS (original) refers to the IMS system trained with the original training instances as reported in (Zhong and Ng, 2010). We also compare our systems with two other configurations obtained from training IMS on SEMCOR, and SEMCOR plus DSO datasets. In Table 4, these two settings are shown by IMS (SC) and IMS (SC+DSO), respectively. Finally, Rank 1 and Rank 2 are the top two participating systems in the respective all-words tasks.

As shown in Table 4, our systems (both with and without the DSO corpus as training instances) perform competitively with and in some cases even better than the original IMS and also the best shared task submissions. This shows that our training set is of high quality and training a supervised WSD system using our training data achieves state-of-the-art results on the all-words tasks. Since the MUN dataset does not cover all target word types in the all-words shared tasks, the accuracy achieved with MUN alone is lower than the SC and SC+DSO settings. However, the evaluation results show that IMS trained on MUN alone often performs better than or is competitive with the WordNet Sense 1 baseline. Finally, it can be seen that adding the training instances from MUN (that is, IMS (MUN+SC) and IMS (MUN+SC+DSO)) often achieves higher accuracy than without MUN instances (IMS (SC) and IMS (SC+DSO)).

3.2 SemEval-2013 Word Sense Induction Task

In order to evaluate our system on a word sense induction task, we selected SemEval-2013 task 13, the latest WSI shared task. Unlike most other tasks that assume a single sense is sufficient for representing word senses, this task allows each instance to be associated with multiple sense labels with their applicability weights. This WSI task considers 50 lemmas, including 20 nouns, 20 verbs, and 10 adjectives, annotated with the WordNet 3.1

[5] http://wordnet.princeton.edu/wordnet/download/current-version/

	noun	verb	adjective	adverb	total
MUN (before sampling)	649	190	319	0	1,158
MUN	649	190	319	0	1,158
MUN+SC	11,446	4,705	5,129	28	21,308
MUN+SC+DSO	11,446	4,705	5,129	28	21,308

Table 1: Number of word types in each part-of-speech (WordNet 1.7.1)

	number of training samples					
	noun	verb	adjective	adverb	total	size
MUN (before sampling)	14,492,639	4,400,813	4,078,543	0	22,971,995	17.7 GB
MUN	503,408	265,785	218,046	0	987,239	745 MB
MUN+SC	582,028	341,141	251,362	6,207	1,180,738	872 MB
MUN+SC+DSO	687,871	412,482	251,362	6,207	1,357,922	939 MB

Table 2: Number of training samples in each part-of-speech (WordNet 1.7.1). The size column shows the total size of each dataset in megabytes or gigabytes.

sense inventory. We use WordNet 3.0 in our experiments on this task.

We evaluated our system using all measures used in the shared task. The results are presented in Table 5. The columns in this table denote the scores of the various systems according to the different evaluation metrics used in the WSI shared task, which are Jaccard Index, K_{δ}^{sim}, WNDCG, Fuzzy NMI, and Fuzzy B-Cubed. See (Jurgens and Klapaftis, 2013) for details of the evaluation metrics.

This table also includes the top two systems in the shared task, `AI-KU` (Baskaya et al., 2013) and `Unimelb` (Lau et al., 2013), as well as `Wang-15` (Wang et al., 2015). `AI-KU` uses a language model to find the most likely substitutes for a target word to represent the context. The clustering method used in `AI-KU` is K-means and the system gives good performance in the shared task. `Unimelb` relies on Hierarchical Dirichlet Process (Teh et al., 2006) to identify the sense of a target word using positional word features. Finally, `Wang-15` uses Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model the word sense and topic jointly. This system obtains high scores, according to Fuzzy B-Cubed and Fuzzy NMI measures. The last three rows are some baseline systems: grouping all instances into one cluster, grouping each instance into a cluster of its own, and assigning the most frequent sense in SEMCOR to all instances. As shown in Table 5, training IMS on our training data outperforms all other systems on three out of five evaluation metrics, and performs competitively on the remaining two metrics.

IMS trained on MUN alone (IMS (MUN)) outperforms IMS (SC) and IMS (SC+DSO) in terms of the first three evaluation measures, and achieves comparable Fuzzy NMI and Fuzzy B-Cubed scores. Moreover, the evaluation results show that IMS (MUN) often performs better than the SEMCOR most frequent sense baseline. Finally, it can be observed that in most cases, adding training instances from MUN significantly improves IMS (SC) and IMS (SC+DSO).

4 Conclusion

One of the major problems in building supervised word sense disambiguation systems is the training data acquisition bottleneck. In this paper, we semi-automatically extracted and sense-tagged an English corpus containing one million sense-tagged instances. This large sense-tagged corpus can be used for training any supervised WSD systems. We then evaluated the performance of IMS trained on our sense-tagged corpus in several WSD and WSI shared tasks. Our sense-tagged dataset has been released publicly[6]. We believe our dataset is the largest publicly available annotated dataset for WSD at present.

After training a supervised WSD system using our training set, we evaluated the system using standard benchmarks. The evaluation results show that our sense-tagged corpus can be used to build a WSD system that performs competitively with the

[6]http://www.comp.nus.edu.sg/∼nlp/corpora.html

	SensEval-2	SensEval-3	SemEval-2007	
	Fine-grained	**Fine-grained**	**Fine-grained**	**Coarse-grained**
IMS (MUN)	64.5	60.6	52.7	78.7
IMS (MUN+SC)	68.2	67.4	58.5	81.6
IMS (MUN+SC+DSO)	68.0	66.6	58.9	82.3
IMS (original)	68.2	**67.6**	58.3	**82.6**
IMS (SC)	66.1	67.0	58.7	81.9
IMS (SC+DSO)	66.5	67.0	57.8	81.6
Rank 1	**69.0**	65.2	**59.1**	82.5
Rank 2	63.6	64.6	58.7	81.6
WordNet Sense 1	61.9	62.4	51.4	78.9

Table 4: Accuracy (in %) on all-words word sense disambiguation tasks

	Jac. Ind.	K_{δ}^{sim}	**WNDCG**	**Fuzzy NMI**	**Fuzzy B-Cubed**
IMS (MUN)	24.6	64.9	33.0	6.9	57.1
IMS (MUN+SC)	25.0	**65.4**	34.2	9.1	55.9
IMS (MUN+SC+DSO)	**25.5**	**65.4**	35.1	**9.7**	55.4
IMS (original)	23.4	64.5	34.0	8.6	59.0
IMS (SC)	22.9	63.5	32.4	6.8	57.3
IMS (SC+DSO)	23.4	63.6	32.9	7.1	57.6
Wang-15 (ukWac)	-	-	-	**9.7**	54.5
Wang-15 (actual)	-	-	-	9.4	59.1
AI-KU (base)	19.7	62.0	**38.7**	6.5	39.0
AI-KU (add1000)	19.7	60.6	21.5	3.5	32.0
AI-KU (remove5-add1000)	24.4	64.2	33.2	3.9	45.1
Unimelb (5p)	21.8	61.4	36.5	5.6	45.9
Unimelb (50k)	21.3	62.0	37.1	6.0	48.3
all-instances-1cluster	19.2	60.9	28.8	0.0	**62.3**
each-instance-1cluster	0.0	0.0	0.0	7.1	0.0
SEMCOR most freq sense	19.2	60.9	28.8	0.0	**62.3**

Table 5: Supervised and unsupervised evaluation results (in %) on SemEval-2013 word sense induction task

top performing WSD systems in the SensEval-2, SensEval-3, and SemEval-2007 fine-grained and coarse-grained all-words tasks, as well as the top systems in the SemEval-2013 WSI task.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We are also grateful to Christian Hadiwinoto and Benjamin Yap for assistance with performing the error analysis, and to the anonymous reviewers for their helpful comments.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Osman Baskaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. AI-KU: using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In

Proceedings of the 20th National Conference on Artificial Intelligence, pages 1037–1042.

Yee Seng Chan. 2008. *Word Sense Disambiguation: Scaling up, Domain Adaptation, and Application to Machine Translation*. Ph.D. thesis, National University of Singapore.

Mona Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 303–310.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2868–2872.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299.

Sandra Kübler and Desislava Zhekova. 2009. Semi-supervised learning for word sense disambiguation: Quality vs. quantity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 197–202.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013. unimelb: topic modelling-based word sense induction. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311.

Yoong Keok Lee, Hwee Tou Ng, and Tee Kiah Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, pages 303–308.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

Rebecca Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC word sense sentence corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3025–3030.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71.

Zhi Zhong and Hwee Tou Ng. 2009. Word sense disambiguation for all words without hard labor. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pages 1616–1621.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics System Demonstrations*, pages 78–83.