Word Sense Disambiguation for All Words using Tree-Structured Conditional Random Fields

Jun Hatori* **Yusuke Miyao*** **Jun'ichi Tsujii***†‡
*Graduate School of Interdisciplinary Information Studies, University of Tokyo
*Graduate School of Information Science and Technology, University of Tokyo
†National Centre for Text Mining / 131 Princess Street, Manchester, M1 7DN, UK
‡School of Computer Science, University of Manchester
{hatori,yusuke,tsujii}@is.s.u-tokyo.ac.jp

Abstract

We propose a supervised word sense disambiguation (WSD) method using tree-structured conditional random fields (TCRFs). By applying TCRFs to a sentence described as a dependency tree structure, we conduct WSD as a labeling problem on tree structures. To incorporate dependencies between word senses, we introduce a set of features on tree edges, in combination with coarse-grained tagsets, and show that these contribute to an improvement in WSD accuracy. We also show that the tree-structured model outperforms the linear-chain model. Experiments on the SENSEVAL-3 data set show that our TCRF model performs comparably with state-of-the-art WSD systems.

1 Introduction

Word sense disambiguation (WSD) is one of the fundamental underlying problems in computational linguistics. The task of WSD is to determine the appropriate sense for each polysemous word within a given text.

Traditionally, there are two task settings for WSD: the *lexical sample task*, in which only one targeted word is disambiguated given its context, and the *all-words task*, in which all content words within a text are disambiguated. Whilst most of the WSD research so far has been toward the lexical sample task, the all-words task has received relatively less attention, suffering from a serious knowledge bottleneck problem. Since it is considered to be a necessary step toward practical applications, there is an urgent need to improve the performance of WSD systems that can handle the all-words task.

In this paper, we propose a novel approach for the all-words task based on *tree-structured conditional random fields* (TCRFs). Our TCRF model incorporates the inter-word sense dependencies, in combination with WORDNET hierarchical information and a coarse-grained tagset, namely supersenses, by which we can alleviate the data sparseness problem.

2 Background

2.1 Inter-word sense dependencies

Since the all-words task requires us to disambiguate all content words, it seems reasonable to assume that we could perform better WSD by considering the sense dependencies among words, and optimizing word senses over the whole sentence. Specifically, we base our model on the assumption that there are strong sense dependencies between a head word and its dependents in a dependency tree; therefore, we employ the dependency tree structures for modeling the sense dependencies.

There have been a few WSD systems that incorporate the inter-word sense dependencies (e.g. Mihalcea and Faruque (2004)). However, to the extent of our knowledge, their effectiveness has not explicitly examined thus far for supervised WSD.

2.2 WORDNET information

Supersense A *supersense* corresponds to the lexicographers' file ID in WORDNET, with which each noun or verb synset is associated. Since they are originally introduced for ease of lexicographers' work, their classification is fairly general, but not too abstract, and is hence expected to act as good coarse-grained semantic categories. The numbers of the supersenses are 26 and 15 for nouns and verbs. The effectiveness of the use of supersenses and other coarse-grained tagsets for WSD has been recently shown by several researchers (e.g. Kohomban and Lee (2005), Ciaramita and Altun (2006), and Mihalcea et al. (2007)).

Sense number A *sense number* is the number of a sense of a word in WordNet. Since senses of a word are ordered according to frequency, the sense number can act as a powerful feature for WSD, which offers a preference for frequent senses, and especially as a back-off feature, which enables our model to output the first sense when no other feature is available for that word.

2.3 Tree-structured CRFs

Conditional Random Fields (CRFs) are graph-based probabilistic discriminative models proposed by Lafferty et al. (2001).

Tree-structured CRFs (TCRFs) are different from widely used linear-chain CRFs, in that the probabilistic variables are organized in a tree structure rather than in a linear sequence. Therefore, we can consider them more appropriate for modeling the semantics of sentences, which cannot be represented by linear structures.

Although TCRFs have not yet been applied to WSD, they have already been applied to some NLP tasks, such as semantic annotation (Tang et al., 2006), proving to be useful in modeling the semantic structure of a text.

Formulation In CRFs, the conditional probability of a label set $\mathbf{y}$ for an observation sequence $\mathbf{x}$ is calculated by

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\Big[\sum_{e \in E, j} \lambda_j f_j(e, \mathbf{x}, \mathbf{y}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{x}, \mathbf{y})\Big] \quad (1)$$

where E and V are the sets of edges and vertices, f_j and g_k are the feature vectors for an edge and a vertex, λ_j and μ_k are the weight vectors for them, and $Z(\mathbf{x})$ is the normalization function. For a detailed description of TCRFs, see Tang et al. (2006).

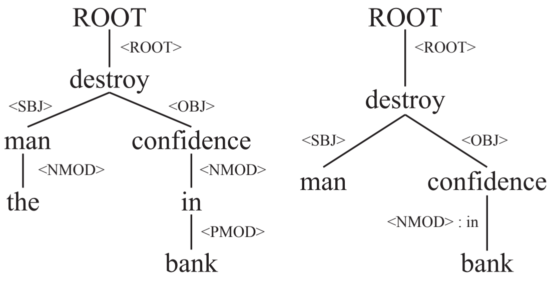

Figure 1: An example sentence described as a dependency tree structure.

3 WSD Model using Tree-structured CRFs

3.1 Overview

Let us consider the following sentence.

(i) *The man destroys confidence in banks.*

In the beginning, we parse a given sentence by using a dependency parser. The left-hand side of Figure 1 shows the dependency tree for Sentence (i) in the CoNLL-X dependency format.

Next, we convert the outputted tree into a tree of content words, as illustrated in the right-hand side of Figure 1, since our WSD task does not focus on the disambiguation of function words.

Finally, we conduct WSD as a labeling task on tree structures, by maximizing the probability of a tree of word senses, given scores for vertex and edge features.

3.2 Sense Labels

Using the information in WordNet, we define four sense labels for a word: a *sense* $s_1(v)$, a *synset* $s_2(v)$, a *topmost synset* $s_3(v)$, and a *supersense* $s_4(v)$. A topmost synset $s_3(v)$ is the superordinate synset at the topmost level in the WordNet hierarchy, and note that a supersense $s_4(v)$ is only available for nouns and verbs. We incorporate all these labels together into the vertex and edge features described in the following sections.

3.3 Vertex features

Most of the vertex features we use are those used by Lee and Ng (2002). All these features are combined with each of the four sense labels $s_n(v)$, and incorporated as g_k in Equation (1).

- Word form, lemma, and part of speech.
- Word forms, lemmas, and parts of speech of the head and dependents in a dependency tree.

	#sentences	#words
Development	470	5,178
Brown-1	10,712	100,804
Brown-2	8,956	85,481
SENSEVAL-3	300	2,081

Table 1: Statistics of the corpora.

- Bag-of-words within 60-words window.
- Parts-of-speech of neighboring six words.
- Local n-gram within neighboring six words.

Additionally, we include as a vertex feature the sense number, introduced in Section 2.2.

3.4 Edge features

For each edge, all possible sense bigrams (i.e. $s_1(v)$-$s_1(v')$,$s_1(v)$-$s_2(v')$,$\cdots$,$s_4(v)$-$s_4(v')$), and the combination of sense bigrams with dependency relation labels (e.g. 'SUB,' 'NMOD') and/or removed function words in between (e.g. 'of,' 'in') are defined as edge features, which correspond to f_j in Equation (1).

4 Experiment

4.1 Experimental settings

In the experiment, we use as our main evaluation data set the *Brown-1* and *Brown-2* sections of SEMCOR. The last files in the five largest categories in Brown-1 are used for development, and the rest of Brown-1 and all files in Brown-2 are alternately used for training and testing. We also use the SENSEVAL-3 English all-words data (Snyder and Palmer, 2004) for testing, in order to compare the performance of our model with other systems. The statistics of the data sets are shown in Table 1.

All sentences are parsed by the Sagae's dependency parser (Sagae and Tsujii, 2007), and the TCRF model is trained using Amis (Miyao and Tsujii, 2002). During the development phase, we tune the parameter of L_2 regularization for CRFs. Note that, in all experiments, we try all content words annotated with WORDNET synsets; therefore, the recalls are always equal to the precisions.

4.2 Results

First, we trained and evaluated our models on SEMCOR. Table 2 shows the overall performance of our models. BASELINE model is the first sense baseline. NO-EDGE model uses only the vertex features, while each of the Sn-EDGE models makes use of the edge features associated with a sense label s_n, where $n \in \{1, 2, 3, 4\}$. The ALL-EDGE model incorporates all possible combinations of sense labels. The only difference in the ALL-EDGE' model is that it omits features associated with dependency relation labels, so that we can compare the performance with the ALL-EDGE'(Linear) model, which is based on the linear-chain model.

In the experiment, all models with one or more edge features outperformed both the NO-EDGE and BASELINE model. The ALL-EDGE model achieved 75.78% and 77.49% recalls for the two data sets, with 0.41% and 0.43% improvements over the NO-EDGE model. By the stratified shuffling test (Cohen, 1995), these differences are shown to be statistically significant[1], with the exception of S3-EDGE model. Also, the tree-structured model ALL-EDGE' is shown to outperform the linear-chain model ALL-EDGE'(Linear) by 0.13% for both data sets ($p = 0.013, 0.006$).

System	Recall
PNNL (Tratz et al., 2007)	67.0%
Simil-Prime (Kohomban and Lee, 2005)	66.1%
ALL-EDGE	65.5%
GAMBL (Decadt et al., 2004)	65.2%
SENSELEARNER (Mihalcea et al.,2004)	64.6%
BASELINE	62.2%

Table 3: The comparison of the performance of WSD systems evaluated on the SENSEVAL-3 English all-words test set.

Finally, we trained our models on the Brown-1 and Brown-2 sections, and evaluated them on the SENSEVAL-3 English all-words task data. Table 3 shows the comparison of our model with the state-of-the-art WSD systems. Considering the difference in the amount of training data, we can conclude that the performance of our TCRF model is comparable to state-of-the-art WSD systems, for all systems in Table 3 other than Simil-Prime (Kohomban and Lee, 2005)[2] utilizes other sense-annotated data, such as the SENSEVAL data sets and example sentences in WORDNET.

[1] Although some of the improvements seem marginal, they are still statistically significant. This is probably because sense bigram features are rarely active, given the size of the training corpus, and most of the system outputs are the first senses. Indeed, 91.3% of the outputs of ALL-EDGE model are the first senses, for example.

[2] Kohomban and Lee (2005) used almost the same training data as our system, but they utilize the instance weighting technique and the combination of several classifiers, which our system does not.

Training set	Brown-1				Brown-2			
Testing set	Brown-2				Brown-1			
Model	Recall	Offset		#correct	Recall	Offset		#correct
ALL-EDGE’	75.77%	0.40%	$\gg$	64766/85481	77.45%	0.39%	$\gg$	78077/100804
ALL-EDGE’ (Linear)	75.64%	0.27%	$\gg$	64662/85481	77.32%	0.26%	$\gg$	77944/100804
ALL-EDGE	75.78%	0.41%	$\gg$	64779/85481	77.49%	0.43%	$\gg$	78114/100804
S4-EDGE	75.46%	0.09%	$\gg$	64507/85481	77.15%	0.09%	$\gg$	77769/100804
S3-EDGE	75.40%	0.03%	$\sim$	64452/85481	77.13%	0.07%	$\gg$	77750/100804
S2-EDGE	75.45%	0.08%	$\gg$	64494/85481	77.12%	0.06%	$\gg$	77738/100804
S1-EDGE	75.44%	0.07%	$\gg$	64491/85481	77.10%	0.04%	$>$	77724/100804
NO-EDGE	75.37%	0.00%		64427/85481	77.06%	0.00%		77677/100804
BASELINE	74.36%			63567/85481	75.91%			76524/100804

Table 2: The performance of our system trained and evaluated on SEMCOR. The statistical significance of the improvement over NO-EDGE model is shown in the ‘Offset’ fields, where ‘$\gg$,’ ‘$>$,’ and ‘$\sim$’ denote $p < 0.01$, $p < 0.05$, and $p \geq 0.05$, respectively.

5 Conclusion

In this paper, we proposed a novel approach for the all-words WSD based on TCRFs. Our proposals are twofold: one is to apply tree-structured CRFs to dependency trees, and the other is to use bigrams of fine- and coarse-grained senses as edge features.

In our experiment, the sense dependency features are shown to improve the WSD accuracy. Since the combination with coarse-grained tagsets are also proved to be effective, they can be used to alleviate the data sparseness problem. Moreover, we explicitly proved that the tree-structured model outperforms the linear-chain model, indicating that dependency trees are more appropriate for representing semantic dependencies.

Although our model is based on a simple framework, its performance is comparable to state-of-the-art WSD systems. Since we can use additionally other sense-annotated resources and sophisticated machine learning techniques, our model still has a great potential for improvement.

References

Ciaramita, M. and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Cohen, P. R. 1995. *Empirical methods for artificial intelligence*. MIT Press.

Decadt, B., V. Hoste, W. Daelemans, and A. V. den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In *Senseval-3: Third Int’l Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Kohomban, U. S. and W. S. Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*.

Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of 18th Int’l Conf. on Machine Learning (ICML)*.

Lee, Y. K. and H. T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Mihalcea, R. and E. Faruque. 2004. SenseLearner: Minimally supervised word sense disambiguation for all words in open text. In *Proc. of ACL/SIGLEX Senseval-3*, Barcelona, Spain, July.

Mihalcea, R., A. Csomai, and M. Ciaramita. 2007. UNT-Yahoo: SuperSenseLearner: Combining SenseLearner with SuperSense and other coarse semantic features. In *Proc. of the 4th Int’l Workshop on the Semantic Evaluations (SemEval-2007)*.

Miyao, Y. and J. Tsujii. 2002. Maximum entropy estimation for feature forests. In *Proc. of Human Language Technology Conf. (HLT 2002)*.

Sagae, K. and J. Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

Snyder, B. and M. Palmer. 2004. The english all-words task. In *Senseval-3: Third Int’l Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Tang, J., M. Hong, J. Li, and B. Liang. 2006. Tree-structured conditional random fields for semantic annotation. In *Proc. of the 5th Int’l Semantic Web Conf.*

Tratz, S., A. Sanfilippo, M. Gregory, A. Chappell, C. Posse, and P. Whitney. 2007. PNNL: A supervised maximum entropy approach to word sense disambiguation. In *Proc. of the 4th Int’l Workshop on Semantic Evaluations (SemEval-2007)*.