PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation

Stephen Tratz, Antonio Sanfilippo, Michelle Gregory, Alan Chappell, Christian Posse, Paul Whitney
Pacific Northwest National Laboratory
902 Battelle Blvd, PO Box 999
Richland, WA 99352, USA
{stephen.tratz, antonio.sanfilippo, michelle, alan.chappell, christian.posse, paul.whitney}@pnl.gov

Abstract

In this paper, we described the PNNL Word Sense Disambiguation system as applied to the English all-word task in SemEval 2007. We use a supervised learning approach, employing a large number of features and using Information Gain for dimension reduction. The rich feature set combined with a Maximum Entropy classifier produces results that are significantly better than baseline and are the highest F-score for the fined-grained English all-words subtask of SemEval.

1 Introduction

Accurate word sense disambiguation (WSD) can support many natural language processing and knowledge management tasks. The main goal of the PNNL WSD system is to support Semantic Web applications, such as semantic-driven search and navigation, through a reliable mapping of words in naturally occurring text to ontological classes. As described in Sanfilippo et al. (2006), this goal is achieved by defining a WordNet-based (Fellbaum, 1998) ontology that offers a manageable set of concept classes, provides an extensive characterization of concept class in terms of lexical instances, and integrates an automated class recognition algorithm. We found that the same features that are useful for predicting word classes are also useful in distinguishing individual word senses.

Our main objective in this paper is to predict individual word senses using a large combination of features including contextual, semantic, and syntactic information. In our earlier paper (Sanfilippo et al., 2006), we reported that the PNNL WSD system exceeded the performance of the best performers for verbs in the SENSEVAL-3 English all-words task dataset. SemEval 2007 is our first opportunity to enter a word sense disambiguation competition.

2 Approach

While many unsupervised word sense disambiguation systems have been created, supervised systems have generally produced superior results (Snyder and Palmer, 2004; Mihalcea et al., 2004). Our system is based on a supervised WSD approach that uses a Maximum Entropy classifier to predict WordNet senses.

We use SemCor[1], OMWE 1.0 (Chklovski and Mihalcea, 2002), and example sentences in WordNet as the training corpus. We utilize the OpenNLP MaxEnt implementation[2] of the maximum entropy classification algorithm (Berger et al., 1996) to train classification models for each lemma and part-of-speech combination in the training corpus. These models are used to predict WordNet senses for words found in natural text. For lemma and part-of-speech combinations that are not present in the training corpus, the PNNL WSD system defaults to the most frequent WordNet sense.

2.1 Features

We use a rich set of features to predict individual word senses. A large number of features are extracted for each word sense instance in the training data. Following Dang & Palmer (2005) and Kohomban & Lee (2005), we use contextual, syntactic and semantic information to inform our word

[1] http://www.cs.unt.edu/~rada/downloads.html.
[2] http://maxent.sourceforge.net/.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 264–267,
Prague, June 2007. ©2007 Association for Computational Linguistics

sense disambiguation system. However, there are significant differences between the specific types of contextual, syntactic and semantic information we use in our system and those proposed by Dang & Palmer (2005) and Kohomban & Lee (2005). More specifically, we employ novel features and feature combinations, as described below.

- *Contextual information.* The contextual information we use includes the word under analysis plus the three tokens found on each side of the word, within sentence boundaries. Tokens include both words and punctuation.
- *Syntactic information*. We include grammatical dependencies (e.g. subject, object) and morpho-syntactic features such as part of speech, case, number and tense. We use the Connexor parser[3] (Tapanainen and Järvinen, 1997) to extract lemma information, parts of speech, syntactic dependencies, tense, case, and number information. A sample output of a Connexor parse is given in Table 1. Features are extracted for all tokens that are related through no more than 3 levels of dependency to the word to be disambiguated.
- *Semantic information*. The semantic information we incorporate includes named entity types (e.g. PERSON, LOCATION, ORGANIZATION) and hypernyms. We use OpenNLP[4] and LingPipe[5] to identify named entities, replacing the strings identified as named entities (e.g., Joe Smith) with the corresponding entity type (PERSON). We also substitute personal pronouns that unambiguously denote people with the entity type PERSON. Numbers in the text are replaced with type label NUMBER. Hypernyms are retrieved from WordNet and added to the feature set for all noun tokens selected by the contextual and syntactic rules. In contrast to Dang & Palmer (2005), we only include the hypernyms of the most frequent sense, and we include the entire hypernym chain (e.g. motor, machine, device, instrumentality, artifact, object, whole, entity).

To address feature extraction processes specific to noun and verbs, we add the following conditions.

- *Syntactic information for verbs*. If the verb does not have a subject, the subject of the closest ancestor verb in the syntax tree is used instead.
- *Syntactic information for nouns*. The first verb ancestor in the syntax tree is also used to generate features.
- *Semantic information for nouns*. A feature indicating whether a token is capitalized for each of the tokens used to generate features.

A sample of the resulting feature vectors that are used by the PNNL word sense disambiguation system is presented in Table 2.

ID	Word	Lemma	Grammatical Dependencies	Morphosyntactic Features
1	the	the	det:>2	@DN> %>N DET
2	engine	engine	subj:>3	@SUBJ %NH N NOM SG
3	throbbe	throb	main:>0	@+FMAINV %VA V PAST
4	d	into	goa:>3	@ADVL %EH PREP
5	into	life	pcomp:>4	@<P %NH N NOM SG
6	life	.		
.				

Table 1. Connexor sample output for the sentence *"The engine throbbed into life"*.

the	pre:2:the, pre:2:pos:DET, det:the, det:pos:DET, hassubj:det:
engine	pre:1:instrumentality, pre:1:object, pre:1:artifact, pre:1:device, pre:1:engine, pre:1:motor, pre:1:whole, pre:1:entity, pre:1:machine, pre:1:pos:N, pre:1:case:NOM, pre:1:num:SG,subj:instrumentality,subj:object, subj:artifact, subj:device, subj:engine, subj:motor, subj:whole, subj:entity, subj:machine, subj:pos:N, hassubj:, subj:case:NOM, subj:num:SG,
throbbed	haspre:1:,haspre:2:,haspost:1:, haspost:2:, haspost:3:, self:throb, self:pos:V, main:,throbbed, self:tense:PAST
into	post:1:into, post:1:pos:PREP, goa:into, goa:pos:PREP,
life	post:2:life, post:2:state, post:2:being, post:2:pos:N, post:2:case:NOM, post:2:num:SG, hasgoa:, pcomp:life, pcomp:state, pcomp:being, pcomp:pos:N, hasgoa:pcomp:, goa:pcomp:case:NOM, goa:pcomp:num:SG
.	post:3:.

Table 2. Feature vector for *throbbed* in the sentence *"The engine throbbed into life"*.

As the example in Table 2 indicates, the combination of contextual, syntactic, and semantic information types results in a large number of features. Inspection of the training data reveals that some features may be more important than others in establishing word sense assignment for each choice of word lemma. We use a feature selection proce-

[3] http://www.connexor.com/.

[4] http://opennlp.sourceforge.nt/.

[5] http://www.alias-i.com/lingpipe/.

dure to reduce the full set of features to the feature subset that is most relevant to word sense assignment for each lemma. This practice improves the efficiency of our word sense disambiguation algorithm. The feature selection procedure we adopted consists of scoring each potential feature according to a particular feature selection metric, and then taking the best *k* features.

We choose Information Gain as our feature selection metric. Information Gain measures the decrease in entropy when the feature is given versus when it is absent. Yang and Pederson (1997) report that Information Gain outperformed other feature selection approaches in their multi-class benchmarks, and Foreman (2003) showed that it performed amongst the best for his 2-class problems.

3 Evaluation

To evaluate our approach and feature set, we ran our model on the SENSEVAL-3 English all-words task test data. Using data provided by the SENSEVAL website[6], we were able to compare our results for verbs to the top performers on verbs alone. Upali S. Kohomban and Wee Sun Lee provided us with the results file for the Simil-Prime system (Kohomban and Lee, 2005). As reported in Sanfilippo et al. (2006) and shown in table 3, our results for verbs rival those of top performers. We had a significant improvement (p-value<0.05) over the baseline of 52.9%, a marginal improvement over the second best performer (SenseLearner) (Mihalcea and Faruque, 2004), and we were as good as the top performer (GAMBL) (Decadt et al., 2004).[7]

System	Precision	Fraction of Recall
Our system	61%	22%
GAMBL	59.0%	21.3%
SenseLearner	56.1%	20.2%
Baseline	52.9%	19.1%

Table 3. Results for verb sense disambiguation on SENSEVAL-3 data, adapted from Sanfilippo et al. (2006).

Since then, we have expanded our evaluation to all parts of speech. Table 4 provides the evaluation of our system as compared to the three top performers on the SENSEVAL-3 data and the baseline. The baseline of 0.631 F-score[8] was computed using the most frequent WordNet sense. The PNNL WSD system performs significantly better than the baseline (p-value<0.05) and rivals the top performers. The performance of the PNNL WSD system relative to the other three systems and the baseline remains unchanged when the unknown sense answers (denoted by a 'U') are excluded from the evaluation.

System	Precision	Recall
PNNL	0.670	0.670
Simil-Prime	0.661	0.663
GAMBL	0.652	0.652
SenseLearner	0.646	0.646
Baseline	0.631	0.631

Table 4. SENSEVAL-3 English all-words**.**

System	Recall	Precision
PNNL	0.669	0.671
GAMBL	0.651	0.651
Simil-Prime	0.644	0.657
SenseLearner	0.642	0.651
Baseline	0.631	0.631

Table 5. SENSEVAL-3 English all-words, No "U"**.**

4 Experimental results on SemEval all-words subtask

This was our first opportunity to test our model in a WSD competition. For this competition, we focused our efforts on the fine-grained English all-words task because our system was set up to perform fine-grained WordNet sense prediction. We are pleased that our system achieved the highest score for this subtask. Our results for the SemEval dataset as compared to baseline are reported in Table 6. The PNNL WSD system did not assign the unknown sense, 'U', to any word instances in the SemEval dataset.

[6] http://www.senseval.org/.

[7] The 2% improvement in precision which our system showed as compared to GAMBL was not statistically significant (p=0.21).

[8] This baseline is slightly higher than that reported by others (Snyder and Palmer 2004).

System	F-score
PNNL	0.591
Baseline	0.514
p-value	<0.01

Table 6. SemEval Results**.**

5 Discussion

Although these results are promising, there is still much work to be done. For example, we need to investigate the contribution of each feature to the overall performance of the system in terms of precision and recall. Such a feature sensitivity analysis will provide us with a better understanding of how the algorithm can be further improved and/or made more efficient by leaving out features whose contribution is negligible**.**

Another important point to make is that, while our system shows the best precision/recall results overall, we can only claim statistical relevance with reference to the baseline and results worse than baseline. The size of the SemEval data set (N=465) is too small to establish whether the difference in precision/recall results with the other top systems is statistically significant.

Acknowledgements

We would like to thank Upali S. Kohomban and Wee Sun Lee for providing us with their SENSEVAL-3 English all-words task results file for SimilPrime. Many thanks also to Patrick Paulson, Bob Baddeley, Ryan Hohimer, and Amanda White for their help in developing the word class disambiguation system on which the work presented in this paper is based.

References

Berger, A., S. Della Pietra and V. Della Pietra (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics,* volume 22, number 1, pages 39-71.

Chklovski, T. and R. Mihalcea (2002) Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*.

Dang, H. T. and M. Palmer (2005) The Role of Semantic Roles in Disambiguating Verb Senses. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor MI, June 26-28, 2005.

Decadt, B., V. Hoste, W. Daelemans and A. Van den Bosch (2004) GAMBL, genetic algorithm optimization of memory-based WSD. *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Fellbaum, C., editor. (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Foreman, G. (2003) An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3, pages 1289-1305.

Kohomban, U. and W. Lee (2005) Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, Ann Arbor, MI.

Mihalcea, R., T. Chklovski, and A. Kilgarriff (2004) The SENSEVAL-3 English Lexical Sample Task, *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelonna, Span.

Mihalcea, R. and E. Faruque (2004) SenseLearner: Minimally supervised word sense disambiguation for all words in open text. *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Sanfilippo, A., S. Tratz, M. Gregory, A. Chappell, P. Whitney, C. Posse, P. Paulson, B. Baddeley, R. Hohimer, A. White (2006) Automating Ontological Annotation with WordNet. *Proceedings to the Third International WordNet Conference*, Jan 22-26, Jeju Island, Korea.

Snyder, B. and M. Palmer. 2004. The English All-Words Task. *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Tapanainen, P. and Timo Järvinen (1997) A nonprojective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington D.C. Association for Computational Linguistics.

Yang, Y. and J. O. Pedersen (1997) A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning* (ICML), pages 412-420, 1997.