CITYU-HIF: WSD with Human-Informed Feature Preference

Oi Yee Kwong
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlolivia@cityu.edu.hk

Abstract

This paper describes our word sense disambiguation (WSD) system participating in the SemEval-2007 tasks. The core system is a fully supervised system based on a Naïve Bayes classifier using multiple knowledge sources. Toward a larger goal of incorporating the intrinsic nature of individual target words in disambiguation, thus introducing a cognitive element in automatic WSD, we tried to fine-tune the results obtained from the core system with human-informed feature preference, and compared it with automatic feature selection as commonly practised in statistical WSD. Despite the insignificant improvement observed in this preliminary attempt, more systematic analysis remains to be done for a cognitively plausible account of the factors underlying the lexical sensitivity of WSD, which would inform and enhance the development of WSD systems in return.

1 Introduction

In recent years, many research teams all over the world have gained rich experience on word sense disambiguation (WSD) from the shared tasks of the SENSEVAL workshops. The need for multiple knowledge sources has become a golden rule, and the "lexical sensitivity" once remarked by Resnik and Yarowsky (1997) is addressed by various means in statistical classifiers, such as learning an optimal combination of the various knowledge sources for individual target words (e.g. Mihalcea, 2002; Escudero *et al.*, 2004). Another common practice is to use an ensemble of classifiers. As pointed out by Mihalcea *et al.* (2004), among the participating systems in the SENSEVAL-3 English lexical sample task, "several of the top performance systems are based on combination of multiple classifiers, which shows once again that voting scheme that combine several learning algorithms outperform the accuracy of individual classifiers". However, the advancement in WSD is rarely accompanied by any extensive account on the cognitive aspects of the task or qualitative analysis of the relation between the disambiguation results and the nature of individual target words underlying the apparent lexical sensitivity of the task.

Given that humans apparently use different strategies in making sense of words, it might be beneficial to have such cognitive aspects, including the type and strength of various kinds of semantic association, realised in NLP systems explicitly. Thus in addition to an optimal combination of classifiers alone, to better understand the contribution of different information types for different types of target words, it is important to look at WSD in relation to the very intrinsic nature of individual target words, which could comprise many factors such as frequency, abstractness, sense relatedness and parts-of-speech (POS). We thus use the concept *Information Susceptibility* (Kwong, 2005) to refer to the relationship between the intrinsic features of a target word and its senses, and the effectiveness of various lexical information to characterise them.

Our current participation in SemEval-2007 is thus intended as a means toward a larger goal, i.e., to incorporate a cognitive element into automatic WSD systems. In particular, we tried to fine-tune the results obtained from the core system with human-informed feature preference.

In Section 2, we will briefly describe the implementation of our disambiguation system and the features used. In Section 3 we will discuss the

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 109–112,
Prague, June 2007.

human input on the target nature and the informativeness of various features. The experiments and results are presented in Section 4, followed by a conclusion in Section 5.

2 System Description

2.1 Core Classifier

The core system is a fully supervised one based on a Naïve Bayes classifier. We made use of the Weka API (Witten and Frank, 2005) in our implementation. According to Yarowsky and Radu (2002), Bayesian classifiers belong to one of the aggregative models which depend heavily on the multiple reinforcing feature clues obtainable from wide context. Thus we use all features described in Section 2.2 below for our core system.

2.2 Knowledge Sources

Only the training data provided by the task organisers was used to train the system. We used four major types of contextual features, which could be classified into *Target features*, *Local features*, *Topical features* and *Syntactic features*, as described in Table 1. All features were converted to binary features.

2.3 Feature Selection

On top of the core system, we tested two value-added steps to accommodate for the lexical sensitivity of WSD. One is automatic feature selection (AFS), for which we used CfsSubsetEval (correlation-based feature selection) as implemented in Weka, based on the training samples of each target word. The other is human-informed feature preference (HIF), for which we ran another Naïve Bayes classifier in parallel with a feature subset deemed informative by human judges to fine-tune the disambiguation results obtained from the core system (see Sections 3 and 4 below).

3 Intrinsic Nature of Target Words

Leacock *et al.* (1998), for example, observed that "the benefits of adding topical to local context alone depend on syntactic category as well as on the characteristics of the individual word". In other words, some target words happen to be more "topical" than others and might therefore be more susceptible to topical contextual features during disambiguation. Others, however, might only be optimally disambiguated with other types of information.

Target Features	
W_0	Word form of the target word
P_0	POS of the target word
Local Features	
P_{-2} P_{-1} P_{+1} P_{+2}	POS of words at fixed positions from the target word, including the first and second word on its left and the first and second word on its right
W_{-2} W_{-1} W_{+1} W_{+2}	Word forms of the words at fixed positions from the target word, including the first and second word on its left and the first and second word on its right
Topical Features	
$W_{-10} \ldots W_{+10}$	Content words appearing within the window of ten words on each side of the target word
Syntactic Features	
$P_{-2}\,P_0$ $P_{-1}\,P_0$ $P_0\,P_{+1}$ $P_0\,P_{+2}$	POS bigrams composed of the target word and its neighbouring words, the non-immediate $P_{-2}\,P_0$ and $P_0\,P_{+2}$ are included to accommodate for some flexibility
$P_{-2}\,P_{-1}\,P_0$ $P_0\,P_{+1}\,P_{+2}$	POS trigrams composed of the target word and its neighbouring words

Table 1 Features Used in the Naïve Bayes Classifer

While statistical WSD has more or less reached its ceiling, it is assumed that a more thorough understanding of the effectiveness of different types of lexical information for characterising a word sense and distinguishing it from others should be able to further inform and enhance the development of WSD systems. To this end, three undergraduate linguistics students in the City University of Hong Kong were asked to go through the training data for the Chinese lexical sample task in SENSEVAL-3 and that for the multilingual Chinese-English lexical sample task (Task 5) in SemEval-2007. For each sense of a given target word, they were asked to rate the *difficulty*, *abstractness*, and *topicality* of the sense on a 3-point scale. At the same time, they were asked to indi-

cate the type of information, among local POS, local words, and contextual words (i.e. the topical features in Table 1), which they reckon to be most useful for disambiguating a given sample of the target word.[1]

While the information collected from the human judges is pending in-depth analysis, the feature preference indicated by them was used to fine-tune the results obtained from our core system. During disambiguation, we run two Naïve Bayes classifiers in parallel, the core one on all features above, and the other only on the type of information deemed most useful by two or more of the human judges, and use the latter to adjust the results from the former, as further discussed in Section 4.2.

4 Experiment and Results

4.1 Datasets

We participated in the Multilingual Chinese-English Lexical Sample Task (Task 5) and the English Lexical Sample Task via English-Chinese Parallel Text (Task 11).

Task 5 consists of 40 Chinese target words, 19 nouns and 21 verbs. The number of senses for the target words ranges from 2 to 8, with an average of 3. There are altogether 2,680 training samples, i.e. on average about 22 for each sense. A total of 935 testing instances were to be tagged, i.e. on average about 23 for each target word. The data were from People's Daily. The sense tags are given in the form of their English translations in the Chinese Semantic Dictionary developed by the Institute of Computational Linguistics of Peking University. The task organiser has provided the data with word segmentation and POS for each segmented word.

Task 11 consists of 40 English target words, including 20 nouns and 20 adjectives. The average number of training samples for each sense is about 42. The number of senses for the target words ranges from 2 to 6, with an average of 3.125. The average number of testing samples for each target word is 68. The data were gathered from word-aligned English-Chinese parallel texts.

In addition, we also used the SENSEVAL-3 Chinese lexical sample data during evaluation, which contains 20 target words.

[1] To simplify the task for the human judges, we did not distinguish between fixed-position local POS and n-gram syntactic features, and only used the former.

4.2 Evaluation

For Task 5, we made use of the segmentation and POS information provided by the task organiser. For Task 11, we first ran the data through the Brill tagger (Brill, 1994) to obtain the POS, from which we then extracted the feature values.

On top of the core system, we also tested two value-added conditions, namely automatic feature selection (AFS) and human-informed feature preference (HIF). For the latter, we run a separate Naïve Bayes classifier in parallel to the core system, using the knowledge source deemed most useful for a given target word by two or more human judges. When the probability of the best guess from the core classifier is under a certain threshold, the best guess from the other is used instead. For the current experiment, the probability of the best guess from the core classifier must at least double that for the next best guess.

For evaluation, we ran a 10-fold cross validation on the SemEval-2007 Task 5 training data, with the core system and AFS. In addition, we tested with the Senseval-3 Chinese lexical sample data. We trained the classifier with the Senseval-3 training data, with the core classifier, AFS, and HIF. The results are discussed below.

4.3 Results

Table 2 shows the evaluation results of the various conditions described above.

Condition	Ave. Precision
SemEval-2007 training data (10-fold CV)	
Core classifier	77.33%
Core classifier + AFS	85.51%
Senseval-3 testing data	
Core classifier	60.2%
Core classifier + AFS	61.7%
Core classifier + HIF	60.7%

Table 2 Evaluation Results

Apparently, and as known and expected, feature selection is useful for choosing an optimal set of features for each target word. How this compares and works together with human intuition and the nature of the individual target words and senses is what we would like to further investigate. In the above experiment, fine-tuning with human-

informed feature preference did not improve the performance as significantly as one would like to see, and the effect varied with individual target words. One possibility is that Naïve Bayes classifiers favour aggregative features, so it might not be most appropriate to do the fine-tuning with a separate classifier. Rather, we could explore the feasibility of adjusting the weights of individual features based on the feature preference.

Our next step is to perform in-depth and systematic analysis on the difficulty, abstractness and topicality of the target words and senses, with the information gathered from the human judges and the confusion matrices generated from the experiment, in association with psychological evidence like semantic activation and the organisation of the mental lexicon (e.g. Kwong, 2007).

4.4 Official Scores in SemEval-2007

The official scores for our system are shown in Table 3.

Task	System	MicroAvg	MacroAvg	Rank
5	HIF	71.0%	74.9%	3 / 6
11	AFS	75.3%[2]	-	3 / 3

Table 3 Official Scores for CITYU in SemEval-2007

Our scores are comparable to the state-of-the-art results. Although the HIF step did not increase the performance significantly, in view of the limitation of state-of-the-art statistical WSD systems, every minor improvement counts. It therefore remains for us to further investigate the cognitive aspects of WSD in relation to target nature and have them systematically realised in WSD systems.

5 Conclusion

In this paper, we have described our system participating in the SemEval-2007 multilingual Chinese-English lexical sample task and English lexical sample task via English-Chinese parallel text. Toward a larger goal of supplementing statistical methods with some cognitive elements of WSD, more systematic analysis of the intrinsic nature of target words underlying the lexical sensitivity of WSD is underway.

[2] A post-hoc analysis reveals a technical problem for six of the target words in Task 11 (educational.a, change.n, future.n, interest.n, need.n, program.n) which were not properly processed by the system in one of the steps, and the most frequent sense was used by default. Ignoring these cases, a precision of 78.3% was obtained using the task organiser's key and scoring program.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1508/06H).

References

Brill, E. (1994) Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pp.722-727.

Escudero, G., Màrquez, L. and Rigau, G. (2004) TALP System for the English Lexical Sample Task. In *Proceedings of SENSEVAL-3*, Barcelona, Spain.

Kwong, O.Y. (2005) Word Sense Classification Based on Information Susceptibility. In A. Lenci, S. Montemagni and V. Pirrelli (Eds.), *Acquisition and Representation of Word Meaning*. Linguistica Computazionale, pp.89-115.

Kwong, O.Y. (2007) Sense Abstractness, Semantic Activation and Word Sense Disambiguation: Implications from Word Association Norms. To appear in *Proceedings of NLPCS-2007*, Madeira, Portugal.

Leacock, C., Miller, G.A. and Chodorow, M. (1998) Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics, 24(1)*:147-166.

Mihalcea, R.F. (2002) Word sense disambiguation with pattern learning and automatic feature selection. *Natural Language Engineering, 8(4)*:343-358.

Mihalcea, R., Chklovski, T. and Kilgarriff, A. (2004) The SENSEVAL-3 English Lexical Sample Task. In *Proceedings of SENSEVAL-3*, Barcelona, Spain.

Resnik, P. and Yarowsky, D. (1997) A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of SIGLEX'97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., pp.79-86.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Yarowsky, D. and Radu, F. (2002) Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering, 8(4)*:293-310.