SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

Joining forces to resolve lexical ambiguity: East meets West in Barcelona

Richard WICENTOWSKI*, Grace NGAI†[1], Dekai WU‡[2]
Marine CARPUAT‡, Emily THOMFORDE*, Adrian PACKEL*

*Swarthmore College
Swarthmore, PA
USA

† Dept. of Computing
† HK Polytechnic University
Hong Kong

‡ HKUST, Dept of Computer Science
Human Language Technology Center
Hong Kong

richardw@cs.swarthmore.edu, csgngai@polyu.edu.hk, dekai@cs.ust.hk
marine@cs.ust.hk, ethomfo1@cs.swarthmore.edu, packel@cs.swarthmore.edu

Abstract

This paper describes the component models and combination model built as a joint effort between Swarthmore College, Hong Kong PolyU, and HKUST. Though other models described elsewhere contributed to the final combination model, this paper focuses solely on the joint contributions to the ”Swat-HK” effort.

1 Introduction

This paper describes the two joint component models of the Swat-HK systems entered into four of the word sense disambiguation lexical sample tasks in Senseval-3: Basque, Catalan, Italian and Romanian, as well as a combination model for each language. The feature engineering (and construction of three other component models which are described in (Wicentowski et al., 2004)) was performed at Swarthmore College, while the Hong Kong team constructed two component models based on well-known machine learning algorithms. The combination model, which was constructed at Swarthmore, uses voting to combine all five models.

2 Experimental Features

A full description of the experimental features for all four tasks can be found in the report submitted by the Swarthmore College Senseval team (Wicentowski et al., 2004). Briefly, the systems used lexical and syntactic features in the context of the target word:

- The “bag of words (and lemmas)” in the context of the ambiguous word.
- Bigrams and trigrams of words (and lemmas, part-of-speech tags, and, for Basque, case information) surrounding the ambiguous word.
- The topic (or code) of the document containing the current instance of the word was extracted. (Basque and Catalan only.)

These features have been shown to be effective in previous WSD research. Since our systems were all supervised, all the data used was provided by the Senseval organizers; no additional (unlabeled) data was included.

3 Methodology

The systems that were constructed by this team included two component models: a boosting model and a maximum entropy model as well as a combination system. The component models were also used in other Senseval-3 tasks: Semantic Role Labeling (Ngai et al., 2004) and the lexical sample tasks for Chinese and English, as well as the Multilingual task (Carpuat et al., 2004).

To perform parameter tuning for the two component models, 20% of the samples from the training set were held out into a validation set. Since we did not expect the senses of different words to share any information, the training data was partitioned by the ambiguous word in question. A model was then trained for each ambiguous word type. In total, we had 40 models for Basque, 27 models for Catalan, 45 models for Italian and 39 models for Romanian.

3.1 Boosting

Boosting is a powerful machine learning algorithm which has been shown to achieve good results on a variety of NLP problems. One known property of boosting is its ability to handle large numbers of features. For this reason, we felt that it would be well suited to the WSD task, which is known to be highly lexicalized with a large number of possible word types.

Our system was constructed around the Boostexter software (Schapire and Singer, 2000), which implements boosting on top of decision stumps (deci-

[1] The author would like to thank the Hong Kong Polytechnic University for supporting this research in part through research grants A-PE37 and 4-Z03S.

[2] The author would like to thank the Hong Kong Research Grants Council (RGC) for supporting this research in part through research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09.

sion trees of one level), and was originally designed for text classification.

Tuning a boosting system mainly lies in modifying the number of iterations, or the number of base models it would learn. Larger number of iterations contribute to the boosting model's power. However, they also make it more prone to overfitting and increase the training time. The latter, a simple disadvantage in another problem, becomes a real issue for Senseval, since large numbers of models (one for each word type) need to be trained in a short period of time.

Since the available features differed from language to language, the optimal number of iterations also varied. Table 1 shows the performance of the model on the validation set with respect to the number of iterations per language.

	Accuracy Number of iterations		
Language	500	1000	2000
Basque	66.12%	67.07%	67.08%
Catalan	84.77%	84.89%	85.02%
Italian	51.11%	50.93%	
Romanian	64.68%	64.52%	

Table 1: Boosting models on the validation sets.

The final systems for the languages used 2000 iterations for Basque and Catalan and 500 iterations for Italian and Romanian. The test set results are shown in Table 4

3.2 Maximum Entropy

The other individual system was based on the maximum entropy model, another machine learning algorithm which has been successfully applied to many NLP problems. Our system was implemented on top of the YASMET package (Och, 2002).

Due to lack of time, we did not manage to fine-tune the maximum entropy model. The YASMET package does provide a number of easily variable parameters, but we were only able to try varying the feature selection count threshold and the smoothing parameter, and only on the Basque data.

Experimentally, however, smoothing did not seem to make a difference. The only change in performance was caused by varying the feature selection count threshold, which controls the number of times a feature has to be seen in the training set in order to be considered. Table 2 shows the performances of the system on the Basque validation set, with count thresholds of 0, 1 and 2.

	Threshold		
	0	1	2
Accuracy	55.62%	66.13%	65.68%

Table 2: Maximum Entropy Models on Basque validation set.

Since word sense disambiguation is known to be a highly lexicalized task involving many feature values and sparse data, it is not too surprising that setting a low threshold of 1 proves to be the most effective. The final system kept this threshold, smoothing was not done and the GIS iterations allowed to proceed until it converged on its own. These parameters were used for all four languages.

The maximum entropy model was not entered into the competition as an official contestant; however, it did participate in the combined system.

3.3 Combined System

Ensemble methods have been widely studied in NLP research, and it is well-known that a set of systems will often combine to produce better results than those achieved by the best individual system alone. The final system contributed by the Swarthmore-Hong Kong team was such an ensemble. In addition to the boosting and maximum entropy models mentioned earlier, three other models were included: a nearest-neighbor clustering model, a decision list, and a Naïve Bayes model. The five models were then combined by a simple weighted majority vote, with an ad-hoc weight of 1.1 given to the boosting and decision lists systems, and 1.0 otherwise, with ties broken arbitrarily.

Due to an unfortunate error with the input data of the voting algorithm (Wicentowski et al., 2004), the official submitted results for the combined system were poorer than they should have been. Table 3 compares the official (submitted) results to the corrected results on the test set. The decrease in performance caused by the error ranged from 0.9% to 3.3%.

Language	official	corrected	net gain
Basque	67.0%	67.9%	0.9%
Catalan	79.5%	80.4%	0.9%
Italian	51.4%	54.7%	3.3%
Romanian	72.4%	73.3%	0.9%

Table 3: Ensemble system results on the test set. Both official and corrected results are included.

System		Acc. (%)
Description	Name	
Basque		
Boosting	**basque-swat_hk-bo**	71.1
Combined	**swat-hk-basque**	67.0 (67.9)
NNC		66.0
DL		64.6
Maxent		62.1
NB		60.4
Baseline		55.8
Catalan		
Boosting	**catalan-swat_hk-bo**	79.6
DL		80.6
Combined	**swat-hk-catalan**	79.5 (80.4)
NNC		77.5
NB		71.3
Maxent		70.9
Baseline		66.4
Italian		
Combined	**swat-hk-italian**	51.4 (54.7)
DL		50.3
Boosting	**italian-swat_hk**	48.3
Maxent		46.9
NNC		44.9
NB		42.1
Baseline		23.7
Romanian		
Boosting	**romanian-swat_hk-bo**	72.7
Combined	**swat-hk-romanian**	72.4 (73.3)
DL		70.9
NNC		67.9
Maxent		66.5
NB		62.8
Baseline		58.4

Table 4: Test set results on 4 languages. Official contestants are in bold; corrected voting results are in parentheses. Key: NB: Naïve Bayes, NNC: Nearest-Neighbor Clustering, DL: Decision List

4 Test Set Results

Final results from all the systems are shown in Table 4. As a reference, the results of a simple baseline system which assigns the most frequent sense as seen in the training set is also provided.

Due to the error in the voting system, the official results for the combination system were lower than they should have been — as a result, boosting was *officially* the top ranked system for 3 of the 4 languages. With the *corrected* results, however, the combined system outperforms the individual models, as expected. The only exception is Basque, where the booster had an exceptionally strong performance. This is probably due to the fact that Basque has a much richer feature set than the other languages, which boosting was better able to take advantage of.

The poor performance of the maximum entropy model was also unexpected at first; however, it is perhaps not too surprising, given the lack of time spent on fine-tuning the model. As a result, most of the parameters were left at their default values.

One thing worth noting is the fact that the systems were combined as “closed systems” — i.e. all that was known about them was the output result, and nothing else. The result was that no confidence measures from the boosting and maximum entropy could be used in the combined system. It is likely that the performance could have been further improved if more information had been available.

5 Conclusions and Discussion

This paper describes the “Swat-HK” systems which were the result of collaborative work between Swarthmore College, Hong Kong Polytechnic University and HKUST. Several base systems were constructed on the same feature set, and a weighed majority voting system was used to combine the results. The individual systems all achieve good results, easily beating the baseline. As expected, the combined system outperforms the best individual system for the majority of the tasks.

References

Marine Carpuat, Weifeng Su, and Dekai Wu. 2004. Augmenting Ensemble Classification for Word Sense Disambiguation with a Kernel PCA Model. In *Proceedings of Senseval-3*, Barcelona.

Grace Ngai, Dekai Wu, Marine Carpuat, Chi-Shing Wang, and Chi-Yung Wang. 2004. Semantic Role Labeling with Boosting, SVMs, Maximum Entropy, SNOW, and Decision Lists. In *Proceedings of Senseval-3*, Barcelona.

Franz Josef Och. 2002. Yet Another Small Maxent Toolkit: Yasmet. http://www-i6.informatik.rwth-aachen.de/Colleagues/och.

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Richard Wicentowski, Emily Thomforde, and Adrian Packel. 2004. The Swarthmore College Senseval-3 system. In *Proceedings of Senseval-3*, Barcelona.