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

The Swarthmore College SENSEVAL3 System

Richard Wicentowski, Emily Thomforde and Adrian Packel
Computer Science Department
Swarthmore College
Swarthmore, Pennsylvania, 19081, USA
{richardw,ethomfo1,packel}@cs.swarthmore.edu

Abstract

This paper presents the Swarthmore College word-sense disambiguation system which was designed for the 2004 SENSEVAL3 competition. Our system participated in five tasks: the lexical sample tasks in Basque, Catalan, Italian, Romanian, and Spanish. For each task, a suite of supervised algorithms were combined using voting to form the final system.

1 Introduction

The Swarthmore College system consisted of three supervised classifiers which were used to perform lexical ambiguity resolution in five languages. A nearest-neighbor clustering classifier, a naïve Bayes classifier, and a decision list classifier were each trained on several permutations of the extracted feature set, then the answers were joined using voting. The training data was limited to the labeled data provided by the organizers; no outside or unlabeled data was used.

The systems presented in this paper were developed by undergraduates as part of a class project at Swarthmore College.

2 Features

Each of the supervised algorithms made use of the same set of features, extracted from only the labeled data provided to us by the task organizers. We used no unlabeled data. We used the tagged and lemmatized data to extract the following features, which were the only features used in our system:

- Bag-of-words and bag-of-lemmas
- Bigrams and trigrams of words, lemmas, part-of-speech, and case (Basque-only) around the target word
- Topic or code (Basque, Catalan and Spanish)

In order to prevent individual features from dominating any individual system, we used up to eight permutations of the above mentioned features (depending on the language) for each of our classifiers.

Catalan and Spanish provided fine-grained part-of-speech tags which we felt would lead to sparse data problems. To reduce this problem, for some feature sets we made the part-of-speech tags more coarse by simplifying the tags to include only the first or first two letters of the tag.

3 Systems

The following systems were used to complete the Basque, Catalan, Italian and Romanian lexical sample tasks. The Spanish lexical sample task was completed before the other four tasks were begun and used only a subset of the systems presented below. Full details on the systems and methods used for the Spanish lexical sample task can be found in Section 7.3.

See Section 4 for details on the classifier combination, and Section 5.2 for information about our use of bagging.

3.1 Cosine-based Clustering

The first system developed was a nearest-neighbor clustering method which used the cosine similarity of feature vectors as the distance metric. A centroid was created for each attested sense in the training data, and each test sample was assigned to a cluster based on its similarity to the existing centroid. Centroids were not recalculated after each added test instance.

3.2 Naïve Bayes

The second system used was a naïve Bayes classifier where the similarity between an instance, I, and a sense class, S_j, is defined as:

$$Sim(I, S_j) = P(I, S_j) = P(S_j)P(I|S_j)$$

We then choose the sense class, S_j, which maximized the similarity function above, making standard independence assumptions.

3.3 Decision List

The final system was a decision list classifier that found the log-likelihoods of the correspondence be-

tween features and senses, using plus-one smoothing (Yarowsky, 1994). The features were ordered from most to least indicative to form the decision list. A separate decision list was constructed for each set of lexical samples in the training data. For each test instance, the first matching feature found in the associated decision list was used to determine the classification of the instance. Instances which failed to match any rule in the decision list were labeled with the most frequent sense, as calculated from the training data.

4 Classifier Combination

Due to time constraints, we were unable to get cross-validation results for all of the systems we created, and therefore all of the final classifier combination was done using simple majority vote, breaking ties arbitrarily. To reach a consensus vote, we combined the multiple decision list systems, which had been run on each of the different subsets of extracted features, into a single system. We then did the same for the clustering system and the naïve Bayes system, yielding a total of three new systems. These three systems were then voted together to form the final system. The two-tiered voting was performed to ensure equal voting in the case of our joint work (Wicentowski et al., 2004) where the five systems that needed to be combined were run on different numbers of feature subsets.

4.1 Combination Errors

There were two mistakes we made when voting our systems together. We caught one mistake after the submission deadline but before notification of results; the other we realized only while evaluating our systems after receiving our results. For this reason, there are three sets of results that we will report here:

- *Official results*: The results from our submission to SENSEVAL3.
- *Unofficial results*: Includes a bug-fix found *before* notification of competition results.
- *Tie-breaking results*: Includes a bug-fix found *after* notification of results.

In doing the evaluation of our system for this paper, we will use the unofficial results[1]. Because of the nature of the bug-fix, evaluating our system based on the official results will yield less informative results than an evaluation of results after fixing the error. Since these unofficial results were obtained before notification of results from the competition organizers, we believe this to be a fair comparison.

4.1.1 Over-weighting part-of-speech n-grams

The bug which yielded our unofficial results occurred when we combined the multiple decision list systems into a single decision list system (and similarly for the multiple clustering and naïve Bayes systems). As discussed in Section 2, we experimented with forming partial labels for the part-of-speech tags to reduce the sparse-data issues: using the full part-of-speech tag, using only the first letter of the tag, and using the first two letters of the tag. However, in the final combination, we ended up including all three methods in the voting, instead of including only one. Obviously, these three classifiers, based solely on part-of-speech n-grams around the target word, had a high rate of agreement and were therefore over-weighted in the final voting. Our systems underperformed where they should have, with the surprising exception of Catalan, which performed better with the mistake than without it. Table 1 compares our official results with our unofficial results.

Language	official	unofficial	change
Basque	64.6%	66.6%	2.0%
Catalan	79.7%	79.5%	(-0.2%)
Italian	46.5%	49.6%	3.1%
Romanian	70.1%	71.3%	1.2%
Spanish	79.5%	–	–

Table 1: Final results, officially and unofficially, from making a bug-fix before notification of results, but after the submission deadline.

4.1.2 Voting without weighting

Our classifier combination used a non-informed method for breaking ties: whichever sense had the first hash code (as determined by Perl's hash function) was chosen. Our inability to complete cross-validation experiments led us to not favor any one classifier over another. Performance would have been improved by using an ad-hoc weighting scheme which took into account the following intuitions:

- Initial experiments indicated that the instances of the classifiers with access to the full set of features would outperform the instances running on limited subsets of the features.
- Empirical evidence suggested that the decision list classifier was the best, the clustering

[1] As mentioned previously, Spanish is a special case and we will report only our official results.

method a strong second, and the naïve Bayes method a distant third.

In fairness, we did not discover this mistake until we were preparing this paper, only after receiving notification of our results. While we report our revised results, we make no further comparisons based on these results. In addition, we ran no extra experiments to determine the weighting scheme listed below, we simply used our intuition based on our earlier experimentation as noted above. These intuitions were not always correct, as indicated in Table 5 and Table 6.

Using very simple ad-hoc weights which back up these intuitions, we changed our classifier combination system to break ties according to the following scheme: In the first tier of voting, we fractionally increased the weight given to the classifiers run on the full-feature set: instead of each system receiving exactly one vote, we gave those systems an extra $\frac{1}{10}$th of a vote. In the second tier of voting, we made the same fractional increase to the weight given to the decision list classifier. Use of this tie breaking scheme increases our results impressively, as shown below in Table 2.

Language	official	tie-breaking	net gain
Basque	64.6%	68.2%	3.6%
Catalan	79.7%	81.0%	1.3%
Italian	46.5%	52.4%	5.9%
Romanian	70.1%	73.2%	3.1%

Table 2: Using simple tie-breakers in voting. The second column also includes the bug fix described in §4.1.1. Note that the tie-breaking error was found after notification of our final results.

5 Additional features

5.1 Collocational Senses

In the Basque and Romanian tasks, senses could be labeled either as numbered alternatives or as a collocational sense. For example, the Basque word `astun` could be labeled with the collocational sense `pisu_astun`.

From the SENSEVAL2 English lexical-sample task, we found there were 175 words labeled with a collocational sense. A lemmatized form of the collocation was found in 96.6% of these when considering a ± 2-word window around the target. To take advantage of this expected behavior in Basque and Romanian, we labeled a target word with a collocational sense if we found the lemmatized collocation in a ± 2-word window. In Romanian, many collocations contained prepositions or other stop-words; therefore, we labeled a target word with the collocational sense only if a non-stop-word from the collocation was found in the ± 2-word window. Overall, as shown in Table 3, this decision proved to be reasonably effective.

Language	Correct	Answered	Precision
Romanian	161	190	84.7%
Basque	70	79	88.6%

Table 3: Precision on likely collocational senses.

Complementary to this issue, a sampling of the same English data indicated that if a target word was part of a previously seen collocation, it was highly unlikely that this word would *not* be tagged with the collocational sense. Therefore, we expected it would be advantageous if we could remove the collocational senses from the training data to prevent target words which were not part of collocations from being tagged as such. Based on cross-validated results, we found that this was worthwhile for Basque, but not for Romanian, where there were many examples of a target word being tagged as collocational sense without the collocation being present.

5.2 Bagging

For the decision list and clustering systems, we used bagging (Breiman, 1996) to train on five randomly sampled instances of the training data which were combined using a simple majority vote. We limited ourselves to five samples due to time limitations imposed by the competition. We found a consistent, but minimal, improvement for each of the four tasks due to our use of bagging, as shown below in Table 4.

Language	no bagging	bagging	net gain
Basque	66.0%	66.6%	0.6%
Catalan	79.4%	79.5%	0.1%
Italian	48.6%	49.6%	1.0%
Romanian	70.9%	71.3%	0.4%

Table 4: Overall impact of using bagging.

6 Evaluation

As previously discussed, we used a combination of three supervised classifiers, each run on a different subset of the features. Here we report the performance of each of the individual classifiers, as well as the features we found to be most indicative of the correct sense.

6.1 Indicative features

As discussed in Section 2, we did not use any external data for the lexical sample tasks, but we did try to use all of the features that were available in the training and test sets. In order to show the effectiveness of each of the features we used, we present the following sample taken from running our decision list system in Basque, Catalan, Italian and Romanian using only one feature at a time.

Basque	Feature	Catalan
55.8%	*mfs baseline*	55.0%
64.6%	*all features*	80.6%
52.9%	case n-grams	-
54.2%	simplified pos n-grams	74.9%
59.2%	topic/code tag	70.5%
54.1%	part-of-speech n-grams	77.5%
61.7%	docsrc tag	69.7%
61.4%	bag of words	78.7%
61.3%	bag of 'forms'	79.7%
62.6%	bag of lemmas	78.7%
65.0%	word n-grams	**81.7%**
66.1%	lemma n-grams	**81.7%**

Italian	Feature	Romanian
27.7%	*mfs baseline*	58.4%
50.3%	*all features*	70.9%
38.4%	simplified pos n-grams	63.7%
38.6%	part-of-speech n-grams	64.2%
41.0%	bag of words	64.7%
41.1%	bag of 'forms'	-
41.1%	bag of lemmas	64.8%
44.4%	word n-grams	**70.0%**
46.5%	lemma n-grams	69.4%

Table 5: Accuracy of the decision list system using each of the available features individually. All of the above features, except 'docsrc' were used in the final system. The features are ordered from least to most informative across the four languages.

With the exception of Romanian, the bigrams and trigrams comprised of the lemmas were the most informative single feature for the decision list system. Surprisingly, in both Catalan and Basque, the decision list system trained only on lemma n-grams outperformed decision list system which used all of the features.

Because the lemmas were so important, we suspect that omitting them in future data sets will favor those systems which can incorporate accurate lemmatizers. Since real world applications will require such lemmatizers, we are in favor of omitting these in future competitions.

6.2 Classifiers

As shown in Table 6, the decision list system was the best single system; however, the nearest-neighbor clustering system outperformed decision lists in Basque. Each of the supervised systems is compared against the baseline most-frequent-sense classifier (as computed from the training data).

Language	MFS	NB	NNC	DL
Basque	55.8%	60.4%	**66.0%**	64.6%
Catalan	55.0%	71.3%	77.5%	**80.6%**
Italian	27.7%	42.1%	44.9%	**50.3%**
Romanian	58.4%	62.8%	67.9%	**70.9%**

Table 6: Accuracies for each of the classifiers: Most Frequent Sense, Naïve Bayes, Nearest-Neighbor Clustering, and Decision Lists.

7 Task-specific Details

7.1 Basque

The Basque data contained the largest number of available features, but in places, the features were incomplete (case markers) or required additional steps to extract. Most notably, though lemmas were provided, the target word was not indicated in either the training or test data; therefore, we performed some simple pre-processing of the Basque data to isolate the target lemma in the training and test data As is shown in Table 5, these lemma n-grams around the target word were the most indicative features for our decision list system.

7.2 Romanian

The Romanian data also provided a large number of available features, however some pre-processing was necessary to change the format of the supplied part-of-speech tagged data into the format supplied by the other tasks.

7.3 Spanish

We were required to submit our results for the Spanish lexical sample task before we had completed writing our system, so the submission includes only two classifiers, a naïve Bayes classifier and a decision list classifier. We ran our decision list on seven permutations of the feature set, and the naïve Bayes on two permutations, for a total of nine systems. These nine systems were joined using a majority-voting scheme. Relative performance on this task is expected to be below that of other tasks.

8 Collaborative Work

This paper refers only to the entries completed exclusively by the Swarthmore College team and

discusses the entries submitted under the label "Swat". The "Swat-HK" and "Swat_HK-Bo" entries were submitted by Swarthmore College in collaboration with a joint team from Hong Kong Polytechnic University and Hong Kong University of Science and Technology. For these entries, Swarthmore College provided the data, with all of the features described, to the Hong Kong team. Their team then sent us back two sets of results: the output of their maximum entropy system and their boosting system. These two results were then combined with the three systems written by Swarthmore College. Details on this joint effort can be found in (Wicentowski et al., 2004).

In addition, the decision list system described here was used in the Semantic Role Labeling task submitted by (Ngai et al., 2004).

9 Acknowledgments

The authors thank the following Swarthmore College students for their assistance and guidance: Ben Mitchell '05, Charles Bell '06, Lisa Spitalewitz '06, and Michael Stone '07. Their efforts as part of the Fall 2003 "Information Retrieval and Natural Language Processing" class laid the foundation for our successful entry into the SENSEVAL3 competition.

In addition, the authors express their gratitude to Grace Ngai, Dekai Wu, and all the members of their joint team, for asking us to participate in their Semantic Role Labeling system.

Finally, the authors thank the organizers, especially Rada Mihalcea, for their support of our participation.

References

L. Breiman. 1996. Bagging predictors. *Machine Learning*, 24:123–140.

G. Ngai, D. Wu, M. Carpuat, C.S. Wang, and C.Y. Wang. 2004. Semantic Role Labeling with Boosting, SVMs, Maximum Entropy, SNOW, and Decision Lists. In *Proceedings of SENSEVAL-3*, Barcelona.

R. Wicentowski, G. Ngai, E. Thomforde, A. Packel, D. Wu, and M. Carpuat. 2004. Joining forces to resolve lexical ambiguity: East meets West in Barcelona. In *Proceedings of SENSEVAL-3*, Barcelona.

D. Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.