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 "Meaning" System on the English Allwords Task

**L. Villarejo**†, **L. Màrquez**†, **E. Agirre**‡, **D. Martínez**‡, , **B. Magnini***,
**C. Strapparava***, **D. McCarthy****, **A. Montoyo***** , and **A. Suárez*****

†TALP Research Center, Universitat Politècnica de Catalunya, {luisv,lluism}@lsi.upc.es
‡IXA Group, University of the Basque Country, {eneko,davidm}@si.ehu.es
*ITC-irst (Istituto per la Ricerca Scientifica e Tecnologica), {magnini,strappa}@itc.it
**University of Sussex, dianam@sussex.ac.uk
***LSI, University of Alicante, montoyo@dlsi.ua.es,armando.suarez@ua.es

## 1 Introduction

The "Meaning" system has been developed within the framework of the Meaning European research project[1]. It is a combined system, which integrates several supervised machine learning word sense disambiguation modules, and several knowledge–based (unsupervised) modules. See section 2 for details. The supervised modules have been trained exclusively on the SemCor corpus, while the unsupervised modules use WordNet-based lexico–semantic resources integrated in the Multilingual Central Repository (MCR) of the Meaning project (Atserias et al., 2004).

The architecture of the system is quite simple. Raw text is passed through a pipeline of linguistic processors (tokenizers, POS tagging, named entity extraction, and parsing) and then a Feature Extraction module codifies examples with features extracted from the linguistic annotation and MCR. The supervised modules have priority over the unsupervised and they are combined using a weighted voting scheme. For the words lacking training examples, the unsupervised modules are applied in a cascade sorted by decreasing precision. The tuning of the combination setting has been performed on the Senseval-2 allwords corpus.

Several research groups have been providers of resources and tools, namely: IXA group from the University of the Basque Country, ITC-irst ("Istituto per la Ricerca Scientifica e Tecnologica"), University of Sussex (UoS), University of Alicante (UoA), and TALP research center at the Technical University of Catalonia. The integration was carried out by the TALP group.

## 2 The WSD Modules

We have used up to seven supervised learning systems and five unsupervised WSD modules. Some of them have also been applied individually to the Senseval-3 lexical sample and allwords tasks.

- **Naive Bayes** (NB) is the well–known Bayesian algorithm that classifies an example by choosing the class that maximizes the product, over all features, of the conditional probability of the class given the feature. The provider of this module is IXA. Conditional probabilities were smoothed by Laplace correction.
- **Decision List** (DL) are lists of weighted classification rules involving the evaluation of one single feature. At classification time, the algorithm applies the rule with the highest weight that matches the test example (Yarowsky, 1994). The provider is IXA and they also applied smoothing to generate more robust decision lists.
- In the **Vector Space Model** method (cosVSM), each example is treated as a binary-valued feature vector. For each sense, one centroid vector is obtained from training. Centroids are compared with the vectors representing test examples, using the cosine similarity function, and the closest centroid is used to classify the example. No smoothing is required for this method provided by IXA.
- **Support Vector Machines** (SVM) find the hyperplane (in a high dimensional feature space) that separates with maximal distance the positive examples from the negatives, i.e., the *maximal margin* hyperplane. Providers are TALP ($SVM_1$) and IXA ($SVM_2$) groups. Both used the freely available implementation by (Joachims, 1999), linear kernels, and one–vs–all binarization, but with different parameter tuning and feature filtering.
- **Maximum Entropy** (ME) are exponential conditional models parametrized by a flexible set of features. When training, an iterative optimization procedure finds the probability distribution over feature coefficients that maximizes

[1] Meaning, Developing Multilingual Web-scale Language Technologies (European Project IST-2001-34460): http://www.lsi.upc.es/∼nlp/meaning/meaning.html.

the entropy on the training data. This system is provided by UoA.

- **AdaBoost** (AB) is a method for learning an ensemble of *weak* classifiers and combine them into a *strong* global classification rule. We have used the implementation described in (Schapire and Singer, 1999) with decision trees of depth fixed to 3. The provider of this system is TALP.

- **Domain Driven Disambiguation** (DDD) is an unsupervised method that makes use of domain information in order to solve lexical ambiguity. The disambiguation of a word in its context is mainly a process of comparison between the domain of the context and the domains of the word's senses (Magnini et al., 2002). ITC-irst provided two variants of the system $\text{DDD}_P$ and $\text{DDD}_{F_1}$, aiming at maximizing precision and $F_1$ score, respectively. The UoA group also provided another domain–based unsupervised classifier (DOM). Their approach exploits information contained in glosses of WordNet Domains and introduces a new lexical resource "Relevant Domains" obtained from Association Ratio over glosses of WordNet Domains.

- **Automatic Predominant Sense** (autoPS) provide an unsupervised first sense heuristic for the polysemous words in WordNet. This is produced by UoS automatically from the BNC (McCarthy et al., 2004). The method uses automatically acquired thesauruses for the main PoS categories. The nearest neighbors for each word are related to its WordNet senses using a WordNet similarity measure.

- We also used a **Most Frequent Sense tagger**, according to the WordNet ranking of senses (MFS).

## 3 Evaluation of Individual Modules

For simplicity, and also due to time constraints, the supervised modules were trained exclusively on the SemCor-1.6 corpus, intentionally avoiding the use of other sources of potential training examples, e.g, other corpora, WordNet examples and glosses, similar/substitutable examples extracted from the same Semcor-1.6, etc. An independent training set was generated for each polysemous word (of a certain part–of–speech) with 10 or more examples in the SemCor-1.6 corpus. This makes a total of 2,440 independent learning problems, on which all supervised WSD systems were trained.

The feature representation of the training examples was shared between all learning modules. It consists of a rich feature representation obtained using the Feature Extraction module of the TALP team in the Senseval-3 English lexical sample task. The feature set includes the classic window–based pattern features extracted from a local context and the "bag–of–words" type of features taken from a broader context. It also contains a set of features representing the syntactic relations involving the target word, and semantic features of the surrounding words extracted from the MCR of the Meaning project. See (Escudero et al., 2004) for more details on the set of features used.

The validation corpus for these classifiers was the Senseval-2 allwords dataset, which contains 2,473 target word occurrences. From those, 2,239 occurrences correspond to polysemous words. We will refer to this subcorpus as S2-pol. Only 1,254 words from S2-pol were actually covered by the classifiers trained on the SemCor-1.6 corpus. We will refer to this subset of words as the S2-pol-sup corpus. The conversion between WordNet-1.6 synsets (SemCor-1.6) and WordNet-1.7 (Senseval-2) was performed on the output of the classifiers by applying an automatically derived mapping provided by TALP[2].

Table 1 shows the results (precision and coverage) obtained by the individual supervised modules on the S2-pol-sup subcorpus, and by the unsupervised modules on the S2-pol subcorpus (i.e., we exclude from evaluation the monosemous words). Support Vector Machines and AdaBoost are the best performing methods, though all of them perform in a small accuracy range from 53.4% to 59.5%.

Regarding the unsupervised methods, DDD is clearly the best performing method, achieving a remarkable precision of 61.9% with the $\text{DDD}_P$ variant, at a cost of a lower coverage. The $\text{DDD}_{F_1}$ appears to be the best system for augmenting the coverage of the former. Note that the autoPS heuristic for ranking senses is a more precise estimator than the WordNet most–frequent–sense (MFS).

## 4 Integration of WSD modules

All the individual modules have to be integrated in order to construct a complete allwords WSD system. Following the architecture described in section 1, we decided to apply the unsupervised modules only to the subset of the corpus not covered by the training examples. Some efforts on applying the unsupervised modules jointly with the supervised failed at improving accuracy. See an example in table 3.

[2] http://www.lsi.upc.es/∼nlp/tools/mapping.html

| | supervised, S2-pol-sup corpus | | | | | | | unsupervised, S2-pol corpus | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $SVM_1$ | AB | cosVSM | $SVM_2$ | ME | NB | DL | $DDD_P$ | $DDD_{F_1}$ | autoPS | MFS | DOM |
| prec. | 59.5 | 59.1 | 57.8 | 57.1 | 56.3 | 54.6 | 53.4 | 61.9 | 50.2 | 45.2 | 32.5 | 23.8 |
| cov. | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 48.8 | 99.6 | 89.6 | 98.0 | 49.1 |

Table 1: Results of individual supervised and unsupervised WSD modules

As a first approach, we devised three baseline systems (Base-1, Base-2, and Base-3), which use the best modules available in both subsets. Base-1 applies the $SVM_1$ supervised method and the MFS for the non supervised part. Base-2 applies also the $SVM_1$ supervised method and the cascade $DDD_P$–MFS for the non supervised part (MFS is used in the cases in which $DDD_P$ abstains). Base-3 shares the same approach but uses a third unsupervised module: $DDD_P$–$DDD_{F_1}$–MFS.

The precision results of the baselines systems can be found in the right hand side of table 3. As it can be observed, the positive contribution of the $DDD_P$ module is very significant since Base-2 performs 2.2 points higher than Base-1. The addition of the third unsupervised module ($DDD_{F_1}$) makes Base-3 to gain 0.4 extra precision points.

As simple combination schemes we considered majority voting and weighted voting. More sophisticated combination schemes are difficult to tune due to the extreme data sparseness on the validation set. In the case of unsupervised systems, these combination schemes degraded accuracy because the least accurate systems perform much worse that the best ones. Thus, we simply decided to apply a cascade of unsupervised modules sorted by precision on the Senseval-2 corpus.

In the case of the supervised classifiers there is a chance of improving the global performance, since there are several modules performing almost as well as the best. Previous to the experiments, we calculated the agreement rates on the outputs of each pair of systems (low agreements increase the probability of uncorrelatedness between errors of different systems). We obtained an average agreement of 83.17%, with values between 64.7% (AB vs $SVM_2$) and 88.4% ($SVM_2$ vs cosVSM).

The ensembles were obtained by incrementally aggregating, to the best performing classifier, the classifiers from a list sorted by decreasing accuracy. The ranking of classifiers can be performed by evaluating them at different levels of granularity: from particular words to the overall accuracy on the whole validation set. The level of granularity defines a tradeoff between classifier specialization and risk of overfitting to the tuning corpus. We decided to take an intermediate level of granularity, and sorted the classifiers according to their performance on word sets based on the number of training examples available[3].

Table 2 contains the results of the ranking experiment, by considering five word-sets of increasing number of training examples: between 10 and 20, between 21 and 40, between 41 and 80, etc. At each cell, the accuracy value is accompanied by the relative position the system achieves in that particular subset. Note that the resulting orderings, though highly correlated, are quite different from the one derived from the overall results.

| | (10,20) | (21,40) | (41,80) | (81,160) | >160 |
|---|---|---|---|---|---|
| $SVM_1$ | 60.9-**1** | 59.1-**1** | 64.2-**2** | 61.1-**2** | 56.4-**1** |
| AB | 60.9-**1** | 56.6-**2** | 60.0-**7** | 64.7-**1** | 56.1-**2** |
| c-VSM | 59.9-**2** | 56.6-**2** | 62.6-**3** | 57.0-**4** | 55.8-**3** |
| $SVM_2$ | 50.8-**5** | 55.1-**4** | 61.6-**4** | 57.4-**3** | 53.1-**5** |
| ME | 56.7-**3** | 55.3-**3** | 65.3-**1** | 53.3-**5** | 53.8-**4** |
| NB | 59.9-**2** | 54.6-**5** | 61.1-**5** | 49.2-**6** | 51.5-**7** |
| DL | 56.4-**4** | 49.9-**6** | 60.5-**6** | 47.2-**7** | 52.5-**6** |

Table 2: Results on frequency–based word sets

Table 3 shows the precision results[4] of the Meaning system obtained on the whole Senseval-2 corpus by combining from 1 to 7 supervised classifiers according to the classifier orderings of table 2 for each subset of words. The unsupervised classifiers are all applied in a cascade sorted by precision. M-Vot stands for a majority voting scheme, while W-Vot refers to the weighted voting scheme. The weights for the classifiers are simply the accuracy values on the validation corpus. As an additional example, the column M-Vot+ shows the results of the voting scheme when the unsupervised $DDD_P$ module is also included in the ensemble. The table also includes the baseline results.

Unfortunately, the ensembles of classifiers did not provide significant improvements on the final precision. Only in the case of weighted voting a slight improvement is observed when adding up to 3 classifiers. From the fourth classifier performance also degrades. The addition of unsupervised systems to the supervised ensemble systematically degraded performance.

As a reference, the best result (67.5% precision

[3]One of the factors that differentiates between learning algorithms is the amount of training examples needed to learn.

[4]Coverage of the combined systems is 98% in all cases.

| | M-Vot | W-Vot | M-Vot+ | Base-1 | Base-2 | Base-3 |
|---|---|---|---|---|---|---|
| 1 | 67.3 | 67.3 | 66.4 | – | – | – |
| 2 | – | 67.4 | 66.3 | – | – | – |
| 3 | 67.2 | 67.5 | 67.1 | – | – | – |
| 4 | – | 67.1 | 66.9 | – | – | – |
| 5 | 66.5 | 66.5 | 66.7 | – | – | – |
| 6 | – | 66.3 | 66.3 | – | – | – |
| 7 | 65.7 | 65.9 | 66.0 | – | – | – |
| best | 67.3 | **67.5** | 67.1 | 64.8 | 67.0 | 67.4 |

Table 3: Results of the combination of systems

| System | prec. | recall | $F_1$ |
|---|---|---|---|
| Meaning-c | 61.1% | 61.0% | 61.05 |
| Meaning-wv | **62.5**% | **62.3**% | **62.40** |

Table 4: Results on the Senseval-3 test corpus

and 98.0% coverage) would have put our combined system in second place in the Senseval-2 allwords task.

## 5 Evaluation on the Senseval-3 Corpus

The Senseval-3 test set contains 2,081 target words, 1,851 of them polysemous. The subset covered by the SemCor-1.6 training contains 1,211 target words (65.42%, compared to the 56.0% of the Senseval-2 corpus). We submitted the outputs of two different configurations of the Meaning system: Meaning-c and Meaning-wv. These systems correspond to Base-3 and W-Vot (in the best configuration) from table 3, respectively. The results from the official evaluation are given in table 4. Again, we applied an automatic mapping from WordNet-1.6 to WordNet-1.7.1 synset labels. However, there are senses in 1.7.1 that do not exist in 1.6, thus our system simply cannot assign them.

It can be observed that, even though on the tuning corpus both variants obtained very similar precision (67.4 and 67.5), on the test set the weighted voting scheme is clearly better than the baseline system, probably due to the robustness achieved by the ensemble. The performance decrease observed on the test set with respect to the Senseval-2 corpus is very significant ($\sim$5 points). Given that the baseline system performs worse than the voted approach, it seems unlikely that there is overfitting during the ensemble tuning. However, we plan to repeat the tuning experiments directly on the Senseval-3 corpus to see if the same behavior and conclusions are observed. Probably, the decrease in performance is due to the differences between the training and test corpora. We intend to investigate the differences between SemCor-1.6, Senseval-2, and Senseval-3 corpora at different levels of linguistic information in order to check the appropriateness of using SemCor-1.6 as the main information source.

## 6 Acknowledgements

This research has been possible thanks to the support of European and Spanish research projects: IST-2001-34460 (Meaning), TIC2000-0335-C03-02 (Hermes). The authors would like to thank also Gerard Escudero for letting us use the Feature Extraction module and German Rigau for helpful suggestions and comments.

## References

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The Meaning multilingual central repository. In *Proceedings of the Second International WordNet Conference*.

G. Escudero, L. Màrquez, and G. Rigau. 2004. TALP system for the english lexical sample task. In *Proceedings of the Senseval-3 ACL Workshop*, Barcelona, Spain.

T. Joachims. 1999. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184. MIT Press, Cambridge, MA.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the Senseval-3 ACL Workshop*, Barcelona, Spain.

R. Schapire and Y. Singer. 1999. Improved boosting algorithms using confidence–rated predictions. *Machine Learning*, 37(3):297–336.

David 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, Las Cruces, NM.