TALP System for the English Lexical Sample Task

Gerard Escudero[‡], **Lluís Màrquez**[§] **and German Rigau**[†]
[‡]TALP Research Center. EUETIB. LSI. UPC. escudero@lsi.upc.es
[§]TALP Research Center. LSI. UPC. lluism@lsi.upc.es
[†] IXA Group. UPV/EHU. rigau@si.ehu.es

1 Introduction

This paper describes the TALP system on the English Lexical Sample task of the Senseval-3[1] event. The system is fully supervised and relies on a particular Machine Learning algorithm, namely Support Vector Machines. It does not use extra examples than those provided by Senseval-3 organisers, though it uses external tools and ontologies to extract part of the representation features.

Three main characteristics have to be pointed out from the system architecture. The first thing is the way in which the multiclass classification problem posed by WSD is addressed using the binary SVM classifiers. Two different approaches for binarizing multiclass problems have been tested: *one–vs–all* and *constraint classification*. In a cross-validation experimental setting the best strategy has been selected at word level. Section 2 is devoted to explain this issue in detail.

The second characteristic is the rich set of features used to represent training and test examples. *Topical* and *local context* features are used as usual, but also syntactic relations and semantic features indicating the predominant semantic classes in the example context are taken into account. A detailed description of the features is presented in section 3.

And finally, since each word represents a learning problem with different characteristics, a per–word feature selection has been applied. This tuning process is explained in detail in section 4.

The last two sections discuss the experimental results (section 5) and present the main conclusions of the work performed (section 6).

2 Learning Framework

The TALP system belongs to the supervised Machine Learning family. Its core algorithm is the Support Vector Machines (SVM) learning algorithm (Cristianini and Shawe-Taylor, 2000). Given a set of binary training examples, SVMs find the hyperplane that maximizes the margin in a high dimensional feature space (transformed from the input space through the use of a non-linear function, and implicitly managed by using the *kernel trick*), i.e., the hyperplane that separates with maximal distance the positive examples from the negatives. This learning bias has proven to be very effective for preventing overfitting and providing good generalisation. SVMs have been also widely used in NLP problems and applications.

One of the problems in using SVM for the WSD problem is how to binarize the multiclass classification problem. The two approximations tested in the TALP system are the usual *one–vs–all* and the recently introduced *constraint–classification* framework (Har-Peled et al., 2002).

In the one–vs–all approach, the problem is decomposed into as many binary problems as classes has the original problem, and one classifier is trained for each class trying to separate the examples of that class (positives) from the examples of all other classes (negatives). This method assumes the existence of a separator between each class and the set of all other classes. When classifying a new example, all binary classifiers predict a class and the one with highest confidence is selected (*winner–take–all* strategy).

2.1 Constraint Classification

Constraint classification (Har-Peled et al., 2002) is a learning framework that generalises many multiclass classification and ranking schemes. It consists of labelling each example with a set of binary constraints indicating the relative order between pairs of classes. For the WSD setting of Senseval-3, we have one constraint for each correct class (sense) with each incorrect class, indicating that the classifier to learn should give highest confidence to the correct classes than to the negatives. For instance, if we have 4 possible senses $\{1, 2, 3, 4\}$ and a training example with labels 2 and 3, the constraints corresponding to the example are $\{(2>1), (2>4), (3>1),$ and $(3>4)\}$. The aim of the methodology is to learn a classifier consistent with the partial order defined

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

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

by the constraints. Note that here we are not assuming that perfect separators can be constructed between each class and the set of all other classes. Instead, the binary decisions imposed are more conservative.

Using Kesler's construction for multiclass classification, each training example is expanded into a set of (longer) binary training examples. Finding a vector–based separator in this new training set is equivalent to find a separator for each of the binary constraints imposed by the problem. The construction is general, so we can use SVMs directly on the expanded training set to solve the multiclass problem. See (Har-Peled et al., 2002) for details.

3 Features

We have divided the features of the system in 4 categories: local, topical, knowledge-based and syntactic features. First section of table 1 shows the **local features**. The basic aim of these features is to modelize the information of the surrounding words of the target word. All these features are extracted from a ±3–word–window centred on the target word. The features also contain the position of all its components. To obtain Part–of–Speech and lemma for each word, we used FreeLing [2]. Most of these features have been doubled for lemma and word form.

Three types of **Topical features** are shown in the second section of table 1. Topical features try to obtain non–local information from the words of the context. For each type, two overlapping sets of redundant topical features are considered: one extracted from a ±10–word–window and another considering all the example.

The third section of table 1 presents the **knowledge–based features**. These features have been obtained using the knowledge contained into the Multilingual Central Repository (MCR) of the MEANING project[3] (Atserias et al., 2004). For each example, the feature extractor obtains, from each context, all nouns, all their synsets and their associated semantic information: Sumo labels, domain labels, WordNet Lexicographic Files, and EuroWordNet Top Ontology. We also assign to each label a weight which depends on the number of labels assigned to each noun and their relative frequencies in the whole WordNet. For each kind of semantic knowledge, summing up all these weights, the program finally selects those semantic labels with higher weights.

local feats.	
Feat.	**Description**
form	form of the target word
locat	all part–of–speech / forms / lemmas in the local context
coll	all collocations of two part–of–speech / forms / lemmas
coll2	all collocations of a form/lemma and a part–of–speech (and the reverse)
first	form/lemma of the first noun / verb / adjective / adverb to the left/right of the target word
topical feats.	
Feat.	**Description**
topic	bag of forms/lemmas
sbig	all form/lemma bigrams of the example
comb	forms/lemmas of consecutive (or not) pairs of the open–class–words in the example
knowledge-based feats.	
Feat.	**Description**
f_sumo	first sumo label
a_sumo	all sumo labels
f_semf	first wn semantic file label
a_semf	all wn semantic file labels
f_tonto	first ewn top ontology label
a_tonto	all ewn top ontology labels
f_magn	first domain label
a_magn	all domain labels
syntactical feats.	
Feat.	**Description**
tgt_mnp	syntactical relations of the target word from minipar
rels_mnp	all syntactical relations from minipar
yar_noun	NounModifier, ObjectTo, SubjectTo for nouns
yar_verb	Object, ObjectToPreposition, Preposition for verbs
yar_adjs	DominatingNoun for adjectives

Table 1: Feature Set

Finally, the last section of table 1 describes the **syntactic features** which contains features extracted using two different tools: Dekang Lin's Minipar[4] and Yarowsky's dependency pattern extractor.

It is worth noting that the set of features presented is highly redundant. Due to this fact, a feature selection process has been applied, which is detailed in the next section.

4 Experimental Setting

For each binarization approach, we performed a feature selection process consisting of two consecutive steps:

[2] http://www.lsi.upc.es/∼nlp/freeling

[3] http://www.lsi.upc.es/˜meaning

[4] http://www.cs.ualberta.ca/∼lindek/minipar.htm

- **POS feature selection**: Using the Senseval–2 corpus, an exhaustive selection of the best set of features for each particular Part–of–Speech was performed. These feature sets were taken as the initial sets in the feature selection process of Senseval-3.

- **Word feature selection**: We applied a forward(selection)–backward(deletion) two–step procedure to obtain the best feature selection per word. For each word, the process starts with the best feature set obtained in the previous step according to its Part–of–Speech. Now, during selection, we consider those features not selected during POS feature selection, adding all features which produce some improvement. During deletion, we consider only those features selected during POS feature selection, removing all features which produces some improvement. Although this addition–deletion procedure could be iterated until no further improvement is achieved, we only performed a unique iteration because of the computational overhead. One brief experiment (not reported here) for *one–vs–all* achieves an increase of 2.63% in accuracy for the first iteration and 0.52% for a second one. First iteration improves the accuracy of 53 words and the second improves only 15. Comparing the evolution of these 15 words, the increase in accuracy is of 2.06% for the first iteration and 1.68% for the second one. These results may suggest that accuracy could be increased by this iteration procedure.

The result of this process is the selection of the best binarization approach and the best feature set for each individual word.

Considering feature selection, we have inspected the selected attributes for all the words and we observed that among these attributes there are features of all four types. The most selected features are the local ones, and among them those of 'first noun/adjective on the left/right'; from topical features the most selected ones are the 'comb' and in a less measure the 'topic'; from the knowledge–based the most selected feature are those of 'sumo' and 'domains labels'; and from syntactical ones, those of 'Yarowsky's patterns'. All the features previously mentioned where selected at least for 50 of the 57 Senseval–3 words. Even so, it is useful the use of all features when a selection procedure is applied. These general features do not work fine for all words. Some words make use of the less selected features; that is, every word is a different problem.

Regarding the implementation details of the system, we used SVM^{light} (Joachims, 2002), a very robust and complete implementation of Support Vector Machines learning algorithms, which is freely available for research purposes[5]. A simple lineal kernel with a regularization C value of 0.1 was applied. This parameter was empirically decided on the basis of our previous experiments on the Senseval–2 corpus. Additionally, previous tests using non–linear kernels did not provide better results.

The selection of the best feature set and the binarization scheme per word described above, have been performed using a 5-fold cross validation procedure on the Senseval-3 training set. The five partitions of the training set were obtained maintaining, as much as possible, the initial distribution of examples per sense.

After several experiments considering the 'U' label as an additional regular class, we found that we obtained better results by simply ignoring it. Then, if a training example was tagged only with this label, it was removed from the training set. If the example was tagged with this label and others, the 'U' label was also removed from the learning example. In that way, the TALP system do not assigns 'U' labels to the test examples.

Due to lack of time, the TALP system presented at the competition time did not include a complete model selection for the constraint classification binarization setting. More precisely, 14 words were processed within the complete model selection framework, and 43 were adjusted with a fixed one–vs–all approach but a complete feature selection. After the competition was closed, we implemented the constraint classification setting more efficiently and we reprocessed again the data. Section 5 shows the results of both variants.

A rough estimation of the complete model selection time for both approaches is the following. The training spent about 12 hours (OVA setting) and 5 days (CC setting) to complete[6], suggesting that the main drawback of these approaches is the computational overhead. Fortunately, the process time can be easily reduced: the CC layer could be ported from Perl to C++ and the model selection could be easily parallelized (since the treatment of each word is independent).

5 Results

Table 2 shows the accuracy obtained on the training set and table 3 the results of our system (SE3,

[5] http://svmlight.joachims.org

[6] These figures were calculated using a 800 MHz Pentium III PC with 320 Mb of memory.

TALP), together with the most frequent sense baseline (mfs), the recall result of the best system in the task (best), and the recall median between all participant systems (avg). These last three figures were provided provided by the organizers of the task.

OVA(base) in table 2 stands for the results of the one–vs–all approach on the starting feature set (5–fold–cross validation on the training set). CC(base) refers to the constrain–classification setting on the starting feature set. OVA(best) and CC(best) mean one–vs–all and constraint–classification with their respective feature selection. Finally, SE3 stands for the system officially presented at competition time[7] and TALP stands for the complete architecture.

method	**accuracy**
OVA(base)	72 38%
CC(base)	72.28%
OVA(best)	75.27%
CC(best)	75.70%
SE3	75.62%
TALP	76.02%

Table 2: Overall results of all system variants on the training set

It can be observed that the feature selection process consistently improves the accuracy by around 3 points, both in OVA and CC binarization settings. Constraint–classification is slightly better than one–vs–all approach when feature selection is performed, though this improvement is not consistent along all individual words (detailed results omitted) neither statistically significant (z–test with 0.95 confidence level). Finally, the combined binarization–feature selection further increases the accuracy in half a point (again this difference is not statistically significant).

measure	**mfs**	**avg**	**best**	**SE3**	**TALP**
fine	55.2	65.1	72.9	71.3	71.6
coarse	64.5	73.7	79.5	78.2	78.2

Table 3: Overall results on the Senseval-3 test set

However, when testing the complete architecture on the official test set, we obtained an accuracy decrease of more than 4 points. It remains to be analyzed if this difference is due to a possible overfitting to the training corpus during model selection, or simply is due to the differences between training and test corpora. Even so, the TALP system achieves a very good performance, since there is a difference of only 1.3 points in fine and coarse recall respect to the best system of the English lexical sample task of Senseval–3.

6 Conclusions

Regarding supervised Word Sense Disambiguation, each word can be considered as a different classification problem. This implies that each word has different feature models to describe its senses.

We have proposed and tested a supervised system in which the examples are represented through a very rich and redundant set of features (using the information content coherently integrated within the Multilingual Central Repository of the MEANING project), and which performs a specialized selection of features and binarization process for each word.

7 Acknowledgments

This research has been partially funded by the European Commission (Meaning Project, IST-2001-34460), and by the Spanish Research Department (Hermes Project: TIC2000-0335-C03-02).

References

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen 2004. The MEANING Multilingual Central Repository. In *Proceedings of the Second International WordNet Conference*.

N. Cristianini and J. Shawe-Taylor 2000. *An Introduction to Support Vector Machines*. Cambridge University Press.

T. Joachims 2002. *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer.

S. Har-Peled and D. Roth and D. Zimak 2002. Constraint Classification for Multiclass Classification and Ranking. In *Proceedings of the 15th Workshop on Neural Information Processing Systems*.

[7] Only 14 words were processed with the full architecture.