Low–cost Named Entity Classification for Catalan: Exploiting Multilingual Resources and Unlabeled Data

Lluís Màrquez, **Adrià de Gispert**, **Xavier Carreras**, and **Lluís Padró**
TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona, 1–3, E-08034, Barcelona
{lluism,agispert,carreras,padro}@talp.upc.es

Abstract

This work studies Named Entity Classification (NEC) for Catalan without making use of large annotated resources of this language. Two views are explored and compared, namely exploiting solely the Catalan resources, and a direct training of bilingual classification models (Spanish and Catalan), given that a large collection of annotated examples is available for Spanish. The empirical results obtained on real data point out that multilingual models clearly outperform monolingual ones, and that the resulting Catalan NEC models are easier to improve by bootstrapping on unlabelled data.

1 Introduction

There is a wide consensus about that Named Entity Recognition and Classification (NERC) are Natural Language Processing tasks which may improve the performance of many applications, such as Information Extraction, Machine Translation, Question Answering, Topic Detection and Tracking, etc. Thus, interest on detecting and classifying those units in a text has kept on growing during the last years.

Previous work in this topic is mainly framed in the *Message Understanding Conferences* (MUC), devoted to Information Extraction, which included a NERC competition task. More recent approaches can be found in the proceedings of the shared task at the 2002 and 2003 editions of the *Conference on Natural Language Learning* (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), where several machine–learning (ML) systems were compared at the NERC task for several languages.

One remarkable aspect of most widely used ML algorithms is that they are *supervised*, that is, they require a set of labelled data to be trained on. This may cause a severe bottleneck when such data is not available or is expensive to obtain, which is usually the case for minority languages with few pre–existing linguistic resources and/or limited funding possibilities. This is one of the main causes for the recent growing interest on developing *language–independent* NERC systems, which may be trained from small training sets by taking advantage of *unlabelled examples* (Collins and Singer, 1999; Abney, 2002), and which are easy to *adapt* to changing domains (being all these aspects closely related).

This work focuses on exploring the construction of a low–cost Named Entity classification (NEC) module for Catalan without making use of large/expensive resources of the language. In doing so, the paper first explores the training of classification models by using only Catalan resources and then proposes a training scheme, in which a Catalan/Spanish bilingual classifier is trained directly from a training set including examples of the two languages. In both cases, the bootstrapping of the resulting classifiers is also explored by using a large unannotated Catalan corpus. The strategy used for training the bilingual NE classification models has been also applied with good results to NE recognition in (Carreras et al., 2003), a work that can be considered complementary to this one.

When considering the training of bilingual models, we take advantage of the facts that Spanish and Catalan are two Romance languages with similar syntactic structure, and that —since Spanish and Catalan social and cultural environments greatly overlap— many Named Entities appear in both languages corpora. Relying on this structural and content similarity, we will build our Catalan NE classifier on the following assumptions: (a) Named Entities appear in the same contexts in both languages, and (b) Named Entities are composed by similar patterns in both languages.

The paper presents an extensive experimental evaluation, giving strong evidence about the advantage of using multilingual models for training on a language with scarce resources. Additionally, the Catalan NEC models resulting from the bilingual training are easier to improve by bootstrapping on unlabelled data.

The paper is organized as follows. Section 2 describes the Catalan and Spanish resources available and the feature codification of examples. Section 3 briefly describes the learning algorithms used to train the classifiers. Section 4 is devoted to the learning of NEC modules using only Catalan resources, while section 5 presents and evaluates the bilingual approach. Finally, the main conclusions of the work are summarized in section 6.

2 Setting

2.1 Corpus and data resources

The experimentation of this work has been carried on two corpora, one for each language. Both corpora consist of sentences extracted from news articles of the year 2,000. The Catalan data, extracted from the Catalan edition of the daily newspaper *El Periódico de Catalunya*, has been randomly divided into three sets: a training set (to train a system) and a test set (to perform evaluation) for manual annotation, and a remaining set left as unlabelled. The Spanish data corresponds to the CoNLL 2002 Shared Task Spanish data, the original source being the EFE Spanish Newswire Agency. The training set has been used to improve classification for Catalan, whereas the test set has been used to evaluate the bilingual classifier. The original development set has not been used. Table 1 shows the number of sentences, words and Named Entities in each set. Although a large amount of Catalan unlabelled NEs is available, it must be observed that these are automatically recognised with a 91.5% accurate NER module, introducing a certain error that might undermine bootstrapping results.

lang.	set	#sent.	#words	#NEs
es	train.	8,322	264,715	18,797
es	test	1,516	51,533	3,558
ca	train.	817	23,177	1,232
ca	test	844	23,595	1,338
ca	unlab.	83,725	2,201,712	75,038*

Table 1: Sizes of Spanish and Catalan data sets

Considered classes include MUC categories PER LOC and ORG, plus a fourth category MIS, including named entities such as documents, measures and taxes, sport competitions, titles of art works and others. For Catalan, we find 33.0% of PER, 17.1% of LOC, 43.5% of ORG and 6.4% of MIS out of the 2,570 manually annotated NEs, whereas for Spanish, out of the 22,355 labelled NEs, 22.6% are PER, 26.8% are LOC, 39.4% are ORG and the remaining 11.2% are MIS.

Additionally, we used a Spanish 7,427 trigger–word list typically accompanying persons, organizations, locations, etc., and an 11,951 entry gazetteer containing geographical and person names. These lists have been semi-automatically extracted from lexical resources and manually enriched afterwards. They have been used in some previous works allowing significant improvements for the Spanish NERC task (Carreras et al., 2002; Carreras et al., 2003).

Trigger–words are annotated with the corresponding Spanish synsets in the EuroWordNet lexical knowledge base. Since there are translation links among Spanish and Catalan (and other languages) for the majority of these words, an equivalent version of the trigger–word list for Catalan has been automatically derived. In this work, we consider the gazetteer as a language independent resource and is indistinctly used for training Catalan and Spanish models.

2.2 Feature codification

The features that characterise the NE examples are defined in a window W anchored at a word w, representing its local context used by a classifier to make a decision. In the window, each word around w is codified with a set of primitive features, requiring no linguistic pre–processing, together with its relative position to w. Each primitive feature with each relative position and each possible value forms a final binary feature for the classifier (e.g., "the **word_form** at **position(-2)** is **street**"). The kind of information coded in these features may be grouped in the following kinds:

- **Lexical**: Word forms and their position in the window (e.g., $W(3)$="bank"), as well as word forms appearing in the named entity under consideration, independent from their position.
- **Orthographic**: Word properties regarding how it is capitalised (*initial-caps*, *all-caps*), the kind of characters that form the word (*contains-digits*, *all-digits*, *alphanumeric*, *roman-number*), the presence of punctuation marks (*contains-dots*, *contains-hyphen*, *acronym*), single character patterns (*lonely-initial*, *punctuation-mark*, *single-char*), or the membership of the word to a predefined class (*functional-word*[1]) or pattern (*URL*).
- **Affixes**: The prefixes and suffixes up to 4 characters of the NE being classified and its internal components.
- **Word Type Patterns**: Type pattern of consecutive words in the context. The type of a word is either *functional* (`f`), *capitalised* (`C`), *lowercased* (`l`), *punctuation mark* (`.`), *quote* (`'`) or *other* (`x`).
- **Bag-of-Words**: Form of the words in the window, without considering positions (e.g., "bank"$\in W$).
- **Trigger Words**: Triggering properties of window words, using an external list to determine whether a word may trigger a certain Named Entity (NE) class (e.g., "president" may trigger class PER). Also context patterns to the left of the NE are considered, where each word is marked with its triggering properties, or with a functional–word tag, if appropriate (e.g., the phrase "the president of United Nations" produces pattern `f_ORG_f` for the NE "United_Nations", assuming that "president" is listed as a possible trigger for ORG).
- **Gazetteer Features**: Gazetteer information for window words. A gazetteer entry consists of a set of possible NE categories.
- Additionally, binary features encoding the length in words of the NE being classified.

All features are computed for a {-3,+3} window around the NE being classified, except for the **Bag-of-Words**, for which a {-5,+5} window is used.

3 Learning Algorithms

As previously said, we compare two learning approaches when learning from Catalan examples: supervised (using the AdaBoost algorithm), and unsupervised (using the Greedy Agreement Algorithm). Both of them are briefly described below.

3.1 Supervised Learning

We use the multilabel multiclass AdaBoost.MH algorithm (with confidence–rated predictions) for learning the classification models. The idea of this algorithm is to learn an accurate strong classifier by linearly combining, in a weighted voting scheme, many simple and moderately–accurate base classifiers or rules. Each base rule is sequentially learned by presenting the base learning algorithm a weighting over the examples (denoting importance of examples), which is dynamically adjusted depending on the behaviour of the previously learned rules. We refer the reader to (Schapire and Singer, 1999) for details about the general algorithm, and to (Schapire, 2002) for successful applications to many areas, including several NLP tasks. Additionally, a NERC system based on the AdaBoost algorithm obtained the best results in the CoNLL'02 Shared Task competition (Carreras et al., 2002).

In our setting, the boosting algorithm combines several small fixed–depth decision trees. Each branch of a tree is, in fact, a conjunction of binary features, allowing the strong boosting classifier to work with complex and expressive rules.

3.2 Unsupervised Learning

We have implemented the Greedy Agreement Algorithm (Abney, 2002) which, based on two independent views of the data, is able to learn two binary

[1]Functional words are determiners and prepositions which typically appear inside NEs.

classifiers from a set of hand-typed seed rules. Each classifier is a majority vote of several atomic rules, which abstains when the voting ends in a tie. The atomic rules are just mappings of a single feature into a class (e.g., if suffix "lez" then PER). When learning, the atomic rule that maximally reduces the disagreement on unlabelled data between both classifiers is added to one of the classifiers, and the process is repeated alternating the classifiers. See (Abney, 2002) for a formal proof that this algorithm tends to gradually reduce the classification error given the adequate seed rules.

For its extreme simplicity and potentially good results, this algorithm is very appealing for the NEC task. In fact, results are reported to be competitive against more sophisticated methods (Co-DL, Co-Boost, etc.) for this specific task in (Abney, 2002).

Three important questions arise from the algorithm. First, what features compose each view. Second, how seed rules should be selected or whether this selection strongly affects the final classifiers. Third, how the algorithm, presented in (Abney, 2002) for binary classification, can be extended to a multiclass problem.

In order to answer these questions and gain some knowledge on how the algorithm works empirically, we performed initial experiments on the big labelled portion of the Spanish data.

When it comes to view selection, we tried two alternatives. The first, suggested in (Collins and Singer, 1999; Abney, 2002), divides into one view capturing internal features of the NE, and the other capturing features of its left-right contexts (hereafter referred to as Greedy Agreement pure, or GA_p). Since the contextual view turned out to be quite limited in performance, we interchanged some feature groups between the views. Specifically, we moved the Lexical features independent of their position to the contextual view, and the the Bag-of-Words features to the internal one (we will refer to this division as Greedy Agreement mixed, or GA_m). The latter, containing redundant and conditionally dependent features, yielded slightly better results in terms of precision–coverage trade–off.

As for seed rules selection, we have tried two different strategies. On the one hand, blindly choosing as many atomic rules as possible that decide at least in 98% of the cases for a class in a small validation set of labelled data, and on the other, manually selecting from these atomic rules only those that might be valid still for a bigger data set. This second approach proved empirically better, as it provided a much higher starting point in the test set (in terms of precision), whereas a just slightly lower coverage value, presenting a better learning curve.

Finally, we have approached the multiclass setting by a *one–vs–all* binarization, that is, dividing the classification problem into four binary decisions (one per class), and combining the resultant rules. Several techniques to combine them have been tested, from making a prediction only when one classifier assigns positive for the given instance and all other classifiers assign negative (very high precision, low coverage), to much unrestrictive approaches, such as combining all votes from each classifier (lower precision, higher coverage). Results proved that the best approach is to sum all votes from all non-abstaining binary classifiers, where a vote of a concrete classifier for the negative class is converted to one vote for each of the other classes.

The best results obtained in terms of coverage/precision and evaluated over the whole set of training data (and thus more significant than over a small test set) are 80.7/84.9. These results are comparable to the ones presented in (Abney, 2002), taking into account, apart from the language change, that we have introduced a fourth class to be treated the same as the other three. Results when using Catalan data are presented in section 4.

4 Using only Catalan resources

This section describes the results obtained by using only the Catalan resources and comparing the fully unsupervised Greedy Agreement algorithm with the AdaBoost supervised learning algorithm.

4.1 Unsupervised vs. supervised learning

In this experiment, we used the Catalan training set for extracting seed rules of the GA algorithm and to train an AdaBoost classifier. The whole unlabelled Catalan corpus was used for bootstrapping the GA algorithm. All the results were computed over the Catalan test set.

Figure 1 shows a precision–coverage plot of AdaBoost (noted as CA, for CAtalan training) and

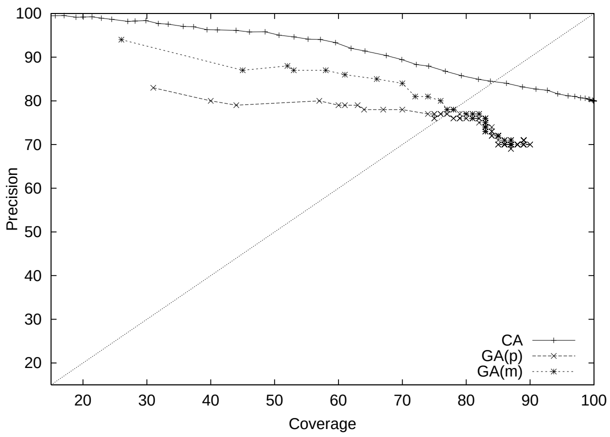

Figure 1: Precision–coverage plot of GA_p, GA_m, and CA models trained on Catalan resources

the Greedy Agreement algorithm for the two views selections (noted GA_p and GA_m, respectively). The curve for CA has been computed by varying a confidence threshold: CA abstains when the highest prediction of AdaBoost is lower than this threshold.

On the one hand, it can be seen that GA_m is more precise than GA_p for low values of coverage but their asymptotic behaviour is quite similar. By stopping at the best point in the validation set, the Greedy Agreement algorithm (GA_m) achieves a precision of 76.53% with a coverage of 83.62% on the test set. On the other hand, the AdaBoost classifier clearly outperforms both GA models at all levels of coverage, indicating that the supervised training is preferable even when using really small training sets (an accuracy around 70% is obtained by training AdaBoost only with the 20% of the learning examples, i.e., 270 examples).

The first three rows of table 2 contain the accuracy of these systems (i.e., precision when coverage is 100%), detailed at the NE type level (best results printed in boldface)[2]. The fourth row (BTS) corresponds to the best results obtained when additional unlabelled Catalan examples are taken into account, as explained below.

It can be observed that the GA models are highly biased towards the most frequent NE types (ORG and PER) and that the accuracy achieved on the less represented categories is very low for LOC and negligible for MIS. The MIS category is rather difficult to learn (also for the supervised algorithm), probably because it does not account for any concrete NE type and does not show many regularities. Considering this fact, we learned the models using only the LOC, ORG, and PER categories and treated the MIS as a default value (assigned whenever the classifier does not have enough evidence for any of the categories). The results obtained were even worse.

[2]In order to obtain a 100% coverage with the GA models we have introduced a naive algorithm for breaking ties in favour of the most frequent categories, in the cases in which the algorithm abstains.

	LOC	ORG	PER	MIS	avg.
GA_p	14.66	83.64	**93.88**	0.00	66.66
GA_m	20.67	**95.30**	76.94	4.00	68.28
CA	61.65	86.84	91.67	**40.00**	79.83
BTS	**65.41**	87.22	91.94	37.33	**80.63**

Table 2: Accuracy results of all models trained on Catalan resources

4.2 Bootstrapping AdaBoost models using unlabelled examples

Ideally, the supervised approach can be boosted by using the unlabelled Catalan examples in a kind of iterative bootstrapping procedure. We have tested a quite simple strategy for bootstrapping. The unlabelled data in Catalan has been randomly divided into a number of equal–size disjoint subsets $S_1 \ldots S_N$, containing 1,000 sentences each. Given the initial training set for Catalan, noted as T_L, the process is as follows:

1. Learn the M_0 classification model from T_L
2. For $i = 1 \cdots N$ do :
   (a) Classify the Named Entities in $S_1 \ldots S_i$ using model M_{i-1}
   (b) Select a subset S of previously classified examples ($S \subseteq \bigcup_{j=1}^{i} S_j$)
   (c) Learn a new model M_i using as training data $T_L \cup S$
3. Output Model M_N.

At each iteration, a new unlabelled fold is included in the learning process. First, the folds are labelled by the current model, and then, a new model is learned using the base training data plus the label–predicted folds.

it.	CA_{bts1}	CA_{bts2}	CA_{bts3}	XL_{bts2}	XL_{bts3}
0	79.83	79.83	79.41	82.63	82.42
1	78.48	79.58	79.46	**82.69**	82.29
2	78.29	79.22	**80.04**	82.45	**82.72**
3	78.13	**79.87**	**79.95**	**82.89**	**82.74**
4	78.01	79.58	79.56	**82.98**	82.45
5	78.73	79.08	79.11	**82.79**	**83.42**
6	78.22	79.07	**79.95**	**83.14**	**82.96**
7	78.25	78.93	**80.63**	**83.73**	**83.12**
8	77.99	79.14	79.65	**82.70**	**83.06**
9	78.17	79.57	79.17	82.37	**83.34**
10	78.30	78.89	79.21	82.10	**82.96**

Table 3: Accuracy results of the bootstrapping procedure for all models

We devised two variants for selecting the subset of labelled instances to include at each iteration. The first one consists of simply selecting all the examples, and the second one consists of choosing only the most confident ones (in order to avoid the addition of many training errors). For the latter, we have used a confidence measure based on the difference between the first and second highest predictions for the example (after normalization in $[-1, +1]$). The confidence parameter has been empirically set to 0.3. These two variants lead to bootstrapping algorithms that will be referred to as CA_{bts1},CA_{bts2}.

Finally, a third variant of the bootstrapping algorithm has been tested, consisting of training the M_0 model using the Catalan training set T_L plus a set of examples (of comparable size and distribution over NE types) selected from the most confidently labelled examples by the GA_{m} model. This strategy, which is applied in combination with the CA_{bts2} selection scheme, will be referred to as CA_{bts3}.

Left–hand side of table 3 contains the results obtained by these bootstrapping techniques for up to 10 iterations. Figures improving the baseline CA model are printed in boldface.

It can be observed that, frequently, the bootstrapping procedure decreases the accuracy of the system. This is probably due to two main factors: the supervised learning algorithm cannot recover from the almost 20% of errors introduced by the initial CA model, and the effect of the recognition errors (mostly in segmentation) that are present in the Catalan unlabelled corpus (recall that our NE recogniser is far from perfect, achieving 91.5 of F_1 measure).

However, significant differences can be observed between the three variants. Firstly, the simple addition of all the examples (CA_{bts1}) systematically decreases performance. Secondly, the selection of confident examples (CA_{bts2}) minimises the loss but does not allow to improve results (probably because most of the selected examples do not provide new information). Finally, the addition of the examples labelled by GA_{m} in the first learning step, though starting with a less accurate classifier, obtains better results in the majority of cases (though the bootstrapping process is certainly unstable). This seems to indicate that the information introduced by these examples is somehow complementary to that of CA. It is worth noting that GA_{m} examples do not cover the most frequent cases, since if we use them to train an AdaBoost classifier, we obtain a very low accuracy of 33%. The best result achieved by CA_{bts3} is detailed in the last row of table 2.

More complex variations to the above bootstrapping strategy have been experimented. Basically, our direction has concentrated on selecting a *right sized* set of confident examples from the unlabelled material by considering the cases in which CA and GA models agree on the prediction. In all cases, results lead to conclusions similar to the ones described above.

5 Using Spanish resources

In this section we extend our previous work on NE recognition (Carreras et al., 2003) to obtain a bilingual NE classification model. The idea is to exploit the large Spanish annotated corpus by learning a Spanish-Catalan bilingual model from the joint set of Spanish and Catalan learning examples. In order to make the model bilingual, we just have to deal with the features that are language dependent, namely the lexical ones (word forms appearing in context patterns and Bag–of–Words). All other features are left unchanged.

A translation dictionary from Spanish to Catalan and vice-versa has been automatically built for the word–form features. It contains a list of translation pairs between Spanish and Catalan words. For instance, an entry in a dictionary is "calle $\sim$ carrer", meaning that the Spanish word "calle" ("street" in

English) corresponds to the Catalan word "carrer".

In order to obtain the relevant vocabulary for the NEC task, we have run several trainings on the Spanish and Catalan training sets by varying the learning parameters, and we have extracted from the learned models all the involved lexical features. This set of relevant words contains 8,042 words (80% coming from Spanish and 20% coming from Catalan).

The translation of these words has been automatically done by applying the InterNOSTRUM Spanish–Catalan machine translation system developed by the Software Department of the University of Alacant[3]. The translations have been resolved without any context information (so, the MT system is often mistaken), and the entries not recognised by InterNOSTRUM have been left unchanged. A very light posterior hand–correcting has been done in order to fix some minor errors coming between different segmentations of translation pairs.

5.1 Cross–Linguistic features

In order to train bilingual classification models, we make use of what we call *cross–linguistic features*, instead of the monolingual word forms specified in section 2.2. This technique is exactly the same we proposed to learn a Catalan–Spanish bilingual NE recognition module (Carreras et al., 2003). Assume a feature *lang* which takes value *es* or *ca*, depending on the language under consideration. A cross–linguistic feature is just a binary feature corresponding to an entry in the translation dictionary, "es_w $\sim$ ca_w", which is satisfied as follows:

$$\text{X–Ling}_{\text{es\_w}\sim\text{ca\_w}}(w) = \begin{cases} 1 & \text{if } w = \text{es\_w and } lang = \text{es} \\ 1 & \text{if } w = \text{ca\_w and } lang = \text{ca} \\ 0 & \text{otherwise} \end{cases}$$

This representation allows to learn from a corpus consisting of mixed Spanish and Catalan examples. When an example, say in Spanish, is codified, each occurrence of a word form is checked in the dictionary and *all* translation pairs that match the Spanish entry are codified as cross–linguistic features.

The idea here is to take advantage of the fact that the concept of NE is mostly shared by both languages, but differs in the lexical information, which we exploit through the lexical translations. With this, we can learn a bilingual model which is able to classify NEs both for Spanish and Catalan, but that may be trained with few —or even any— data of one language, in our case Catalan.

[3]The InterNOSTRUM system is freely available at the following URL: http://www.internostrum.com.

5.2 Results

Table 4 shows accuracy by categories of the multilingual model XL in comparison to the best models trained only with Catalan data, already presented in section 4. As it can be seen in row XL, accuracy is increased by almost 3 points compared to supervised learning for Catalan, CA. Whereas improvement for the easiest categories (ORG and PER) is moderate, it is particularly significant for LOC and MIS, achieving improvements of 7.5 and 5.3 points, respectively.

The multilingual classifier has also been evaluated with the Spanish test set (see table 1). AdaBoost supervised algorithm has been used to learn an Spanish classifier from Spanish training data, which achieves 87.1% average accuracy. Interestingly, the multilingual classifier presents just a slight reduction to 86.9%, which could be considered irrelevant, whereas performance for Catalan is boosted by almost 3 points.

The two best–performing bootstrapping strategies for the case using only Catalan (CA_{bts2} and CA_{bts3}) have also been applied to the multilingual classifier (XL_{bts2} and XL_{bts3}). Table 3 presents the results for the first (the right–hand side of table), while figure 2 depicts the process graphically. It can be observed that both strategies consistently outperform the baseline bilingual model XL as shown in boldface figures. In this case, XL_{bts3}, again starting from a lower accuracy point, proves more stable above the baseline. This is probably due to the fact that Catalan labelled examples introduced at iteration 0 from the unsupervised classifier do not have such big impact in a bilingual model conditioned by Spanish data

	LOC	ORG	PER	MIS	avg.
CA	61.65	86.84	91.67	40.00	79.83
CA_{bts3}	65.41	87.22	91.94	37.33	80.63
XL	69.17	88.16	92.76	**45.33**	82.63
XL_{bts2}	**70.68**	**89.10**	**94.71**	41.33	**83.73**

Table 4: Accuracy results of supervised models trained on Catalan and Spanish resources

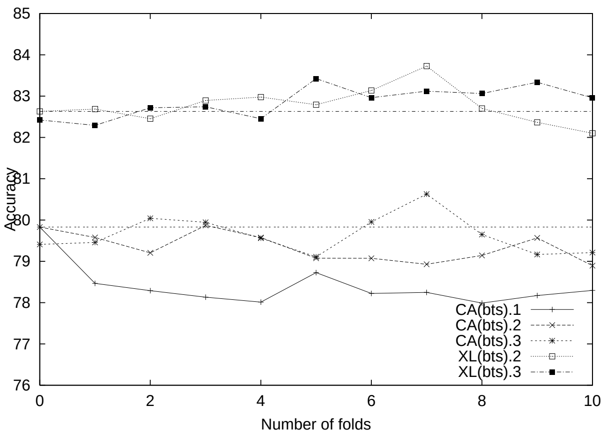

Figure 2: Progress of accuracy through bootstrapping iterations. The horizontal lines correspond to the CA and XL baselines.

than in the CA_{bts3} case. On the other hand, XL_{bts2} achieves a higher peak (increasing accuracy up to 1.1 points more than multilingual baseline XL and 3.9 more than compared to model using only Catalan data, CA) before decreasing below baseline.

6 Conclusions

We have presented a thorough experimental work on developing low–cost Named Entity classifiers for a language with no available annotated resources. Several strategies to build a Catalan NEC system have been devised and evaluated. On the one hand, using only a small initial hand–tagged corpus, supervised (AdaBoost) and fully unsupervised (Greedy Agreement) learning algorithms have been compared. On the other, using existing resources for a similar language as a starting point, a bilingual classifier has been trained. In both cases, bootstrapping strategies have been tested.

The main conclusions drawn form the presented results are:

- Given a small labelled data set, AdaBoost supervised learning algorithm clearly outperforms the fully unsupervised Greedy Agreement algorithm, even when large unlabelled text is available.

- Supervised models trained with few annotated data do not easily profit from bootstrapping strategies, even when using examples with high–confidence for retraining. Examples labelled with unsupervised models provide a complementary boost when bootstrapping.

- Multilingual models, trained with an automatically derived dictionary, are able to significantly improve accuracy for the language with less annotated resources without significantly decreasing performance in the language with more data available. Retraining with unlabelled examples performs a bit better, learning a much accurate classifier than when using only Catalan labelled examples.

Acknowledgments

Research partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02, PETRA TIC2000-1735-C02-02, ALIADO TIC2002-04447-C02), by the European Comission (FAME IST-2000-28323, MEANING IST-2001-34460), and by the Catalan Research Department (CIRIT's consolidated research group 2001SGR-00254 and predoctoral research grants 2001FI-00663 and 2003FI-00433).

References

S. Abney. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Taipei, Taiwan.

X. Carreras, L. Màrquez, and L. Padró. 2002. Named Entity Extraction Using AdaBoost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan.

X. Carreras, L. Màrquez, and L. Padró. 2003. Named Entity Recognition for Catalan Using Spanish Resources. In *Proceedings of EACL'03*, Budapest, Hungary.

M. Collins and Y. Singer. 1999. Unsupervised Models for Named Entity Classification. In *Proceedings of EMNLP/VLC-99*, College Park MD, USA.

R. Schapire and Y. Singer. 1999. Improved Boosting Algorithms Using Confidence–rated Predictions. *Machine Learning*, 37(3):297–336.

R. Schapire. 2002. The Boosting Approach to Machine Learning. An Overview. In *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA.

E. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

E. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.