

# Learning Named Entity Recognition in Portuguese from Spanish

Thamar Solorio and Aurelio López López

Instituto Nacional de Astrofísica Óptica y Electrónica  
Luis Enrique Erro # 1  
Santa María Tonantzintla, Puebla, México 72840  
{thamy,allopez}@inaoep.mx

**Abstract.** We present here a practical method for adapting a NER system for Spanish to Portuguese. The method is based on training a machine learning algorithm, namely a C4.5, using internal and external features. The external features are provided by a NER system for Spanish, while the internal features are automatically extracted from the documents. The experimental results show that the method performs well in both languages Spanish and Portuguese.

## 1 Introduction

Named entities are sequences of words that refer to a concrete entity such as person, organization, location, date and measure [1]. Named Entity Recognition (NER) consists in determining the boundaries of named entities, and even though this task is trivial for a human, the same cannot be said about making computer programs to perform this task. However, it is important to have accurate methods as Named Entities (NEs) can be valuable in several natural language applications. For instance, automatic text summarization systems can be enriched by using NEs, as they provide important cues for identifying relevant segments in text. Other uses of NE taggers are in the fields of information retrieval (i.e. more accurate Internet search engines), automatic speech recognition, question answering and machine translation.

There has been a lot of work in NER, but most approaches are targeted to specific languages, moreover, some are suitable only to narrow domains within that language. We believe this is an important disadvantage, specially considering the fact that all efforts are aimed at developing tools for a handful of languages. In this paper we present results of adapting a NE extractor for Spanish to Portuguese. Our method is based on training a machine learning classifier with the output of the NE extractor and additional lexical attributes. The experimental results are promising and represent an important advance towards cross language NER.

We begin by describing our learning scenario in section 2. We continue presenting some experimental results in section 2.3, where we compare performance

of our method when applied to Spanish and Portuguese corpora. In section 3 we describe some related work and then we conclude in section 4.

## 2 Named Entity Recognition

One of our goals is to develop a method that facilitates the adaptability of a NER system to a new domain or language. In this setting, we assume that we have available one NER system (in this case one that is targeted for Spanish) and we want to adapt it so it can be capable of performing NER accurately in documents from a different language, namely Portuguese. It is important to note here that we try to avoid the use of complex and costly linguistic tools or techniques, apart from the existing NER system, given the language restrictions they pose. Although, we do need a corpus of the target language. However, we consider the task of gathering a corpus much easier and faster than that of developing linguistic tools such as parsers, part-of-speech taggers, grammars and the like.

In our approach NER is tackled as a learning task. The features used as attributes are automatically extracted from the documents. For each word we combined two types of features: internal and external; we consider as internal features the following: the word itself, orthographic information and the position in the sentence; additionally we used some external features that are provided by the NER system for Spanish, and these are the POS tag and the NE tag (where these tags use the BIO scheme explained below). Then, the attributes for a given word  $w$  are extracted using a window of three words anchored in the word  $w$ , each word described by the internal and external features mentioned previously.

Within the orthographic information we consider 6 possible states of a word. A value of 1 in this attribute means that the letters in the word are all capitalized. A value of 2 means the opposite: all letters are lower case. The value 3 is for words that have the initial letter capitalized. 4 means the word has digits, 5 is for punctuation marks and 6 refers to marks representing the beginning and end of sentences.

The most important feature however, is the output of the NER system for Spanish. This system was developed by the TALP research center [2]. They have developed a set of NLP analyzers for Spanish, English and Catalan that include practical tools such as POS taggers, semantic analyzers and NE extractors. This NER system is based on hand-coded grammars, lists of trigger words and gazetteer information.

As in most approaches to NER we used the BIO classification scheme. Here every word in the document must be labeled with one of three tags. The *B* tag is assigned to words believed to be the beginning of a NE, the *I* tag is for words that belong to an entity but that are not at the beginning, and the *O* tag is for all words that do not satisfy any of the previous two conditions. Different from other methods we do not perform binary classifications, and we do not build specialized classifiers for each of the tags. Our classifier learns to discriminate

between the three classes and assigns labels to all the words in a single step. In Table 1 we present an example taken from the data used in the experiments where internal and external features are extracted for each word in a sentence. As we can see there are some misclassifications from the NER system for Spanish. This was expected, as Portuguese was not the target language of the system. We still believe that there is useful information provided by the existing system that will help the recognition of named entities in Portuguese.

**Table 1.** An example of the attributes used in the learning setting for NER in Portuguese

| Word         | Internal Features |      |          | External Features |         | Class |
|--------------|-------------------|------|----------|-------------------|---------|-------|
|              | Word              | Caps | Position | POS tag           | BIO tag |       |
| Somente      | Somente           | 3    | 1        | NP                | B       | O     |
| em           | em                | 2    | 2        | VM                | O       | O     |
| algumas      | algumas           | 2    | 3        | AQ                | O       | O     |
| localidades  | localidades       | 2    | 4        | NC                | O       | O     |
| de           | de                | 2    | 5        | SP                | O       | O     |
| o            | o                 | 2    | 6        | CC                | O       | O     |
| Vale         | Vale              | 3    | 7        | VM                | O       | B     |
| do           | do                | 2    | 8        | NC                | O       | I     |
| Paranapanema | Paranapanema      | 3    | 9        | NP                | B       | I     |

## 2.1 The C4.5 Algorithm

C4.5 is an extension to the decision-tree learning algorithm ID3 [3]. Only a brief description of the method is given here, more information can be found in [4]. The algorithm consists of the following steps:

1. Build the decision tree from the training set (conventional ID3).
2. Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to a leaf node.
3. Prune each rule by removing any preconditions that result in improving its accuracy, according to a validation set.
4. Sort the pruned rules in descending order according to their accuracy, and consider them in this order when classifying subsequent instances.

We used the version of C4.5 implemented in the WEKA environment, it is called J48 and is claimed to have some minor improvements over the C4.5 algorithm [5].

## 2.2 The Data sets

We used two data sets in our experiments. For evaluating NER on Portuguese we downloaded the corpus provided by the Lácio-Web project<sup>1</sup>. This corpus contains newspaper articles published by Folha de São Paulo in 1994. It consists of 1,174,206 words. The data are provided with POS tags, where named entities are labeled as proper names. From these tags we were able to measure accuracy of our system.

The other corpus is that used in the CoNLL 2002 competitions for the Spanish NE extraction task. This corpus is divided in three sets: a training set consisting of 20,308 NEs and two different sets for testing, *testa* which has 4,634 NEs and *testb* with 3,948 NEs, the former was designated to tune the parameters of the classifiers (development set), while *testb* was designated to compare the results of the competitors. As in our setting there is no parameter tuning we performed experiments with the *testb* set.

These available corpora are considerably large, so we were forced to cut down the size for both languages to probe them in a short time. We selected for experimentation smaller subsets consisting of 20,000 words for each language. In the Portuguese corpus the distribution of classes is as follows: for class *O* there were 18,012 instances, for class *B* there were 983 and for class *I* there were 1,005 instances. In the case of the Spanish corpora there were 17,114 instances of class *O* while for classes *B* and *I* there were 1,602 and 1,283 respectively.

## 2.3 Evaluation

We present here our experimental results. For all results reported here we show the overall average of several runs of 10-fold cross-validation. We used common measures from information retrieval: precision, recall and  $F_1$  and we present results from individual classes as we believe it is important in learning settings such as this, where nearly 90% of the instances belong to one class.

Table 3 presents results of using our method for the Spanish corpus. That is, all the training and testing are targeted to the Spanish language. We can see that even though the NER system performs very well by itself, by training the C4.5 algorithm on its outputs we improve performance in all the cases, with the exception of precision for class *B*. In Table 3 we show results of applying our method to the Portuguese corpus. In this case the improvements are much more notorious, particularly for class *B*, in all the cases the best results are obtained from our technique.

From the results presented above, it is clear that the method can perform NER in Portuguese with very high accuracy, however those results give no indication of the usefulness of the Spanish system for NER in Portuguese. In Tables 4 and 5 we can see comparative results of training C4.5 with only the internal features against using only the external features. As mentioned in Section 2, internal features are those extracted automatically from the documents

---

<sup>1</sup> This corpus is freely available at <http://www.nilc.icmc.usp.br/lacioweb/>

**Table 2.** Experimental results for Spanish NER

| Class | NER system for Spanish |        |        | Our method |        |       |
|-------|------------------------|--------|--------|------------|--------|-------|
|       | Precision              | Recall | $F_1$  | Precision  | Recall | $F_1$ |
| B     | 92.89%                 | 89.33% | 91.07% | 92.0%      | 93.6%  | 92.8% |
| I     | 84.36%                 | 85.22% | 84.78% | 91.6%      | 86.5%  | 89.0% |
| O     | 98.67%                 | 98.97% | 98.83% | 99.1%      | 99.3%  | 99.2% |

**Table 3.** Experimental results for Portuguese NER

| Class | NER system for Spanish |        |        | Our method |        |       |
|-------|------------------------|--------|--------|------------|--------|-------|
|       | Precision              | Recall | $F_1$  | Precision  | Recall | $F_1$ |
| B     | 58.68%                 | 91.45% | 71.48% | 87.7%      | 94.0%  | 90.8% |
| I     | 89.71%                 | 72.93% | 80.45% | 94.9%      | 91.6%  | 93.3% |
| O     | 99.03%                 | 97.05% | 98.03% | 99.5%      | 99.3%  | 99.4% |

(the original word, orthographic information, position in the word, etc.) while the external features are the ones provided by the Spanish NER system (POS and BIO tags). In the case of Spanish NER, the features from the NER system (external features) did much better than the internal features. While for Portuguese the opposite occurs, improved results are achieved when using the internal features only. This may be due to the fact that the external features for Spanish are more accurate given that the existing NER system was designed for this language. On the contrary for Portuguese the tags assigned by the NER system are not very accurate, as it is shown in Table 3. However, for both languages, the best results are attained when combining internal and external features, as shown in Tables 2 and 3. Thus, the results suggest that in order to perform NER in Portuguese we do not need an existing NER system for Spanish, but if we have one available, we can use the information provided by it and improve accuracy.

**Table 4.** Comparison of results for Spanish NER using only the internal features, such as the word and orthographic information, against using features from the NER system for Spanish (external features)

| Class | Internal features |        |       | External features |        |       |
|-------|-------------------|--------|-------|-------------------|--------|-------|
|       | Precision         | Recall | $F_1$ | Precision         | Recall | $F_1$ |
| B     | 71%               | 87.5%  | 78.4% | 92.9%             | 89.3%  | 91.1% |
| I     | 90.8%             | 35.7%  | 51.3% | 84.4%             | 85.2%  | 84.8% |
| O     | 96.7%             | 99.2%  | 97.9% | 98.7%             | 99%    | 98.8% |

**Table 5.** Comparison of results for NER in Portuguese using only the internal features, such as the word and orthographic information, against using features from the NER system for Spanish (external features)

| Class | Internal features |        |       | External features |        |       |
|-------|-------------------|--------|-------|-------------------|--------|-------|
|       | Precision         | Recall | $F_1$ | Precision         | Recall | $F_1$ |
| B     | 89.2%             | 85.9%  | 87.5% | 82.2%             | 83.2%  | 82.7% |
| I     | 90.3%             | 91.3%  | 90.8% | 86.2%             | 86.6%  | 86.4% |
| O     | 99.3%             | 99.5%  | 99.4% | 99.3%             | 99.2%  | 99.3% |

### 3 Related Work

Spanish resources for NER have been used previously to perform NER on a different language. Carreras et al. presented results of a NER for Catalan using Spanish resources [1]. They explored several methods for building NER for Catalan. Their best results are achieved using cross-linguistic features. In this method the NER system is trained on mixed corpora and performs reasonably well on both languages. Our work follows Carreras et al. approach, but differs from theirs since we apply directly the NER system for Spanish to Portuguese and train a classifier using the output and the real classes.

There has been a lot of work on NER and classification, and there is a remarkable trend towards the use of machine learning algorithms. For instance, Zhou and Su use Hidden Markov Models where the attributes are a combination of internal features such as gazetteer information, and external features such as the context of other NE already recognized [6].

In [7] a new method for automating the task of extending a proper noun dictionary is presented. The method combines two learning approaches: an inductive decision-tree classifier and unsupervised probabilistic learning of syntactic and semantic context. The attributes selected for the experiments include POS tags as well as morphological information whenever available.

One work focused in NE recognition for Spanish is based on discriminating among different kinds of named entities: core NEs, which contain a trigger word as nucleus, syntactically simple weak NEs, formed by single noun phrases, and syntactically complex named entities, comprised of complex noun phrases. Arévalo et al. focused on the first two kinds of NEs [8]. The method is a sequence of processes that uses simple attributes combined with external information provided by gazetteers and lists of trigger words. A context free grammar, manually coded, is used for recognizing syntactic patterns.

### 4 Conclusions

We present here a fast and easy method for adapting a NER system for Spanish to Portuguese. Our findings are the following:

- This proposal is a good alternative to develop tools for languages for which linguistic resources are underdeveloped. We believe that similar results can be obtained with other languages, such as Italian.
- Our method can also be applied to adapt the NER system to a different domain of the same language. As the results showed, our method outperformed the existing NER system regardless that the target documents were the same for which the NER system was created. So it is very likely that the same will happen when working in a different domain.
- The internal features suggested in this work are sufficient to train a machine learning algorithm to perform NER in Portuguese. However, the features from the NER system for Spanish in combination with the internal features proved to be very useful for enhancing results.

As work in progress, we are exploring the use of this method to named entity classification, which is a more challenging problem, given that orthographic and lexical features are less helpful. Another research direction is the adaptation of this method to cross language NER. We are very interested in exploring if, by training a classifier with mixed language corpora, we can perform NER in more than one language simultaneously.

## Acknowledgements

We would like to thank CONACyT for partially supporting this work under grants 166934 and U39957-Y.

## References

1. X. Carreras, L. Márquez, and L. Padró. Named entity recognition for catalan using spanish resources. In *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, April 2003.
2. X. Carreras and L. Padró. A flexible distributed architecture for natural language analyzers. In *Proceedings of LREC'02*, Las Palmas de Gran Canaria, Spain, 2002.
3. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
4. J. R. Quinlan. C4.5: Programs for machine learning. 1993. San Mateo, CA: Morgan Kaufmann.
5. I. H. Witten and E. Frank. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 1999.
6. G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of ACL'02*, pages 473–480, 2002.
7. G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–135. ACM Press, 2000.
8. M. Arévalo, L. Márquez, M.A. Martí, L. Padró, and M. J. Simón. A proposal for wide-coverage spanish named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, (28):63–80, May 2002.