Exploiting Named Entity Taggers in a Second Language

Thamar Solorio
Computer Science Department
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro #1, Tonantzintla, Puebla
72840, Mexico

Abstract

In this work we present a method for Named Entity Recognition (NER). Our method does not rely on complex linguistic resources, and apart from a hand coded system, we do not use any language-dependent tools. The only information we use is automatically extracted from the documents, without human intervention. Moreover, the method performs well even without the use of the hand coded system. The experimental results are very encouraging. Our approach even outperformed the hand coded system on NER in Spanish, and it achieved high accuracies in Portuguese.

1 Introduction

Given the usefulness of Named Entities (NEs) in many natural language processing tasks, there has been a lot of work aimed at developing accurate named entity extractors (Borthwick, 1999; Velardi et al., 2001; Arévalo et al., 2002; Zhou and Su, 2002; Florian, 2002; Zhang and Johnson, 2003). Most approaches however, have very low portability, they are designed to perform well over a particular collection or type of document, and their accuracies will drop considerably when used in different domains. The reason for this is that many NE extractor systems rely heavily on complex linguistic resources, which are typically hand coded, for example regular expressions, grammars, gazetteers and the like. Adapting a system of this nature to a different collection or language requires a lot of human effort, involving tasks such as rewriting the grammars, acquiring new dictionaries, searching trigger words, and so on. Even if one has the human resources and the time needed for the adaptation process, there are languages that lack the linguistic resources needed, for instance, dictionaries are available in electronic form for only a handful of languages. We believe that, by using machine learning techniques, we can adapt an existing hand coded system to different domains and languages with little human effort.

Our goal is to present a method that will facilitate the task of increasing the coverage of named entity extractor systems. In this setting, we assume that we have available an NE extractor system for Spanish, and we want to adapt it so that it can perform NER accurately in documents from a different language, namely Portuguese. It is important to emphasize here that we try to avoid the use of complex and costly linguistic tools or techniques, besides 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 the next section we present some recent work related to NER. Section 3 describes the data sets used in our experiments. Section 4 introduces our approach to NER, and we conclude in Section 5 giving a brief discussion of our findings and proposing research lines for future work.

Proceedings of the ACL Student Research Workshop, pages 25–30,
Ann Arbor, Michigan, June 2005.

2 Related Work

There has been a lot of work on NER, and there is a remarkable trend towards the use of machine learning algorithms. Hidden Markov Models (HMM) are a common choice in this setting. For instance, Zhou and Su trained HMM with a set of attributes combining internal features such as gazetteer information, and external features such as the context of other NEs already recognized (Zhou and Su, 2002). (Bikel et al., 1997) and (Bikel et al., 1999) are other examples of the use of HMMs.

Previous methods for increasing the coverage of hand coded systems include that of Borthwick, he used a maximum entropy approach where he combined the output of three hand coded systems with dictionaries and other orthographic information (Borthwick, 1999). He also adapted his system to perform NER in Japanese achieving impressive results.

Spanish resources for NER have been used previously to perform NER on a different language. Carreras et al. presented results of a NER system for Catalan using Spanish resources (Carreras et al., 2003a). 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 in that we apply directly the NER system for Spanish to Portuguese and train a classifier using the output and the real classes.

In (Petasis et al., 2000) 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 on 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 and colleagues focused on the first two kinds of NEs (Arévalo et al., 2002). 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.

3 Data sets

In this paper we report results of experimenting with two data sets. The corpus in Spanish is that used in the CoNLL 2002 competitions for the NE extraction task. This corpus is divided into 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. We performed experiments with *testa* only.

For evaluating NER on Portuguese we used the corpus provided by "HAREM: Evaluation contest on named entity recognition for Portuguese". This corpus contains newspaper articles and consists of 8,551 words with 648 NEs.

4 Two-step Named Entity Recognition

Our approach to NER consists in dividing the problem into two subproblems that are addressed sequentially. We first solve the problem of determining boundaries of named entities, we called this process Named Entity Delimitation (NED). Once we have determined which words belong to named entities, we then get to the task of classifying the named entities into categories, this process is what we called Named Entity Classification (NEC). We explain the two procedures in the following subsections.

4.1 Named Entity Delimitation

We used the BIO scheme for delimiting named entities. In this approach each word in the text is labeled with one out of three possible classes: 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.

Table 1: An example of the attributes used in the learning setting for NER in Spanish. The fragment presented in the table, *"El Ejército Mexicano puso en marcha el Plan DN-III"*, translates as "The Mexican Army launched the DN-III plan"

Internal Features			External Features		
Word	Caps	Position	POS tag	BIO tag	Class
El	3	1	DA	O	O
Ejército	2	2	NC	B	B
Mexicano	2	3	NC	I	I
puso	2	4	VM	O	O
en	2	5	SP	O	O
marcha	2	6	NC	O	O
el	3	7	DA	O	O
Plan	2	8	NC	B	B
DN-III	3	9	NC	I	I

In our approach, NED is tackled as a learning task. The features used as attributes are automatically extracted from the documents and are used to train a machine learning algorithm. We used a modified version of C4.5 algorithm (Quinlan, 1993) implemented within the WEKA environment (Witten and Frank, 1999).

For each word we combined two types of features: internal and external; we consider as internal features the word itself, orthographic information and the position in the sentence. The external features are provided by the hand coded NER system for Spanish, these are the Part-of-Speech tag and the BIO tag. Then, the attributes for a given word w are extracted using a window of five 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 hand coded system used in this work was developed by the TALP research center (Carreras and Padró, 2002). 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.

In contrast to other methods we do not perform binary classifications, as (Carreras et al., 2003b), thus we do not build specialized classifiers for each of the tags. Our classifier learns to discriminate among the three classes and assigns labels to all the words, processing them sequentially. 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.

4.1.1 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 a learning setting such as this, where nearly 90% of the instances belong to one class.

Table 2 presents comparative results using the Spanish corpus. We show four different sets of results, the first ones are from the hand coded system, they are labeled *NER system for Spanish*. Then we present results of training a classifier with only the internal features described above, these results are labeled *Internal features*. In a third experiment we trained the classifier using only the output of the NER system, these are under column *External features*. Finally, the results of our system are presented in column labeled *Our method*. 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. Given that the hand coded system was built for this collection, it is very encouraging to see our method outperforming this system. In Table 3 we show results of applying our method to the Portuguese corpus. In this case the improvements are much more impressive, particularly for class B, in all the cases the best results are obtained from our technique. This was expected as we are using a system developed for a different language. But we can see that our method yields very competitive results for Portuguese, and although by using only the internal features we can outperform the hand coded system, by combining the information using our method we can increase accuracies.

Table 2: Comparison of results for Spanish NE delimitation

	NER system for Spanish			Internal features			External features			Our method		
Class	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
B	92.8	89.3	91.7	87.1	89.3	88.2	93.9	91.5	92.7	93.5	92.9	**93.2**
I	84.3	85.2	84.7	89.5	77.1	82.9	87.8	87.8	85.7	90.6	87.4	**89.0**
O	98.6	98.9	98.8	98.1	98.9	98.5	98.7	99	98.9	98.9	99.2	**99.1**
overall	**91.9**	**91.1**	**91.7**	**91.5**	**88.4**	**89.8**	**93.4**	**92.7**	**92.4**	**94.3**	**93.1**	**93.7**

Table 3: Experimental results for NE delimitation in Portuguese

	NER system for Spanish			Internal features			External features			Our method		
Class	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
B	60.0	68.8	64.1	82.4	85.8	84.1	75.9	81.0	78.4	82.1	87.8	**84.9**
I	64.5	73.3	68.6	80.1	76.8	78.4	73.8	70.3	72.0	80.9	77.8	**79.3**
O	97.2	95.5	96.4	98.7	98.5	**98.6**	98.1	97.7	97.9	98.8	98.4	**98.6**
overall	**73.9**	**79.2**	**76.3**	**87.0**	**87.0**	**87.0**	**82.6**	**83.0**	**82.7**	**87.2**	**88.0**	**87.6**

From the results presented above, it is clear that the method can perform NED in Spanish and Portuguese with very high accuracy. Another insight suggested by these results is that in order to perform NED in Portuguese we do not need an existing NED system for Spanish, the internal features performed well by themselves, but if we have one available, we can use the information provided by it to build a more accurate NED method.

4.2 Named Entity Classification

As mentioned previously, we build our NE classifiers using the output of a hand coded system. Our assumption is that by using machine learning algorithms we can improve performance of NE extractors without a considerable effort, as opposed to that involved in extending or rewriting grammars and lists of trigger words and gazetteers. Another assumption underlying this approach is that of believing that the misclassifications of the hand coded system for Spanish will not affect the learner. We believe that by having available the correct NE classes in the training corpus, the learner will be capable of generalizing error patterns that will be used to assign the correct NE. If this assumption holds, learning from other's mistakes, the learner will end up outperforming the hand coded system.

In order to build a training set for the learner, each instance is described with the same attributes as for the NED task described in section 4.1, with the addition of a new attribute. Since NEC is a more difficult task, we consider useful adding as attribute the suffix of each word. Then, for each instance word we consider its suffix, with a maximum size of 5 characters.

Another important difference between this classification task and NED relies in the set of target values. For the Spanish corpus the possible class values are the same as those used in CoNLL-2002 competition task: *person*, *organization*, *location* and *miscellaneous*. However, for the Portuguese corpus we have 10 possible classes: *person*, *object*, *quantity*, *event*, *organization*, *artifact*, *location*, *date*, *abstraction* and *miscellaneous*. Thus the task of adapting the system for Spanish to perform NEC in Portuguese is much more complex than that of NED given that the Spanish system only discerns the four NE classes defined on the CoNLL-2002. Regardless of this, we believe that the learner will be capable of achieving good accuracies by using the other attributes in the learning task.

4.2.1 Experimental Results

Similarly to the NED case we trained C4.5 classifiers for the NEC task, results are presented in Tables 4 and 5. Again, we perform comparisons between the hand coded system and the use of different subsets of attributes. For the case of Spanish NEC, we can see in Table 4, that our method using internal and external features presents the best results. The improvements are impressive, specially for the NE class *Miscellaneous* where the hand coded system achieved an F measure below 1 while our system achieved an F measure of 56.7. In the case of NEC in Portuguese the results are very encouraging. The

Table 4: NEC performance on the Spanish development set

	NER system for Spanish			Internal features			External features			Our method		
Class	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
Per	84.7	93.2	88.2	94.0	62.9	75.3	88.3	93.1	90.6	88.2	95.4	**91.7**
Org	78.7	88.7	82.9	61.7	90.0	73.2	77.7	91.9	84.2	83.4	89.0	**86.1**
Loc	78.7	76.2	76.9	78.4	65.1	71.2	80.3	80.3	80.3	82.0	82.5	**82.2**
Misc	24.9	.004	.008	75.5	42.0	54.0	52.9	23.4	33.5	71.6	46.9	**56.7**
overall	**66.7**	**64.5**	**62.0**	**77.4**	**65.0**	**68.4**	**74.8**	**72.1**	**72.1**	**81.3**	**78.4**	**79.1**

hand coded system performed poorly but by training a C4.5 algorithm results are improved considerably, even for the classes that the hand coded system was not capable of recognizing. As expected, the external features did not solve the NEC by themselves but contribute for improving the performance. This, and the results from using only internal features, suggest that we do not need complex linguistic resources in order to achieve good results. Additionally, we can see that for some cases the classifiers were not able of performing an accurate classification, as in the case of classes *object* and *miscellaneous*. This may be due to a poor representation of the classes in the training set, for instance the class *object* has only 4 instances. We believe that if we have more instances available the learners will improve these results.

5 Conclusions

Named entities have a wide usage in natural language processing tasks. For instance, it has been shown that indexing NEs within documents can help increase precision of information retrieval systems (Mihalcea and Moldovan, 2001). Other applications of NEs are in Question Answering (Mann, 2002; Pérez-Coutiño et al., 2004) and Machine Translation (Babych and Hartley, 2003). Thus it is important to have accurate NER systems, but these systems must be easy to port and robust, given the great variety of documents and languages for which it is desirable to have these tools available.

In this work we have presented a method for performing named entity recognition. The method uses a hand coded system and a set of lexical and orthographic features to train a machine learning algorithm. Apart from the hand coded system our method does not require any language dependent features, we do not make use of lists of trigger words, neither we use any gazetteer information. The only information used in this approach is automatically extracted from the documents, without human intervention. Yet, the results presented here are very encouraging. We were able to achieve good accuracies for NEC in Portuguese, where we needed to classify NEs into 10 possible classes, by exploiting a hand-coded system for Spanish targeted to only 4 classes. This achievement gives evidence of the flexibility of our method. Additionally we outperform the hand coded system on NER in Spanish. Thus, our method has shown to be robust and easy to port to other languages. The only requirement for using our method is a tokenizer for languages that do not separate words with white spaces, the rest can be used pretty straightforward.

We are interested in exploring the use of this method to perform NER in English, we would like to determine to what extent our system is capable of achieving competitive results without the use of language dependent resources, such as dictionaries and lists of words. 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.

References

Montse Arévalo, Xavier Carreras, Lluís Màrquez, Toni Martí, Lluís Padró, and Maria José Simon. 2002. A proposal for wide-coverage Spanish named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, (28):63–80, May.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the EACL 2003 Workshop on MT and Other Language Technology Tools*, pages 1–8.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high perfor-

Table 5: NEC performance on the Portuguese set

	NER system for Spanish			Internal features			External features			Our method		
Class	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
Pessoa (Person)	34.8	72.5	46.6	49.1	92.0	**64.0**	46.9	64.6	54.4	45.5	91.1	60.7
Coisa (Object)	0	0	0	0	0	0	0	0	0	0	0	0
Valor (Quantity)	0	0	0	82.1	47.1	59.8	74.6	69.1	71.8	77.6	76.5	**77.0**
Acontecimento (Event)	0	0	0	33.3	21.4	26.1	14.3	7.1	9.5	50.0	21.4	**30.0**
Organização (Organization)	41.4	38.4	39.3	70.7	56.9	**63.1**	45.7	56.9	50.7	79.3	49.2	60.8
Obra (Artifact)	0	0	0	76.6	64.3	**69.9**	29.4	8.9	13.7	74.4	57.1	64.6
Local (Location)	52.5	16.5	24.8	72.6	32.6	**45.0**	43.6	38.5	40.9	67.4	32.1	43.5
Tempo (Date)	0	0	0	74.0	86.6	79.8	85.5	83.9	84.7	87.0	83.9	**85.5**
Abstracção (Abstraction)	0	0	0	82.1	41.8	**55.4**	22.2	3.6	6.3	79.3	41.8	54.8
Variado (Miscellaneous)	0	0	0	1	15.4	**26.7**	0	0	0	1	15.4	**26.7**
overall	**12.8**	**12.7**	**11.0**	**54.1**	**45.8**	**48.9**	**36.2**	**33.2**	**33.2**	**56.1**	**46.8**	**50.3**

mance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201.

Daniel M. Bikel, Richard Schwartz, and Ralph Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning, Special Issue on Natural Language Learning*, 34(1–3):211–231, February.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University, New York, September.

Xavier Carreras and Lluís Padró. 2002. A flexible distributed architecture for natural language analyzers. In *Proceedings of LREC'02*, Las Palmas de Gran Canaria, Spain.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003a. 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.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003b. A simple named entity extractor using adaboost. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 152–155. Edmonton, Canada.

Radu Florian. 2002. Named entity recognition as a house of cards: Classifier stacking. In *Proceedings of CoNLL-2002*, pages 175–178. Taipei, Taiwan.

Gideon S. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan.

Rada Mihalcea and Dan Moldovan. 2001. Document indexing using named entities. *Studies in Informatics and Control*, 10(1), January.

Manuel Pérez-Coutiño, Thamar Solorio, Manuel Montes y Gómez, Aurelio López López, and Luis Villaseñor Pineda. 2004. Question answering for Spanish based on lexical and context annotation. In Christian Lemaître, Carlos Reyes, and Jesús A. González, editors, *Advances in Artificial Intelligence – IBERAMIA 2004*, Lecture Notes in Artificial Intelligence 3315, pages 325–333, Puebla, Mexico, November. Springer.

Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D. Spyropoulos. 2000. 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.

J. R. Quinlan. 1993. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Thamar Solorio. 2005. *Improvement of Named Entity Tagging by Machine Learning*. Ph.D. thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, Mexico, (to appear).

Paola Velardi, Paolo Fabriani, and Michel Missikoff. 2001. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems*, pages 270–284. ACM Press.

Ian H. Witten and Eibe Frank. 1999. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Tong Zhang and David Johnson. 2003. A robust risk minimization based named entity recognition system. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 204–207. Edmonton, Canada.

Guodong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of ACL'02*, pages 473–480.