Procesamiento del Lenguaje Natural, núm. 36 (2006), pp. 25-37 recibido 01-03-2006; aceptado 01-06-2006

# Lessons from the Development of a Named Entity Recognizer for Basque

**Alegría I.**
Informatika F.
Donostia
i.alegria@si.ehu.es

**Arregi O.**
Informatika F.
Donostia
acparuro@si.ehu.es

**Ezeiza N.**
Informatika F.
Donostia
n.ezeiza@si.ehu.es

**Fernandez I.**
Informatika F.
Donostia
acbfegoi@si.ehu.es

**Resumen:** En este trabajo se presenta el diseño de un sistema de Reocnocimiento de Entidades para textos escritos en vasco. Para el desarrollo de dicho sistema se han probado diferentes técnicas, algunas de ellas basadas en información linguística y otras en cambio aplicando diversos métodos de aprendizaje automático. Además de presentar cada técnica por separado y sus correspondientes experiementos con diferentes fuentes de información, proponemos también una serie de combinaciones con diferentes metodos para obtener así un sistema más completo y robusto. Para concluir, presentamos las conclusiones y reflexiones concluidas de todos estos experimentos, especialmente válidas para aquellos sistemas que traten el reconocimiento de entidades en textos escritos en otros idiomas que no sea el inglés.

**Palabras clave:** Reconocimiento de Entidades con nombre, conocimiento linguístico, Aprendizaje automático, Métodos combinados

**Abstract:** This paper presents the conclusions reached from the development of a system for Named Entity recognition in written Basque. In order to obtain this recognizer we have worked with different types of classifiers, one of them based on linguistic information and others constructed using machine learning methods. Taking these classifiers as starting point, and once we explain the different attempts done with each simple method using different information sources, we present the experiments we did combining those single methods in order to improve the performance and obtain a more robust system. Finally, we explain some conclusions and lessons we have learned from all these experiments, especially useful when dealing with named entity recognition in languages others than English.

**Keywords:** Named Entity Recognition, Linguistic Knowlege, Machine Learning and Combined methods

## 1 *Introduction and Related Work*

Named Entity Recognition and Classification (NERC) constitutes a very important task in Natural Language Processing (NLP) and specifically in tasks related to Information Extraction.

As defined in the *Message Understanding Conference* (MUC) (Chinchor, 1998), Named Entity (NE) recognition consists on the identification and categorization of entity names (person, organization and location), temporal expressions (dates and times), and some types of numerical expressions (percentages, monetary values and so on). Sometimes two different subtasks are distinguished: Named Entity Recognition (NER) which determines the boundaries of the entities, and Named Entity Classification (NEC) which assigns the corresponding type of entity. When we refer to NERC as a task, it involves the sequential application of both NER and NEC tasks. Among the different techniques used to process these data, we find some systems based on statistical methods, some based on strictly linguistic methods that make use of grammar rules (Magnini *et al.*, 2002), and finally the ones that combine rules and statistics (Mikheev *et al.*, 1998). Some previous work involving an analysis of the written text is generally required before applying these techniques. In the simplest cases, only tokenization is applied, but in other cases, morphological analysis, disambiguation, and the attachment of semantic features must also be carried out.

The machine learning (ML) paradigm has been very successful on this task and recently language-independent shared tasks have been evaluated with different systems using the same corpus (Tjong, 2002) (Tjong & De Meulder, 2003). Observing these workshops we can say that Maximum Entropy, AdaBoost or the combination of several methods achieve good results. The features used are quite well-defined too, such as a window with the previous and next words, including information about spelling (mainly capital letters), part-of-speech (PoS), affixes, chunk tags and occurrence in gazetteers. Anyway, results in CoNLL-2003 " *do not reveal a single feature that would be ideal for NER* " (Tjong & De Meulder, 2003).

The paper is structured as follows: Section 2 presents the aims and design of the project. Section 3 deals with the first step, a linguistic recognizer based on a grammar for the identification and a heuristic for the classification. In section 4 we describe the different experiments in order to obtain the best possible tool using ML based methods. In section 5 we present the experiments for NERC, and finally, section 6 presents conclusions and future work.

## 2 Experimental Context

The need of a NE recognizer for Basque was observed in various projects of the IXA Group (*ixa.si.ehu.es*). As it can be used as a basic tool for future projects in Language Engineering, we decided to build a system for identifying and classifying NEs in written Basque texts.

Basque has a very rich inflectional system in which the article and the case marks (corresponding to prepositions in other languages) are attached to the word stem (Alegria *et al.*, 1996). This also applies to proper nouns (even to foreign ones). Thus, proper nouns can occur in several forms e.g. the person name Izaskun can occur in the form of Izaskun, Izaskunek, Izaskuni, Izaskunentzat and so on.

Although the forms *Izaskunek*, *Izaskuni* and *Izaskunentzat* have a clear inflectional case, this does not happen in the form *Izaskun.* Izaskun has two possible morphological analysis: One with no-declension case and the other with the absolutive case. This is the case for 'Errusia' in the sentence: 'Gaizki egon da Errusia Jeltsin pean' (*Russia has not been well governed by Jeltsin*). As we read this sentence we tend to think that 'Errusia Jeltsin' is one entity, as the absolutive case is not visible, but actually we have two separate entities. So this non-visible absolutive cases can create ambiguity when dealing with entity recognition.

Regarding the structure and the nature of the elements of Basque named entities, the following aspects can be underlined, some of which are shared by other languages:

- Many family names and even first names can be common nouns or adjectives. For example the adjective 'zabala' ('wide') is a common surname.
- A lot of location names can be person surnames too. In the Basque Country, people have been given the name of their birthplace or living place (house or quarters) for centuries.
- Interferences from dialectal uses and from spelling rules in neighbouring languages are also frequent. The late standardization process of Basque and its wide dialectal use may cause the same name to be used in different forms e.g. 'Etxeberria' and 'Etxebarria'. In addition to this, it is not strange to find Basque names in Basque texts written with Spanish or French spelling rules e.g. 'Echeverria', 'Echevarria' or 'Echeverria'.
- Obviously, the use of foreign names is also essential, for example, in news reports. Foreign acronyms are especially frequent since it takes longer to adapt these neologisms to Basque. For instance, ISDN/RDSI or DNA/ADN may

appear in Basque written texts either in English or in Spanish, depending on the author.

## 2.1 Resources

According to (McDonald, 1996), there are two kinds of data that should be taken into account in order to identify and classify the possible NEs: internal evidence and external evidence. The former is provided by the expression itself and its elements, and the latter by the context in which it occurs.

In order to exploit the information provided by the entity itself and its elements as much as possible, we have used a lemmatizer/tagger for Basque (Ezeiza *et al.*, 1998). This tool performs the tokenization, the morphosyntactic analysis and the PoS disambiguation on a text, so it allows us work not only with the shallow information of words but also with specific linguistic information. This kind of information is very useful for both NER and NEC task, specially when we are dealing with an agglutinative language such as Basque.

We have used two types of lists as external information: trigger word lists and gazetteers. While the former provide the information about the words around the entity, the latter concentrate on some typical classification of the entities. We borrowed some gazetteers for different categories (Person (PER), Organization (ORG) and Location (LOC)) from *Euskaldunon Egunkaria*, a newspaper entirely written in Basque, and the gazetteer for PER was enriched with information taken from the census of the local government. Thus, we obtained a PER gazetteer with about 121,503 instances, a LOC list with 1,444 instances and an ORG one with 973.

The trigger word list used in the first approaches was extracted from a small news corpus on economics at the beginning of the project. But as we were going ahead, we observed that this information was not enough and we relied on WordNet (WN) information in order to enrich it. For that purpose, we used the work developed in (Magnini *et al.*, 2002). Concretly, as WN synsets represent exactly the same word for every language, we borrowed the synset lists of trigger words from the work mentioned above, and we used EuskalWordNet[1] to obtain the corresponding Basque forms. With this approach we increased trigger words of PER category from 29 to 3,825, LOC category trigger words from 71 to 895 and finally ORG category trigger words from 11 to 594.

The approach made using WN was not only used for trigger words, we also tried to enrich the information of gazetteers using it, but the results were not as good as with triggers. We only obtained 174 new instances for PER, 103 for ORG, while the new instances for LOC were 322.

## 2.2 Project Design

The main goal of our system is to recognize expressions referring to Persons, Locations, Organizations and Others. Numerical and temporal expressions are already captured by the lemmatizer/tagger used in the pre-process. The problem of nested named entities will not be solved in this work, but it would be considered on further works.

When we faced the task we considered two possibilities:

- to tag a corpus directly by hand and to construct the tool using exclusively ML techniques.
- to build a linguistic tool which will help us tag a corpus.

Although ML techniques offer robustness and good results, we finally decided to build a linguistic recognizer, bearing in mind that:

- it can offer good results too.
- it can be combined with ML based methods and they can be complementary.
- the linguistic features defined for this method can be useful for further steps.
- the necessary linguistic pre-process has already been carried out.
- it permits a semiautomatic way to annotate the corpus for evaluation (and for supervised learning with the ML methods), making this task simpler and faster.

So the project was designed in four steps. Firstly, we developed a tool based on linguistic information. Secondly, we generated

[1] WordNet for Basque. http://sisx03.si.ehu.es/cgi-bin/privat-synset-mysql/wei2.editsynset.perl

semi-automatically annotated corpus (hand-reviewed). Thirdly, we trained different ML techniques on these corpus in order to obtain the best possible recognizer. And finally, the recognizers obtained were combined so as to improve the results, of both identification and classification.

## *3 The Linguistic Tool*

The linguistic version, named Eihera, is the result of the first step. It consists of a grammar for identification and its rules are based on the morphological information provided by the lemmatizer/tagger and a heuristic that uses internal and external evidence for classification.

### 3.1 Identification

XFST (*Xerox Finite State Transducer*) (Beesley & Karttunen, 2003) is the tool used to develop the grammar for the identification of entities. It allows us to define both the structure of entity names (organization, location and person names) and the rules for their identification.

#### 3.1.1 Main elements of the grammar

We find entity names and trigger words among the main elements of our grammar. Although we will not consider the latter ones relevant for the identification of the entities, they will be helpful for classification.

Most of the words in named entity expressions are capitalized in Basque, as it happens in English and in Spanish. That is why the main feature of the whole entity is the use of capital letters. But there are other features that should be taken into account, for instance the category and subcategory of the elements and their inflection.

The entity elements can be divided into these main category/subcategories: IZE (common noun), IZB (proper noun), LIB (location/organization proper noun), ADJ (adjective), SIG (acronym) and BST (particle[2]).

The elements in the entity must be written in capitals except the cases of some BST. For the identification of entities we make a distinction among non-case elements, elements in genitive, and others, because the first two can indicate the continuation of the entity, while other cases point out the boundary of the entity.

#### 3.1.2 Main patterns

Two patterns of named entities are distinguished in the grammar: entities containing a single element (*Europa(n)*[3] LOC) and entities composed of more than one element (*Europako Banku Zentral(e)a(n)*[4] ORG).

For entities consisting on a single element, the element must be a proper noun (IZB PoS tag, which will be assigned at the preprocess by the lemmatizer/tagger), a location/organization proper noun (LIB) or an acronym (SIG). Words with other PoS tag will not be considered as candidate for this kind of entities. There is no restriction related to declension. The element can bear any case when it is declined because elements with no-declension case are also accepted. For example, "EHU"[5] is an acronym with no-declension case. "Bilbon"[6] is a location name declined in the inesive. Both will be considered single entities.

The entities with more than one element follow a more complex pattern in both PoS tags and declension restrictions. The last element of the entity and the rest of the elements are different. While the last element can bear any case, the other elements have more declension restriction. They can only be declined in genitive. Regarding the PoS, more tags are accepted. In addition to PoS tags accepted by single-element entities, elements can also be common nouns (IZE), adjectives (ADJ) or some special particles (BST).

Examples of the second pattern are:

- *Europako* (LIB PoS tag element + *ko* LOCATION-GENITIVE affix[7]) *Banku* (IZE + non declension case) *Zentralean* (ADJ+ *ean* INESIVE declension affix)[8]
- *Alex* (IZB) *de* (BST) *la* (BST) *Iglesiak* (IZB + *k* ERGATIVE affix)[9]

Those two patterns described above, delimit the structure of the entities we have to deal with. When we apply them in a text and when both of them can be applied, the

[2] BST stands for particles that occur in some entities borrowed from Spanish, such as 'de' in *Santiago de Compostela*

[3] In Europe

[4] At the European Central Bank

[5] University of the Basque Country

[6] Bilbo is a location name + 'n' inesive declension affix which adds some conceptual information to the element. The complete meaning of the entity is 'in Bilbo'

[7] European or From Europe

[8] At the European Central Bank.

[9] Alex de la Iglesia, a spanish cinema director.

longest element sequence that matches them will be marked as entity candidate. That is the only way XFST allows us to work without any pre-defined word sequence length, and that is why we decided not to treat nested named entities in this first approach.

## 3.2 Classification

For the classification, we apply a heuristic that uses simple trigger words list and gazetteers.

Apart from these sources, the heuristic makes use of linguistic information provided by the entity itself, in the following way:

Step 1:

The identified entities are matched up to the ones in the gazetteers, and when coincidences occur they are assigned the category of the corresponding gazetteer. If no matches are found, the process goes to

Step 2:

The heuristic selects the elements in the entity one by one: first it selects the last element and then it goes leftwards analyzing their declension and PoS. Depending on the information it gets, different weights are assigned to the classification categories. For example, if the element has been tagged with IZB, then person's weight will be increased while location and organization categories will not be changed, because it is more probable that the whole entity refers to a person when proper noun elements appear in it, than to a location. In case there is any genitive among the elements, a further analysis is applied. For instance, in *Europako Banku Zentralean*[10], *Europako* has genitive declension, so we only consider the elements *Banku Zentralean* for classification. This is due to the fact that Basque is a head-final language.

Therefore, the heuristic considers the words to the left until it finds a genitive and when it finds one, this and every element preceding it will not be relevant for the weight assignment.

Step 3:

If there is some trigger word identified together with the entity, it is selected and searched for in the corresponding list of trigger words. In case some of them matches the corresponding list, the weight for assigning its category increases.

Step 4:

The heuristic finds out which category has obtained higher weight and assigns it to the entity.

## 3.3 Results

Since we have made a distinction between identification and classification when we designed Eihera, this distinction is also reflected in the evaluation process.

In a test corpus of 931 entities, Eihera identifies 951 from which 805 are correct. Results are summarized in Table 1.

| | Precision | Recall | F-score |
|---|---|---|---|
| Eihera | 84.65 | 86.10 | 85.37 |

Table 1: Results for Eihera in NER

We revised 100 incorrect NEs, randomly selected, in order to find causes of errors. As shown in Table 2, half of the identification errors (35%+29%) are due to external input reasons such as typographic errors, capitalization errors, and so on, that frequently appear in newspaper articles; thus, they would not occur in accurately written texts. Most of the remaining errors could be corrected with the improvement of the tagger.

| Reason | Percentage |
|---|---|
| Errors in capital letters | 35% |
| Bad analyses in pre-process | 29% |
| Errors in the input format | 22% |
| Nested NEs | 8% |
| Others | 6% |

Table 2: Source of errors in identification

Results are poorer in classification than in identification but they can not be considered bad results. It classifies 679 expressions with the correct category out of 931 entities in the test corpus, so it obtains 72.93% precision.

As it was planned in the design, we will see in the next section how the system becomes more robust and better results can be obtained when ML techniques are used.

## *4 Experiments with ML Methods*

The corpus obtained by applying Eihera on articles from *Euskaldunon Egunkaria* was hand-reviewed and then BIO tagged[11]. We divided it in the following way in order to use it for ML techniques: 46,227 words and 3,817

[10] *In the European Central Bank.*

[11] B-begin of entity, I-intermediate, O-out of entity for NER task and LOC, PER, ORG and OTH for classification.

entities were used for the training corpus and 15,960 words and 931 entities for testing[12].

We considered Weka a good choice for ML experiments (Whitelaw & Patrick, 2003), because, apart from being a free software with GPL license, it is also a very flexible tool that allows us to experiment with different methods.

In order to compare the results to the state of the art in the task, we trained Abionet (Carreras *et al.*, 2003) with our corpus. This method, based on the AdaBoost algorithm was one of the best methods in both CoNLL-2002 and CoNLL-2003 for NERC in several languages.

Our work with these tools has been focused on the selection and tunning of features.

A language independent system[13], based on multiple orthographic tries which are combined with a Hidden Markov model framework (Whitelaw & Patrick, 2003) has been also used for the NER task. Although this tool achieves poorer results than the methods mentioned before, we thought that a different approach could help to improve the results when combining methods.

Once we have learned from the training corpus and tagged the test corpus, the evaluation can be carried out. We always assign one and only one category to each entity expression. That is why only one measure (precision) can be estimated for NEC and there are some measures for NER:

- Average precision, recall and F-score for all the BIO categories.
- Precision, recall and F-score for B and I categories, since they are the target of the system.
- Precision, recall and F-score of the identified entities calculated as in CoNLL.

The last measure is the most useful in order to compare different methods and languages. So this will be the one shown in the evaluation tables.

## 4.1 Identification

There are many variants when using Weka to train a NER classifier. Now we will present the features and methods which best perform for an agglutinative language such as Basque.

### 4.1.1 Features

Based on the related work and the relevant information in the grammar, different features were proposed and tested using ML methods.

The key-features we have applied are the following (similar results were obtained with all the methods): word, lemma, PoS, declension case, capitalized word, capitalized lemma, word in capitals, and lemma in capitals.

Other features, such as the explicit indication of punctuation signs, were excluded because they did not improve the results.

A [-3,+3] window was applied, so 56 features per word were obtained (8 features in 7 words).

Some systems in CoNLL-2003 use external information in order to extract capitalization features, but we think that the information of the capitalized lemmas obtained from the pre-process is equivalent.

### 4.1.2 Methods

We considered the following: for the first attemps on Named Entities recognition at the ML area: Naive Bayes (NB), C4.5 and Support Vector Machines (SVM)[14] from Weka, and AdaBoost from Abionet (Carreras *et al.*, 2003). We knew that some of them performed well for the task we were dealing with and all of them seemed to be quite simple to use.

| | Precision | Recall | F-score |
|---|---|---|---|
| NB | 46.08 | 54.76 | 50.05 |
| Sydney | 74.74 | 77.54 | 76.12 |
| SVM | 85.02 | 81.93 | 83.44 |
| C4.5 | 86.74 | 82.57 | 84.60 |
| Eihera | 84.65 | 86.10 | 85.37 |
| Abionet-BIO [15] | 89.22 | 85.88 | 87.52 |
| Abionet | 89.81 | 85.78 | 87.75 |

Table 3: NER results of different methods

When we selected NB and C4.5 we knew they were not the best performing methods for the identification task; however, we wanted to test them with a double goal: to compare the results with the ones obtained

[12] The same test-corpus is used in all experiments.

[13] We will refer to it as Sydney in the tables.

[14] SVM is a binary classifier but in Weka is extended to handle multi-class using pair-wise classification.

[15] Abionet-BIO represents the result obtained in Abionet without using BIO-tags of previous words. This measure is taken because this features can not be used in NB, C4.5 and SVM, because their implementation does not permit it.

with Abionet, and to have different recognizers to combine.

Aditionaly, the University of Sydney evaluated their language independent system on our corpus.

The results of the different systems are shown in Table 3.

Some conclusion can be drawn:

- SVM does not overcome the results of C4.5 and it needs a very long time for training.
- The results obtained by C4.5 are 3% lower than the results of Abionet, which represents not only state of the art for English and Spanish as it was proved in CoNLL02 competition, but also for Basque as we have just seen in our experiments' results.
- The performance of the linguistic system (Eihera) is poorer than Abionet but better than the others.

### 4.1.3 Selection of features

The selection of a set of features can improve the results. The use of words and/or lemmas was a challenge that we had studied in documents' classification and we considered interesting to test it on NE identification. A lemma can concentrate information of several words (specially in agglutinative languages), but they lose other kind of information (declension, time, ...).

| | w+l | words | lemmas |
|---|---|---|---|
| NB | 50.05 | 50.02 | 62.36 |
| C4.5 | 84.60 | 84.59 | 84.94 |
| SVM | 83.44 | 84.25 | 83.38 |

Table 4: NER results (F-score) with words and/or lemmas

In the evaluated version we took advantage of both words and lemmas but this seemed not to be a good option for feature selection. So, we tested the results with only words, only lemmas and compare them to the previous results got by using both. The results for F-score are in Table 4.

Looking at the table, we can conclude that the use of only lemmas improves the results fairly in Naive Bayes. This improvement applies only to precision, which raises from 46.08 to 72.6 using Bayes, and only from 50 to 62.36 in F-score. Using C4.5 the gain is weaker (only in precision too).

### 4.1.4 Selection of instances

Another technique used to improve the results is to try to set a better sampling, discarding errors and avoiding overfitting. As a consequence of this selection, the learning process can become faster too[16].

We thought that the O category may unbalance the corpus. Researches in the MIT (Mikheev *et al.*, 1999) show the negative effect of classes with sample of different quantity of instances using Naive Bayes. In the NER task this effect can be even worse because, although the system classifies BIO categories, the evaluation is only done over correctly classified entities. In fact the I category shows less performance than the others. That is why we decided to remove some examples of words that were not part of the entities, and to analyse the results. A simple heuristic can discard, with very high confidence, words that do not belong to entities. In our case, the elimination of all words with non-capital letters, except for nouns, conjunctions and BST tagged words rules out almost half of the instances (the number of instances falls from 46,227 words to 22,264 in the training corpus), while only 0.4% of the entities are left out. We have the same proportion in the test corpus.

There are two possible ways of discarding words:

- applying the feature-extracion window so as to obtain the examples, after removing the words.
- removing the examples after applying the feature-extraction window.

| | F-score(whole) | F-score(short)[17] |
|---|---|---|
| NB | 50.05 | 62.10 |
| C4.5 | 84.60 | 84.30 |
| SVM | 83.44 | 84.45 |

Table 5: Effect of reducing the examples using both words and lemmas in NER

The second strategy provides better results (about 3% better precision and recall) since there is richer information about the context of words.

[16] This is very important in methods like SVM since they are very sensitive to changes in the number of examples.

[17] The results are slightly better (about the 0,2%) because the number of correct entities are 935 but after the selection only 931 appear in test corpus.

Results on F-score and time are shown in Table 5 and 6. The former considers words and lemmas and the latter considers only lemmas.

Further conclusions can be extracted from these figures:

- The reduction of the training-set achieves good results with Naive Bayes and SVM. In the last case, the results are now similar to those obtained using C4.5. This behaviour might be due to the small size of the training corpus.
- The training time is drastically reduced, and, since time is a limiting factor for experiments, this can be considered an interesting feature for SVM.

These results are still 3% poorer than the ones obtained with Abionet, the winner in CoNLL02 for Spanish and Dutch.

| | F-score(whole) | F-score(short) |
|---|---|---|
| NB | 62.36 | 69.68 |
| C4.5 | 84.94 | 84.74 |
| SVM | 83.38 | 84.36 |

Table 6: Effect of reducing the examples using only lemmas in NER

### 4.1.5 Combining methods for the identification

Combining methods is another strategy to improve the results. We had the option of combining the same method with different features and training sets, or combining different kinds of methods.

| | Precision | Recall | F-score |
|---|---|---|---|
| Sydney | 74.74 | 77.54 | 76.12 |
| C4.5 | 86.74 | 82.57 | 84.60 |
| C4.5-lemma | 85.73 | 84.17 | 84.94 |
| Eihera | 84.65 | 86.10 | 85.37 |
| Abionet | 89.81 | 85.78 | **87.75** |

Table 7: NER results of each method

The last choice is the most attractive because we already have three different kinds of methods that offer quite good results:

- A Language independent method (Sydney).
- A Linguistic method (Eihera).
- ML based methods (C4.5 and SVM from Weka, and AdaBoost from Abionet).

The results achieved with each method are shown in Table 7. We also tested different combinations of methods using simple voting. The most important results are described in Table 8.

| | Pr. | Rc. | F-sc |
|---|---|---|---|
| Syd.,C4.5,C4.5-lem | 85.59 | 84.28 | 85.42 |
| C4.5,C4.5-lem,Eih. | 86.65 | 84.71 | 85.67 |
| Syd.,C4.5-lem,Eih. | 88.71 | **87.38** | 88.04 |
| Syd.,C4.5,Eih. | 89.49 | **87.38** | **88.42** |
| Abio.,Syd.,C4.5 | **90.72** | 85.78 | 88.18 |
| Abio.,Syd.,Eih. | 90.08 | **87.38** | 88.71 |
| Abio.,C4.5,Eih. | 90.47 | 87.27 | **88.84** |

Table 8: NER results combining methods using voting

These results confirm the hypothesis that results can be improved when combining different kinds of recognizers. We get a slight improvement when combining the same kind of recognizers; however, combining the three kinds of methods (Sydney, Eihera and C4.5) more improvement and better results than the state of the art single method (Abionet) are obtained (88.42 vs. 87.75).

All results increase when Eihera is added, which confirms that exploiting linguistic information in a specific way, as using rules, helps identifying some structures that machine learning algortihms are not able to capture. As it was expected, the best results are obtained when Abionet is included in the combination, by joining the three methods that produced the best results separately. However, the best recall (87.38) and the best precision (90.72) are obtained with Sydney, although this gets worse results when tested individually. This means there is a small overlapping in the decisions taken by Sydney identification system with regard to the other systems.

## 4.2 Classification

When we started working in the NEC task, we knew which were the best performing methods for NER. So only the best performing methods in NER will be used for NE classification task.

We also used new information sources and these have been useful as we will see in the evaluation.

### 4.2.1 Features

We have extracted some attributes based on the heuristic of Eihera. We have word, lemma, PoS, and declension case of each component of the entity as internal evidence. Those attributes are also extracted from the words surrounding the entity (a window of [-2,+2] in the experiments) in order to use contextual information. We used the information obtained from different gazetteers and a list of trigger words as external evidence.

While in the linguistic tool we have only used the gazetteers borrowed from the Euskaldunon Egunkaria newspaper (the PER list enriched with the census of the local government) and a small list of trigger words, we have used the ones enriched with WordNet information for ML experiments too.

To sum up, the number of attributes rises from about 80 to 120 when WordNet information is applied, because various attributes have to be added in the input for each consulted list in order to represent that external evidence. And so as number of lists increase, attributes increase linearly.

### 4.2.2 Methods

The methods we have tested for the classification task are C4.5 and SVM from Weka and AdaBoost implementation of Abionet.

Comparing NEC experiments with NER ones, C4.5 and SVM work better than AdaBoost, which was the state of the art for identification task. The results are shown in Table 9, and the main conclusion we can draw from them is that the ORG category is the most difficult to classify. Comparing them to LOC category the difference is approximately of 10% F-score.

| | ORG | LOC | PER | ALL |
|---|---|---|---|---|
| C4.5 | 71.25 | 81.22 | 78.82 | 75.93 |
| SVM | 69.18 | 79.81 | 76.95 | 74.43 |
| Abio. | 68.82 | 77.61 | 76.55 | 72.93 |
| Eih. | 67.28 | 76.49 | 77.68 | 72.93 |
| C4.5+WN | 71.66 | 82.70 | 79.01 | **76.69** |
| SVM+WN | 69.43 | 78.83 | 77.65 | 74.43 |
| Abio.+WN | 65.65 | 77.11 | 76.76 | 71.75 |

Table 9: NEC results in F-score measure for each category

The classification of all categories was better When we applied WordNet knowledge in C4.5. In the case of SVM the performance was exactly the same on average but not for each category. In contrast to both, Abionet worked worse. But doing an exhaustive analysis for each category we saw that the results obtained for PER and LOC are better or similar using WordNet, respectively, while for ORG the performance decreases significantly.

We can verify in these results that Eihera is again close to the rest of sophisticated methods used.

### 4.2.3 Selection of features

On some occasions the results can be improved selecting appropriate features. However, this time we have applied this technique not only with that purpose, but also to determine which are the most relevant features for the NEC task.

We experimented omitting gazetteer and/or trigger words information with C4.5 and Adaboost methods (results are shown in Table 10) and we saw that the results decrease significantly when the omission of gazetteers was applied, while only a slight decrease happened when omitting trigger words.

If we focus on each possible classification category, and we test methods without gazetteer information results of LOC entities remain similar, but the performance of PER and ORG categories decreases in 10% for C4.5 and in 2% for SVM.

Thus, we can conclude that trigger words are helpful for the NEC task but they are not crucial, while information of gazetteers is very relevant. Their influence would be more or less evident depending on the robustness of the method. Looking at our results it is clear that Abionet is more robust than C4.5 in this aspect.

| | ORG | LOC | PER | ALL |
|---|---|---|---|---|
| C4.5 | 71.25 | 81.22 | 78.82 | 75.93 |
| C4.5-tr | 70.95 | 81.08 | 78.89 | 75.83 |
| C4.5-gz | 61.02 | 75.41 | 68.35 | 67.13 |
| Abio. | 68.82 | 77.61 | 76.55 | 72.93 |
| Abio.-tr | 67.69 | 76.99 | 76.50 | 72.28 |
| Abio.-tr-gz | 65.53 | 76.99 | 73.11 | 70.67 |

Table 10: NEC results with selection of features in F-score measure

### 4.2.4 Combining methods

We have seen in the NER task that results improve when combining methods, so we applied the same approach in this task.

We used different strategies for doing the combinations using classifiers obtained from Eihera, C4.5, SVM and Abionet. In all experiments we applied 3 different classifiers and a simple voting decision.

Firstly, we combined different classifiers constructed with the same method but using different information sources (trigger words/gazetteers/WordNet). The result is exactly the same obtained with the best single classifier used in this combination.

Then, we combined classifiers obtained from different methods. In two experiments (rows 2 and 3 in Table 11) we applied models constructed using similar information sources and another one using information from feature selection (4th row in Table 11).

When we combine C4.5 using WordNet information with SVM and Abionet, the results are the same as the ones obtained when only C.45+WN is applied. Nevertheless, when combining that classifier with the linguistic tool and Abionet (rows 3 and 4) the performance increases.

Observing the results obtained combining C4.5+WN and Eihera with different Abionet classifiers (with or without external information), we can see that omitting external evidence is still reflected on lower results. In those experiments LOC and PER results remain similar while ORG performance decreases. But in this case the difference is lower than in the results of simple methods (2% loss vs 0.75% loss). Table 11 describes the evaluation in detail.

| | Precision |
|---|---|
| C4.5+WN/C4.5-tr/C4.5 | 76.69 |
| C4.5+WN/SVM/Abio. | 76.69 |
| Abionet/Eihera/C4.5+WN | 78.73 |
| Abio.-tr-gz/Eih./C4.5+WN | 77.98 |

Table 11: NEC results combining methods

We wanted to know which was the influence of Eihera in those experiments. For that purpose we chose the best combination (3rd row in Table 11), we tested it without Eihera and we achieved 5% lower results.

In order to find an explanation to this performance, we analysed information used in each system. On the one hand, Eihera uses explicitly the syntactic structure of the whole entity based on morphological information.On the other hand, ML methods are provided with the same morphological information, but it seems that they are not able to extract that structure. So the improvement caused by adding Eihera might be oriented by the information provided by this syntactic structure, as it was mentioned for identification.

## 5 NERC system

We have presented NER and NEC tasks as separate works, but if we want a Named Entities recognizer, we have to apply both. In this section we will present the results we obtained when we apply NER and NEC sequentially with different methods and combinations.

Note that if we apply NEC in the output of NER task, classifiers will treat not only the correctly identified entities, but also some uncompleted expressions and some others which are captured as entities although they are not. So the performance goes down both in precision and recall.

| NERC =>NER\|NEC | Pr. | Rc. | F-sc. |
|---|---|---|---|
| Eih.\|Eih. | 62.14 | 63.48 | 62.80 |
| C4.5\|C4.5+WN | 67.97 | 64.98 | 66.44 |
| Abio.\|Abio. | 66.62 | 63.90 | 65.24 |
| Abio.\|C4.5+WN | 71.19 | 68.20 | 69.62 |
| Syd.+C4.5+Eih.\|<br>Abio.+Eih.+C4.5+WN | 71.69 | 69.92 | 70.79 |
| Abio.+Syd.+C4.5\|<br>Abio.+Eih.+C4.5+WN | 73.07 | 69.38 | 71.18 |
| Abio.+C4.5+Eih.\|<br>Abio.+Eih.+C4.5+WN | 72.50 | 70.24 | 71.35 |

Table 12: NERC experiments

We can distinguish two different kinds of NERC experiments: some systems constructed applying simple methods for both NER and NEC tasks, and others obtained combining different classifiers.

The results of Eihera in Table 12 reveal that the information used is not enough to construct a robust NERC system. Applying the best classifiers obtained from Abionet and C4.5 independently the performance increases. We did not only achieve better results using one for NER and the other for NEC, but we can also consider them the baseline.

In order to continue improving our results we tried constructing more complex systems. For that purpose we chose the best classifiers constructed for each task and the performance increased in almost 8% with regard to Eihera, and in 5% with the rest. We made a detailed analysis based on categories (see Table 13 for the results of the best combination) and it was confirmed that ORG is the most difficult category for recognition, with

10% lower performance than the others. So in the future new information sources must be studied and used in order to correct this phenomenon and construct a robust system for all categories.

| | Precision | Recall | F-score |
|---|---|---|---|
| PER | 77.90 | 73.38 | 75.13 |
| LOC | 75.83 | 79.68 | 77.71 |
| ORG | 65.49 | 63.48 | 64.47 |

Table 13: The results of the best combination for each category(Abionet, C4.5 and Eihera for NER, and Abionet, Eihera and C4.5+WN for NEC)

We have said that Eihera's information is not good enough for a robust NERC system but note that the best results are obtained by using the linguistic tool in NER and NEC combinations. So, although not enough by itself, the information is very useful.

In order to compare Basque with other laguanges, we will take Abionet as reference not only because its performance in Basque is familiar for us, but also because this system is one of the best systems presented in both CoNLL02[18] and CoNLL03[19] competitions. In fact, it was the winner in CoNLL02 for Spanish and Dutch, while in CoNLL03 it got the 5th position for English and German.

We achieved a 65.24% F-score applying Abionet for NERC in Basque, while using Eihera we reached a 62.8% F-score. So there is a difference of around 2.5% between both systems

The state of the art in CoNLL03 (Florian *et al.*, 2003) combines 4 classifiers, obtaining a 3% better performance than Abionet for English and German.

We used 3 different methods for the experiments we carried out for Basque. In this case, we reached around 5% better performance than Abionet.

So, although we have not applied some of the state of the art methods of the CoNLL03 competition, and considering Abionet a good reference, we can say that our results for NERC task with Basque are comparable with the ones obtained with languages others than English, and that it seems we are following the right way.

## 6 Conclusions and Future Work

We have presented the methodology and the results of a NE recognizer for Basque. The contribution of this work can be divided into two main areas: on the one hand it can be used for information extraction and retrieval in Basque written texts, and more concretly for those who work with Basque named entities recognition; and on the other hand it can be useful for named entities recognition in general.

We tried different feature sets for Basque NE recognition, and although we cannot say that we have found the ideal feature set for this task, we have found the very relevant ones while we avoided some others.

In general, we have shown that a linguistic tool can be very useful in NERC task at any language, because it allows us simplifying the annotation, learning about features and getting a good tool to be combined with ML-based methods for both identification and classification of named entities.

We want to emphasize the relevance of the linguistic features, among the ones used in our experiments. While lemmas are sufficient for NER, in NEC it is necessary to add also inflectional cases to obtain a good performance.

Observing the experiments done to select examples for NER, we can conclude that the selection can be useful for both speeding the training and improving the results.

We have also observed that the errors in the pre-process have a significant impact on the coverage of the recognizer. Thus, we want to lay stress on the improvement of the lemmatizer/tagger.

We have tried increasing the size of both trigger and gazetteer lists using WN. We have experimented the different methods and techniques with and without that enrichment. And observing the performances, we can conclude that WordNet seems to be very helpful for entity contextualization using trigger words list, but providing the whole entity classification with only a gazetteer access seems to be not so useful.

Anyway, in the near future we also want to approach different research lines in order to improve this recognizer. One of them these approaches is to measure the effect of the gazetteer size, trying different sizes starting from scratch, as proposed in (Mikheev *et al.*, 1999).

[18] http://cnts.uia.ac.be/conll2002/

[19] http://cnts.uia.ac.be/conll2003/

About corpus, we think that the training corpus is not big enough to reach relevant conclusions about the performance of ML techniques applied to Basque. So we will also test the performance of the recognizer on different sizes of the training set.

A priori the test corpus we have used is quite small and it seems not to be a significant set. But, we used both separated train/test corpus methodology and cross-validation for our first machine learning experiments' evaluations, and we saw that two methodologies' results were very similar. So in spite of its small size, the test corpus seems to be a good sampling set to work with, and we can say that there is no overfitting on the training set. On the other hand, since results reveal that only test and cross-validation methodologies are equivalent and have the same significance for our test set, we decided to only apply test corpus evaluation methodology for further evaluations.

Finally, we want to improve the individual performance of each entity class. For that purpose we will try to develop class-expert-classifiers, selecting appropriate feature sets and learning methods for each one. Another approach we want to experiment is the use of more complex voting methods.

In a further future, we want to have a multilingual named entity recognizer and we have already started doing first approaches. Taking Basque-English and Basque-Spanish comparable bilingual corpus as source, we are designing a system capable of translating entities based on (Al-Onaizan *et al.*, 2002) related work.

## *Acknowledgements*

The research of Iñaki Alegria was carried out during his visit to the University of Sydney supported by a grant from the Basque Government. We would like to thank Jon Patrick, Casey Whitelaw and the ScamSeek group for their support and helpful discussions, to Xavier Carreras and Lluis Màrquez for their help with Abionet and to Itsaso Esparza for her help writing the final version of the paper.

This research has been partially funded by the European Commision (MEANING IST-2001-34460) and the Basque Government (Saiotek-Ihardetsi and BFI04.432).

## *References*

Alegria I., Artola X., Sarasola K.(1996) Automatic morphological analysis of Basque *Literary & Linguistic Computing Vol. 11 No. 4, 193-203. Oxford University Press. Oxford. 1996*

Alegria I., Balza I., Ezeiza N., Fernandez I., Urizar R. (2003) Named Entity Recognition and Classification for texts in Basque. *Proceedings of II Jornadas de Tratamiento y Recuperación de Información 2003*

Al-Onaizan Y., Knight K. (2002) Translating Named Entities using monolingual and bilingual resources. *Proceedings of ACL-2002*

Beesley K.R., Karttunen L. (2003) Finite State Morphology. *CSLI*

Carreras X., Márquez L., Padró L., 2003 A Simple Named Entity Extractor Using AdaBoost. CoNLL-2003 Shared Task Contribution. *Proceedings of CoNLL-2003*

Chinchor N. (1998) Overview of MUC-7. *Proceedings of the 7th Message Understanding Conference (MUC-7)*

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. (1998) Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *COLING-ACL 1998*

Florian R., Ittycheriah A., Jing H. and Zang T., (2003) Named Entities Recognition through Classifier Combination. *Proceedings of CoNLL-2003*

Magnini B., Negri M., Prevete R., Tanev H. A. (2002) WordNet Approach to Names Entity Recognition. *Proceeding of the Workshop SemaNet 2002: Binding and Using Semantic Networks*

McDonald D. (1996) Internal and external evidence in the Identification and Semantic Categorization of Proper Names. *Corpus Processing for Lexical Acquisition (Boguraev and Pustejovsky, eds.).* The MIT Press, Massachusetts.

Mikheev A., Grover C., Moens M. (1998) Description of the LTG system used for MUC-7. *Proceeding of Message Understanding Conference (MUC-7)*

Mikheev A., Moens M., Grover C. (1999) Named Entity Entity recognizer without gazetteers. *Proceeding of EACL-1999*

Rennie J.D.M., Shih L., Teevan J., Karger D.R. (2003) Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Proceedings of the 20th Conference on Machine Learning*

Tjong Kim Sang E. (2002) Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002*

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

Whitelaw C. and Patrick J. (2003) Named Entity Recognition Using a Character-based Probabilistic Approach. *Proceedings of CoNLL-2003.*

Witten I.H. and Eibe F., 1999 Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufm