Named Entity Recognition and Classification for texts in Basque[1]

Alegria Loinaz, Iñaki; Balza Tardaguila, Irene; Ezeiza Ramos, Nerea;
Fernández Gonzalez, Izaskun; Urizar Enbeita, Ruben
IXA taldea. Euskal Herriko Unibertsitatea.
i.alegria@si.ehu.es

This paper presents a system for Named Entity (NE) recognition in written Basque to be used in a CLIR application. Being an agglutinative language, Basque has highly inflected forms, so a previous linguistic preprocess is required. The tool we present relies on a combined method that carries out the identification and recognition of entity names in two subsequent steps. First, a grammar based on morphological information is applied in order to extract the entity names of the text, and then, the identified entities are classified by applying a heuristic that combines contextual information and gazetteers.

Keywords: Named Entity Recognition, IR, IE.

1 INTRODUCTION

Named Entity (NE) recognition constitutes a very important aspect in Natural Language Understanding (NLP) and more specifically in the tasks of Information Extraction and Information Retrieval.

As defined in the *Message Understanding Conference* (MUC) [4], NE recognition consists in identifying and categorizing entity names (person, organization and location), temporal expressions (dates and times), and some types of numerical expressions (percentages, monetary values and so on), which are considered to constitute up to %10 of written texts [5].

According to [8], 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 the latter by the context in which it occurs.

Among the different techniques used to process these data, we find some systems based on statistical methods, such us Hidden Markov Models (HMM) [3], some based on strictly linguistic methods which make use of grammar rules [7], and finally the ones that combine rules and statistics [9].

Before applying these techniques, some previous work involving a more or less deep analysis of the written text is sometimes required. In the simplest cases, only tokenization is applied, but in other cases, also a morphological analysis, disambiguation, and the attachment of semantic features must be carried out.

The tool we present in this paper requires a complex previous process, due to the highly inflected forms of an agglutinative language like Basque, and relies on a combined method that realizes the identification and recognition of entity names for Basque in two subsequent steps. First, we apply a grammar based on morphological information to extract the entity names from the text, and then the identified entities are classified by applying a heuristic that combines context information and gazetteers. The tool is used for Cross-Language Information Retrieval in the Hermes project.

The paper is structured as follows: Section 1 presents our system and its goals. Section 2 deals with the grammar used for the recognition of entity names. In section 3 we describe the classification process. Section 4 shows the results achieved and section 5 presents some conclusions.

1.1 Aims

Many systems consider only proper nouns as entity names. But in fact they can be much more complex. For instance, in the case of "*LABeko Lan Osasuneko arduraduna*"[2] a typical system could extract *LABeko* as an ORGANIZATION entity when in fact the entity involved in the described example is a PERSON entity.

In the HERMES project, as pointed out in [1], we distinguish two classes of NEs: On the one hand, there are Strong NEs, which consist of a single proper noun (*Europako Banku Zentrala*[3]) and, on the other hand, Weak NEs which combine a Strong NE with others and/or

with a trigger word (*Win Duisenberg Europako Banku Zentraleko lehendakaria*[4]).

The priority goal of our system is to capture strong NEs. Numerical and temporal expressions are captured by the lemmatizer/tagger [6] used in the preprocess, which performs the tokenization, the morphosyntactic analysis and the disambiguation of the text. Based on this output entity names are treated by the other modules of the system.

As weak NEs contain strong ones, first we decided to achieve the latter, leaving the treatment of the former for the future.

We have developed a first prototype that provides a corpus for a semi-automatic process. After manual correction, the corpus has been used for the evaluation of the prototype and, in the future, it will be the source for a system based on machine learning.

1.2 Design.

Our system has a modular design that has two main modules: one performs the identification of entity names, and the other one classifies them. These two modules are sequentially executed.

The first module consists of a grammar whose rules are based on the morphological information of text provided by the lemmatizer/tagger.

The classifying module applies a heuristic that combines linguistic information, contextual information such as trigger words, and gazetteers.

Finally, combining the outputs of the previous modules we obtain an XML document in which the entities are marked with an special tag.

The architecture of our system is summarized in Figure 1:

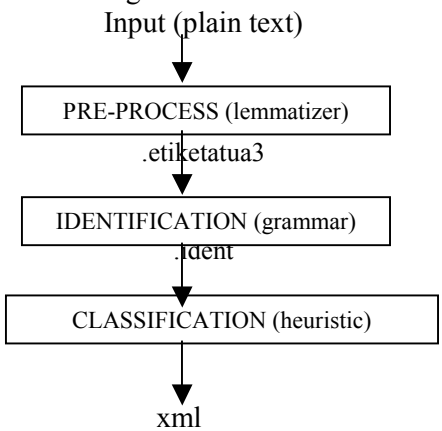

Figure 1: The architecture of the system

2 THE GRAMMAR

The tool used to develop the grammar for the identification of entities is XFST (*Xerox Finite State Transducer*) [2]. XFST permits us to define both the structure of entity names and the rules for their identification.

2.1 Main elements

Among the main elements of our grammar we find entity names and trigger words. Although the latter are not relevant for the identification of the strong entities, they are helpful for their classification.

The main feature of all the entity elements in Basque written texts, as in many other languages, is the use of capital letters. But, apart from this restriction, there are others that should be taken into account in Basque, for instance the PoS and subcategory of the elements and their inflection.

The main Parts of speech/subcategories we must distinguish for entity elements are the following: IZE (common noun), IZB (proper noun), LIB (location/organization proper noun), ADJ (adjective), SIG (acronym) and BST (particle[5]). Except for the case of some BST, the rest of the elements in the entity must be written in capitals.

For the identification of entities we make a distinction between non-case elements, genitive, and others.

When identifying trigger words, we must specify whether they occur before or after an entity name. These trigger words are also restricted to a certain type of PoS and can bear some specific inflection.

2.2 Main patterns

In the grammar for identification we distinguish between two patterns of entity names: entities containing a single element (*Europan* LOCATION) and entities composed of more than one element (*Europako Banku Zentralean* ORGANIZATION).

In the first case (Figure 2), the element PoS assigned by the lemmatizer/tagger must be SIG (acronym), IZB (proper noun) or LIB (location/organization proper noun). In case the element is declined, it can bear any case.

```
define PAT1 [TokenSIG | TokenIZB | TokenLIB];
```

Figure 2: One-element entity

Examples of one-element entities are *EHUn (SIG+inesive)* or *Ibarretxeri (IZB+dative), Bilbora (LIB+adlative)*

The entities with more than one element have a more complex pattern (Figure 3). First, we must distinguish between the last element of the entity and the rest, since the latter have a more restricted declension (only genitive), while the former can appear in any case. In contrast with one-element entities, the elements contained in this second type, can belong to different parts of speech: IZE (common noun), IZB (proper noun), LIB (location/organization proper noun), ADJ (adjective), SIG (acronym) or BST (particle).

```
define PAT2 [TokenLEFT [[TokenMID]* TokenLEFT]*
 [[TokenMID]* TokenLEFT | TokenLAST]];
```

Figure 3: More-than-one-element entity

The meaning of the tags in Figure 3 is the following: `TokenLAST` represents the last element in the entity, `TokenMID` stands for BST PoS elements and `TokenLEFT` represents the rest of the elements.

Examples of the second pattern are:

- *Europako* (LIB+GENITIVE) *Banku* (IZE) *Zentralean* (ADJ+INESIVE)
- *Alex* (IZB) *de* (BST) *la* (BST) *Iglesiak* (IZB+ERGATIVE)

In any case, the grammar captures and puts into brackets the longest sequence of possible entity elements that matches any of the patterns defined above.

Once the grammar has identified the entity expressions and bracketed them, it identifies as trigger words the forms that occur immediately before or after the entity, provided that they meet the following restrictions:

- They must be nouns (IZE).
- When they occur at the left of an entity, they cannot be declined.
- When they occur at the right, there is no restriction on declension, but they are only considered if the last element of the entity is not inflected.

Therefore, the components of the output of this module are the following: the form and lemma of the identified entity and its corresponding trigger words (if any).

3 THE CLASSIFICATION PROCESS

For the classification of the already identified entity names, we apply a heuristic in which different information sources are used.

On the one hand, we borrowed some gazetteers for different categories (PERSON, ORGANIZATION and LOCATION) from *Euskaldunon Egunkaria*, the only newspaper written entirely in Basque. The gazetteer for PERSON entities was enriched with information taken from the census of the local government.

On the other hand, we use lists containing trigger words and information on the type of entity normally associated to them. There is one list for trigger words occurring before the entity and another for those occurring after.

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.

Step 2:

The heuristic selects one by one the elements in the entity: 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 categories. In case there is any genitive among the elements, a further analysis is applied. For instance, in *Europako Banku Zentralean*[6], *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 won't be relevant for the weight assignment.

Step 3:

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

Step 4:

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

4 EVALUATION

In order to evaluate the performance of our system, we have compared the automatically

tagged corpora and the hand tagged corpus mentioned before.

The evaluation corpus consists of 383 articles of different sections published in *Euskaldunon Egunkaria* newspaper, in which 7550 entities have been hand tagged.

Since we have distinguished between identification and classification in NE recognition, we have made the same distinction in the evaluation process.

In order to assess NE identification, both precision and recall have been measured, whereas for NE classification only the precision parameter has been considered, since its output depends on the results of the identification module (no wrongly identified entity can be correctly classified).

However, it would be interesting to assess the performance of the system as a whole. For that purpose, we have compared the hand tagged entities with the ones correctly identified and classified by the system (identification's recall * classification's precision).

	well identified	hand-tagged	%
Id_recall	860	1051	83.73

Table 1: recall parameter

	well treated	automatically recognized	%
Id_precision	880	1114	78.99
Class_precision	716	880	81.36

Table 2: precision parameter

	well identified & classified	hand-tagged	%
Total_measure	716	1051	68.13

Table 3: total recall

The results achieved, including recall and precision for the identification process, and precision for classification, are shown in Tables 1, 2, and 3.

5 CONCLUSIONS

As it can be concluded from the evaluation data above, results are quite good in classification but worse in the identification process.

Most of the errors in the classification task are basically mixing up Place and Organization categories and not detecting miscellaneous entities. This last problem can be relaxed by detecting titles of books and films.

With regard to the problems in the identification process, we have examined the reason for the errors in 100 NEs. As shown in table 4, most of the errors are due to reasons external to the developed system.

Reason	**Percentage**
Errors in capital letters	35 %
Bad analyses in preprocess	29 %
Errors in the input format	22 %
Weak NEs	8 %
Others	6 %

Table 4: Source of errors in identification

Let's examine the different kind errors:

- *Errors in capital letters*: an element of the entity name was not capitalized. This is a difficult problem to solve.
- *Bad analyses in the preprocess*: Most of the errors made in Person and Place names are due to the great number of analyses the guesser module of the tagger generates for words not included in the lexicon. We are currently working to improve the tagging of these elements.
- *Errors in the input format*: The corpus was converted to HTML from Quark. Some surface errors were produced in this process, for example, sometimes, new line characters disappeared. Most of these errors were automatically corrected but some still remain.
- *Weak NEs*: Although the grammar tries to identify strong NEs, sometimes, complex ones are detected instead. This is not an easy problem to solve since any changes in the grammar can cause to exclude correct identifications.

To sum up, we can say that half of the identification errors are due to reasons external to the system and so they would not occur in accurately written texts. Most of the remaining ones could be corrected if the tagger was improved.

In the future, we intend to use the corpus that we have produced in a semiautomatic way as a source for a system based on machine learning. Results might be improved combining both methods.

An example of the output of our system is presented in Appendix 1.

6 REFERENCES

[1] ARÉVALO M., CARRERAS X., MÁRQUEZ L., MARTÍ M.A., PADRÓ L., SIMÓN M.J. (2002): A Proposal for Wide-Coverage Spanish Named Entity Recognition. SEPLN 28.

[2] BEESLEY K.R., KARTTUNEN L. (2001): *Finite State Morphology: Xerox Tools and Techniques*.

[3] BIKEL D., MILLER S., SWATCH R., WEISCHEDEL R. (1999): An Algorithm that Learns What's in a Name. *Machine Learning: Special Issue on Natural Language Learning,* 34.

[4] CHINCHOR N. (1998): Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7).*

[5] COATES-STEPHENS S. (1992): *The Analysis and Acquisition of Proper Names for Robust Text Understanding.* PhD thesis, Department of Computer Science, City University, London.

[6] 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'98, Montreal (Canada).

[7] 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.*

[8] 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.

[9] MIKHEEV A., GROVER C., MOENS M. (1998): Description f the LTG system used for MUC-7. *Proceeding of Message Understanding Conference (MUC-7).*

APPENDIX 1

```
<MW NETYPE="STRONG" FRM="Mikhail_Gorbatxov">
<CAT SCHEME="HERMES-MUC" CODE="PERSON" />
<LEX LEM="Mikhail_Gorbatxov"> </LEX>
  <W FRM="Mikhail">
      <LEX LEM="Mikhail" PAR="IZEIZB"> </LEX>
  </W>
  <W FRM="Gorbatxov">
      <LEX LEM="Gorbatxov" PAR="IZEIZB"> </LEX>
  </W>
</MW>
<MW NETYPE="STRONG" FRM="Sobiet_Batasuneko">
<CAT SCHEME="HERMES-MUC" CODE="LOCATION" />
<LEX LEM="Sobiet_Batasun"> </LEX>
  <W FRM="Sobiet_Batasuneko">
      <LEX LEM="Sobiet_Batasun" PAR="IZELIB"> </LEX>
  </W>
</MW>
<W FRM="presidenteak">
      <LEX LEM="presidente" PAR="IZEARR"> </LEX>
</W>
<MW NETYPE="STRONG" FRM="Moskuko_Alderdi_Komunistako">
<CAT SCHEME="HERMES-MUC" CODE="LOCATION" />
<LEX LEM="Moskuko_Alderdi_Komunista"> </LEX>
  <W FRM="Moskuko">
      <LEX LEM="Mosku" PAR="IZELIB"> </LEX>
  </W>
  <W FRM="Alderdi">
      <LEX LEM="alderdi" PAR="IZEARR"> </LEX>
  </W>
  <W FRM="Komunistako">
      <LEX LEM="komunista" PAR="ADJIZO"> </LEX>
  </W>
```

```
</MW>
<W FRM="idazkari">
      <LEX LEM="idazkari" PAR="IZEARR"> </LEX>
</W>
<W FRM="izendatu">
      <LEX LEM="izendatu" PAR="ADI"> </LEX>
</W>
<W FRM="zuenean">
      <LEX LEM="edun" PAR="ADL"> </LEX>
</W>
```

---

[1] This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-CO3-02) and by the Basque Government (UE02/B11).

[2] *The responsible for Health at Work from LAB (syndicate)*

[3] *European Central Bank*

[4] *The president of the European Central Bank Win Duisenberg*

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

[6] *In the European Central Bank*