Named Entities Translation Based on Comparable Corpora

Iñaki Alegria
IXA NLP Group
EHU
Donostia, Basque Country
i.alegria@ehu.es

Nerea Ezeiza
IXA NLP Group
EHU
Donostia, Basque Country
n.ezeiza@ehu.es

Izaskun Fernandez
IXA NLP Group
EHU
Donostia, Basque Country
acbfegoi@si.ehu.es

Abstract

In this paper we present a system for translating named entities from Basque to Spanish based on comparable corpora. For that purpose we have tried two approaches: one based on Basque linguistic features, and a language-independent tool. For both tools we have used Basque-Spanish comparable corpora, a bilingual dictionary and the web as resources.

1 Introduction

Person, location and organization names, main types of named entities (NEs), are expressions commonly used in all kinds of written texts. Recently, these expressions have become indispensable units of information for many applications in the area of information extraction as well as for many searching engines. A lot of tools that deal with the identification and classification of named entities for a specific language have been presented (CoNLL[1]). But there are few researches for translation of NEs.

Our main goal is to get a multilingual NE database, which can be very useful for translation systems, multilingual information extraction tools (i.e. Question Answering) or many multilingual systems in general. As getting that multilingual source is a complex task, we have started designing a system for translating named entities from Basque to Spanish based on comparable corpora.

Looking at the works published on NE translation, we can distinguish 3 types of systems: the systems more often used are the ones based on parallel corpora; then the ones based on comparable corpora; and finally the ones that only use the web as an open corpus.

As we have mentioned before, most of the related works use parallel corpora. However and as it is widely known, obtaining parallel corpus is not an easy task, and it becomes harder when one of the languages in the pair is a minority language, as is the case of Basque. We can avoid working with parallel corpora using comparable corpora. Comparable corpora are those data sets which are written in different languages, treat similar subjects and are written in a similar style, but are not necessarily texts' translations. Obtaining that kind of corpora is much easier than obtaining parallel one, although sometimes it is not possible to get neither of them. In this case, we can use the web as a multilingual corpus, in order to search it for any possible entity translation.

We have a comparable data set available for Basque and Spanish. But besides using that data source, we decided also to resort to the web as a complementary data set too, as in (Moore, 2003).

Apart from these two data sets, we have also used some other information sources to develop the Basque-Spanish bilingual NE translation system. We have carried out two main different experiments: one using a language-dependent grammar, implementing transliteration transformations (Al-Onaizan *et al.*, 2002b) and rules related to elements' order; and another one based on the edition distance (Kukich, 1992) grammar, simulating simple cognates and transliteration transformations, but in a language-independent way. In both experiments, we have used a Basque-Spanish bilingual dictionary for the words in which transliteration transformations were not enough to obtain the correct translated form.

Furthermore, we have always worked using

[1] http://www.cnts.ua.ac.be/conll2003/ner/

Basque as source language, and Spanish as target language.

Since Basque and Spanish do not follow the same syntactic pattern, entity elements may occur in different positions in both languages. That is why the elements need to be arranged when translating Basque entities into Spanish.

The paper is structured as follows. Section 2 presents the related works. Section 3 presents the experimental settings. In section 4 we describe the development of NE translation system explaining both possible systems, the language-dependent system and the language-independent one, and the system that combines both language-dependent and independent sources. In section 5, we present the results of the experiments, and finally, section 6 presents some conclusions and future works.

2 Related Works

Despite the difficulty of getting bilingual parallel corpus, most of the NE translation researches carried out work with parallel data-sets. Furthermore, those bilingual corpora are used to be aligned at paragraph or even at phrase level. For example, Moore's work (Moore, 2003) uses a bilingual parallel aligned English-French corpora, and applying different statistical techniques, he obtains a French form for each English entity.

Although it has been less experimented with comparable corpora there are some known systems designed to work with them as well. Most of them deal with language pairs that have different kinds of alphabets. For instance, the Chinese-English translation tool presented in ACL 2003 (Chen *et al.*, 2003), or the one published in the ACL 2002 edition for translating entity names from Arabic to English (Al-Onaizan *et al.*, 2002a). The main goal of both systems is to obtain the corresponding form for English, taking Chinese and Arabic respectively as source languages. Two kinds of translations can be distinguished in both systems: direct/simple translations and transliterations (Al-Onaizan *et al.*, 2002b). However, the techniques used by each tool for both kinds of translations are different. Frequency based methods are used in Chinese-English translations, while in the Arabic-English language pair, a more complex process is applied, which involves the combination of different kinds of techniques.

In this paper, we present the research carried out for translating entity names from Basque into Spanish. For the first step, we have based on the system presented by Y. Al Onaizan and K. Knight in ACL 2002. With this system, they first obtain a candidate translation list for the entity in the target language, using both monolingual and bilingual resources. Once they have this list, they build a ranking with candidates applying different methods (such as statistical measures, web-counting, etc.). Finally, if they consider that the correct translation does not appear in the list, they extract an extended list version using the web and they apply again the ranking step.

3 Experimental settings

We have obtained a Basque-Spanish comparable corpora processing news from two newspapers, one for each language: *Euskaldunon Egunkaria*, the only newspaper written entirely in Basque for Basque texts, and *EFE* for Spanish texts. We have collected the articles written in the 2002 year in both newspapers and we have obtained 40,648 articles with 9,655,559 words for Basque and 16,914 with 5,192,567 words for Spanish. Both newspapers deal with similar topics: international news, sports, politics, economy, culture, local issues and opinion articles, but with different scope.

In order to extract Basque NEs, we have used *Eihera* (Alegria *et al.*, 2003), a Basque NE recognizer developed in the IXA Group. Giving a written text in Basque as input, this tool applies a grammar based on linguistic features in order to identify the entities in the text. For the classification of the identified expressions, we use a heuristic that combines both internal and external evidence. We labeled this corpus for the HERMES project[2](news databases: cross-lingual information retrieval and semantic extraction). Thus, we obtained automatically 142,464 different person, location and organization names.

Since we have participated at the HERMES project, we have available labeled corpora for the other languages processed by other participants. It was the TALP[3] research group the one that was in charge of labeling *EFE 2002* newspaper's articles for the Spanish version, in which 106,473 different named entities were dealt with. We have built the comparable corpus using this data-set together with the Basque set mentioned above.

[2] http://nlp.uned.es/hermes/

[3] http://www.lsi.upc.edu/ nlp/web/

Being Basque an agglutinative language, entity elements may contain more than just lexical information. So before doing any translation attempt a morphosyntactic analysis is required in order to obtain all the information from each element. Furthermore, Eihera works on a lemmatized text, so lematizing the input text is a strong requirement. For that purpose, we apply the lemmatizer/tagger for Basque (Alegria *et al.*, 1998) developed by the IXA group.

The goal of our system is to translate Basque person, location and organization names into Spanish entities. These two languages share a lot of cognates, that is, words that are similar in both languages and only have small, usually predictable spelling differences. Two experts have reviewed an extended list of word pairs[4] extracted from *EDBL (Basque Lexical Data-base)* in order to detect these differences. All the observed variations have been listed in a spelling-rule list. These rules are in fact the ones that will be applied for the translation of some of the words, but obviously not for all.

When translating Basque words into Spanish, usually the correct form is not obtained by applying the rules mentioned before, and a different strategy is required. For these words in particular, we have used bilingual dictionaries as in Al-Onaizan and Knight's work.

We have used the *Elhuyar 2000* bilingual dictionary, one of the most popular for that language pair. This dictionary has 74,331 Basque entries, and it contains the corresponding Spanish synonyms.

For the evaluation, we have used a set of 180 named entity-pairs. We have borrowed that set from the *Euskaldunon Egunkaria 2002* newspaper. Concretely we applied *Eihera*, the Basque NE recognizer, to extract all the named entities in the corpus. Then we estimated the normalized frequency of each entity in the corpus, and we selected the most common ones. Finally we translated them manually into Spanish.

In order to carry out an evaluation starting from correct Basque NEs, although the NEs were automatically extracted from the corpus, we verified that all the entities were correctly identified. Because if the original entity was not a correct expression, the translation system could not get a correct translation.

4 Systems' Development

As we have mentioned before, we have done two different experiments in order to get a Basque-Spanish NE translation tool. For both trials we have used bilingual dictionaries and grammars to translate and transliterate entity elements, respectively. But the methodologies used to implement each transliteration grammar are different: on the one hand, we have used Basque linguistic knowledge to develop the grammar; on the other hand, we have defined a language-independent grammar based on edition distance.

Those dictionaries and grammars have been used in order to obtain translation proposals for each entity element. But another methodology is needed for the system to propose the translation of whole entities. For the system based on linguistic information, a specific arranging rule set has been applied getting a candidate list. In order to decide which is the most suitable one, we have created a ranked list based on a simple web count.

For the language-independent system a more simple methodology has been applied. We have generated all the possible candidate combinations, considering that every element can appear at any position in the entity. Then, a comparable corpus has been used in order to decide which is the most probable candidate.

Now we will present the design of each experiment in detail.

4.1 Linguistic Tool

We can see the pseudo-code of the linguistic tool at Figure 1.

```
Input: Basque NE composed of n element.
Output: An ordered list of Spanish translations
Initialize: $proposals_i$ (i=1..n)
foreach $element_i$ ; i=1..n do
  $proposals_i$ = Search $element_i$ in a bilingual dictionary
  if (!$proposals_i$) then
     $proposals_i$ = Apply transliteration grammar to $element_i$
  end if
end for
entity_translations = entity construction ($proposals_{1..n}$)
arranged_entity_translations = Web search
```

Figure 1: Linguistic Tool

The linguistic tool, first tries to obtain a translation proposal for each entity element using bilingual dictionaries. If no candidate is obtained from

[4] One expert has revised adjective and nouns in general, while the other one has only treated proper noun pairs

that search, the transliteration grammar is applied. Once the system has obtained at least one proposal for each element, the arranging grammar is applied, and finally, the resultant entire entity proposals are ranked based on their occurrence on the web.

4.1.1 Transliteration

Reviewing the extended list of words from *EDBL* (a Basque Lexical Data-base) we have obtained 24 common phonologic/spelling transformations, some of which depend on others, and can usually be used together, although not always. We have implemented these 24 transformations using the *XFST* (Beesley and Karttunen, 2003) tool and we have defined 30 rules. These rules have been ordered in such a way that rules with possible interactions are firstly applied and then the rest of them. This way we have avoided interaction problems.

For instance, lets say that we want to translate *Kolonbia* into *Colombia* and that our grammar has the following two simple transformation rules: *nb* → *mb* and *b* → *v*. If we apply the first rule and then the second one, the candidate we will obtain is *Colomvia*, and this is not the correct translation. However, if we do not allow to apply the second rule after the *nb* → *mb* transformation, the grammar will propose the following candidates: *Colonvia* and *Colombia*. So it would generate bad forms but the correct forms too.

We can conclude from this fact that it is necessary to apply the rules in a given order.

The possible combinations of rules are so wide that it causes an overgeneration of candidates. To avoid working with such a big number of candidates in the following steps, we have decided to rank and select candidates using some kind of measure.

We have estimated rules probabilities using the bilingual dictionary *Elhuyar 2000*. We have simply apply all possible rule combinations on every Basque word in the dictionary, and measured the normalized frequency of each rule and each rule pair. Thus, translation proposals are attached a probability based on the probability of a rule being applied, and only the most probable ones are proposed for the following steps.

4.1.2 Entire Entity Construction

At this point, we have N translation candidates for each input entity element at the most, and they have been obtained applying the grammar or from the dictionary search. Our next goal is to create entire entity translation proposals combining all these candidates. But some words features, such as gender and number, must be considered and treated beforehand.

The number of an entity element will be reflected in the whole entity. Let's say, for instance, translate the organization name *Nazio Batuak*[5]. The translation proposals from the previous modules for these two words are *Nación* (for *Nazio*) and *Unida* (for *Batuak*). If we do not consider that the corresponding Basque word of the *Unida* element is in the plural form, then the whole translation candidate will not be correct. In this case, we will need to pluralize the corresponding Spanish words.

Unlike Spanish, Basque has no morphological gender. This means that for some Basque words the generation of both male and female form is required. The word *idazkari*, for example, has no morphological gender, and it has two corresponding Spanish words: the masculine *secretario* and the feminine *secretaria*. If we search for *idazkari* on the bilingual dictionary, we will only obtain the masculine form, but the feminine form is needed for some entities , as it is the case with *Janet Reno Idazkaria*[6]. Since Janet Reno is a woman's proper name, the correct translation of *Idazkaria* would be *Secretaria*. So before constructing the entire entity translation, both male and female forms have been generated for each element.

The simplest entities to construct are the ones whose elements keep the same order in both the Basque and the Spanish forms. Person names usually follow this pattern.

However, there are some translations that are not as regular and easy to translate as the previous ones. Suppose that we want to translate the Basque entity *Lomeko Bake Akordio*[7] into the Spanish form *Acuerdo de Paz de Lome*. After applying grammar and bilingual dictionaries, we obtain the following translated elements (in order to simplify the explanation, we have assumed that the system will only return one translation candidate per element): *Lome Acuerdo* and *Paz*. As you can see, if we do not arrange those elements, the proposal will not be the appropriate Spanish transla-

[5] United Nations

[6] Secretary Janet Reno

[7] Lome Peace Agreement

tion.

An expert's manual work has been carried out in order to define the element arranging needed when turning from one language to the other. The morphosyntactic information of the Basque entity elements (such as PoS, declension, and so on) has been used in this task.

Using this manual work, we have defined 10 element-arranging rules using the *XFST* tool. In the example above, it is clear that some element-arranging rules are needed in order to obtain the correct translation. Let's see how our grammar's rules arranges those elements.

When the system starts arranging the *Lome Acuerdo* and *Paz* Spanish words to get the correct translation for the Basque named entity *Lomeko Bake Akordio* it starts from the right to the left using the Basque elements' morphosyntactic information. So it will start arranging the translated elements for *Bake Akordio* from right to left. Both forms are common nouns with no declension case. Looking at the grammar the system will find a rule for this structure that switches position of the elements and inserts the preposition *de* in between. So the partial translation would be *Acuerdo de Paz*. The next step is to find the correct position for the translation of *Lomeko*, which is a location name declined in genitive. There is a rule in the grammar, that places the elements declined in genitive at the end of the partial entity and adds the preposition *de* before this element. So, the system will apply that rule, obtaining the Spanish translation of the whole entity *Acuerdo de Paz de Lome*, which is the correct form.

4.1.3 Web Search

As we have explained, we combine at the most the N translation candidates per entity elements with each other using the corresponding arranging rule to get the translation of the whole entity. So, at the most we will obtain NxN entity translation proposals. In order to know which candidate is the correct one, the tool makes a web search, but as the number of candidates is so high, we use the same candidate selection technique applied previously for element selection.

This time we will use elements probability in order to obtain a measured proposal list. The *x* candidates with the highest probability are searched and ranked in a final candidate list of translated entities.

In our experiments, we have used the Google API to consult the web. Searching entities in Google has the advantage of getting the most common forms for entities in any type of document. But if you prefer to get a higher precision (rather than a good recall), you can obtain a higher certainty rate by making a specialized search in the web. For those specialized searches we have used Wikipedia, a free encyclopedia written collaboratively by many of its readers in many languages.

4.2 Language Independent Tool

Since creating transformation rules for every language pairs is not always a viable task, we have designed a general transformation grammar, which fits well for most language pairs that use the same alphabetical system. All we need is a written corpus for each language and a bilingual dictionary.

```
Input: Basque NE composed of n element.
Output: An ordered list of Spanish translations
Initialize: $proposals_i$ (i=1..n)
foreach $element_i$ ; i=1..n do
   $proposals_i$ = Apply pseudo-transliteration grammar
                   to $element_i$
end for
entity_translations = entity construction ($proposals_{1..n}$)
arranged_entity_translations = Search in comparable
                                corpora
```

Figure 2: Language Independent Tool

We have constructed a NE translation tool based on comparable corpora using that general grammar. As you can see in Figure 2, the system finds Basque translation proposals for entity elements applying the pseudo-transliteration module. Once it gets at least one translation candidate for each element, it applies the whole entity construction module obtaining all the possible whole entity candidates. Finally, it searches each candidate in the corresponding comparable corpus and returns a ranked candidate list based on that search, in order to obtain the correct translation form.

4.2.1 Pseudo-transliteration module

The pseudo-transliteration module has two main sources: an edition distance (Kukich, 1992) grammar and a Spanish lexicon.

The edition distance grammar is composed of three main rules:

1. a character can be replaced in a word
2. a character can disappear from a word
3. a new character can be inserted in a word

There is no specific rule in the grammar for switching adjacent characters, because we can simulate that transformation just combining the deleting and inserting rules mentioned above.

Since each rule can be applied *n* times for each word, the set of all translated words that we obtain, applying rules independently and combining them, is too extent.

In order to reduce the output proposal-set, we have combined the grammar with a Spanish lexicon, and we have restricted the transformation rules to two applications. So words with more than two transformations have been avoided. Thus, when the system applies the resultant automaton of this combination, only the Spanish words that can be obtained with a maximum of two transformations would be proposed as pseudo-transliterations of a Basque entity element.

The Spanish lexicon has been constructed with all the words of *EFE 2002* (the Spanish corpus of the 2002 year) and the bilingual dictionary *Elhuyar 2000*. And as we have considered this corpus as a comparable corpus with regard to the *Euskaldunon Egunkaria 2002*, Basque corpus version, we assume that most of the Basque words would have their corresponding translation in the Spanish set.

However, there are some words that do not have their corresponding translation at *EFE 2002*, or their translation cannot be obtained applying only two transformations. In order to obtain their translations in a different way, we have used the Basque-Spanish *Elhuyar 2000* bilingual dictionary. To be precise, we have converted the bilingual dictionary into an automaton, and we combined it with the resultant automaton obtained from applying the transliteration grammar in the Spanish lexicon.

In this way the system is able to translate not only the transliterated words in *EFE 2002* corpus, but also, the words that cannot be translated using transformation knowledge and that need information from a bilingual dictionary, such as ’Erakunde’ vs. ’Organización’[8].

4.2.2 Entire Entity Construction

Since we want to build a language independent system that works just having two different language data-sets, we cannot use any linguistic feature for arranging entity elements and getting the correct whole translated entity.

We might use many approaches to arrange elements, but we have chosen the simplest one: combining each proposed element with the rest, considering that each proposal can appear in any position within the entity. Thus, the system will return a large list of candidates, but we have ensured that it will include the correct one, when the independent translation of all the elements has been correctly done.

Although in some cases prepositions and articles are needed to obtain the correct Spanish form, the translation candidates for the whole entity will not contain any element apart from the translated words of the original entity. So, in the following step the lack of these elements will be taken into account.

4.2.3 Comparable Corpus Search

Once the system has calculated all possible translation candidates for the whole entity , the following step is to select the most suitable proposal. For that purpose, we have used the web in the linguistic tool. But this time, we have made used of the data-set in the Spanish-news articles, in which entities were tagged. This set is smaller and permits faster searching; furthermore, since Basque and Spanish-sets are comparable, the correct translation form is expected to occur in this smaller corpus, so it is very probable that the system will propose us the right translation.

Therefore, every translation proposal will be searched in the Spanish data-set and will be positioned at the ranked list according to their frequency. Thus, the most repeated entities in the corpus would appear on the top of the list.

4.2.4 Combining web and comparable corpus rankings

Both *Euskaldunon Egunkaria 2002* and *EFE 2002* data-sets are 2002 year news-sets, and a lot of named entities are due to occur in both sets. But since they are articles taken from newspaper of different countries, there may be some non-shared named entities.

When the system finds these special entities in the Spanish comparable corpus, it is very probable that it will find none of the candidates, and so, the list will not be arranged.

To avoid that random list ranking, when all translation candidates have a very low frequency, we propose to use the web to do a better rank-

[8] Organization

ing. As we will present below, this optional second ranking step improves final results.

5 Experiments

As we have mentioned before, we have first extracted a set of 180 person, location and organization name-pairs from *Euskaldunon Egunkaria 2002* newspaper and then we have translated them manually.

We have used three evaluation measures to present the result of all the experiments:

- $Precision = \frac{correctly\_translated\_NEs}{Translated\_NEs}$
- $Recall = \frac{correctly\_translated\_NEs}{All\_NEs}$
- $F - score = \frac{2*Precision*Recall}{Precision+Recall}$

For the evaluation of the linguistic tool, we have used a parameter (x in the tables) which determines how many translation candidates will be used in each module at the most. This threshold is necessary since the output of both transliteration and arranging grammar is too big to work with in the next modules.

The fr-min parameter in the tables specifies how often a candidate must occur in a data-set to be considered a likely NE translation proposal.

fr. min — x	**Precision**	**Recall**	**F-score**
10 — 1	73.96%	69.44%	71.63%
100 — 1	75.75%	69.44%	72.25%
250 — 1	**78.71%**	**67.77%**	**72.83%**
500 — 1	79.86%	61.66%	69.59%
10 — 3	79.29%	74.44%	76.79%
100 — 3	80.6%	73.88%	77.09%
250 — 3	**83.87%**	**72.22%**	**77.61%**
500 — 3	83.45%	64.4%	72.7%
10 — 10	79.88%	75%	77.36%
100 — 10	81.21%	74.44%	77.68%
250 — 10	84.52%	72.78%	78.21%
500 — 10	84.17%	65%	73.35%

Table 1: Linguistic knowledge + Google

Table 1 presents the results obtained applying the linguistic tool, and searching its proposals in Google. If we observe these results taking into account the values of the x parameter, it seems that the bigger the x value is, the better results we get. But note that the best improvement is obtained when we use the maximum of 3 candidate instead of using just 1. We improved the system performance in 5%. While using 10 candidates, the performance increases in less than 1% compared to the results obtained when x value is 3.

Regarding to the fr-min parameter, it seems that the best value is around 250. Moreover, duplicating this value, performance decreases. So we can say that when fr-min value exceeds 250, the system performs worse.

For next comparatives, we will take the results given by the experiments using the values fr-min=250 and x=1 as reference.

When we search Wikipedia instead of Google (see Table 2), the system's recall decreases from 69.44% to 66.67%. This time the only searching restriction is that the candidate occurs at least once, and not n times. This is because the data-set offered by Wikipedia is significantly smaller than the one given by Google. Moreover, precision remains similar. So although it is a smaller data-set, Wikipedia seems to be similar to Google as far as the information significance of terms is concerned.

fr. min — x	**Precision**	**Recall**	**F-score**
1 — 1	81.63%	66.67%	73.4%
1 — 3	**83.67%**	**68.33%**	**75.23%**
1 — 10	84.35%	68.88%	75.83%

Table 2: Linguistic knowledge + Wikipedia

When we use the comparable corpus instead of the web, the linguistic tool performs a considerable enhancement in precision, a 13% improvement, but gets worse coverage. On the other hand, the language-independent tool achieves similar results with regard to the linguistic tool searching in the web. So the language-independent tool seems to be a good alternative for dealing with NE translation without no exhaustive linguistic work. Those results are detailed in Table 3.

System	**Precision**	**Recall**	**F-score**
Ling. Tool	**91.85%**	68.8%	78.67%
Lang. Indep.	83.3%	72.2%	77.35%

Table 3: Results using comparable corpus

Finally, we have tried searching the proposals from the linguistic tool first in the comparable corpus. When no successful candidate is found in it, the system tries searching the web, in both Google and Wikipedia (See Table 4). In both experiments, precision is significantly lower than the one obtained when the system proposes candidates found in the comparable corpus, without no further search. However, the coverage increases in almost 5% in the trials carried out both with Google and Wikipedia. Therefore, the system's F-score remains similar. Note that this time instead of performing better when Google is used, the searches done in Wikipedia give better results. Furthermore, the best results are obtained when combining comparable corpus and Wikipedia searches in the Linguistic tool.

Web search	**Precision**	**Recall**	**F-score**
Google, 250	81.36%	73.3%	77.12%
Wikipedia, 1	**84.21%**	**73.3%**	**78.38%**

Table 4: Ling. Tool + Comp. corpus + Web search

6 Conclusions and Further Works

We have presented an approach for the design and development of an entity translation system from Basque to Spanish and the different techniques and resources we have used for this work.

On the one hand, we have combined bilingual dictionaries with a phonologic/spelling grammar for the entity elements' translation; on the other hand, we have applied a language-independent grammar based on edition distance. Both combinations perform well, and although the linguistic tool obtains better results, the language-independent grammar may be very useful for other experiments carried out with language-pairs others than Basque and Spanish.

Because of the differences of the syntactical structures of Basque and Spanish, it is necessary to arrange the entity elements for the correct translation of whole NEs; in particular, for those entities with more than one element. For that purpose, we have used two different techniques: probabilistic rules and a simple combination method (all candidates combined with all).

Finally, we have applied different resources and techniques for the selection of the best candidates. On the one hand, we have tried searching the web (Google and Wikipedia); on the other hand, we have used a comparable Basque-Spanish corpus. We have verified, that although Google is a bigger data-set, the significance of the information for NE translation task is similar to the information given by Wikipedia.

All the experiments carried out with comparable corpus have performed very well, and the best results have been obtained when combining it with Wikipedia. So developing a NE translation system based on comparable information have proved to be a good way to build a robust system.

However, some modules can be improved. Firstly, the methods to rank and select candidates are very simple, so if we use more complex ones, the number of candidates for the following modules would decrease considerably, and so, the system's final selection would be easier and more precise.

Regarding to the use of the web, actually we have only used Google and Wikipedia. Searches in Wikipedia are more precise than the ones made in Google and so the information they offer can be considered complementary. Furthermore, we can obtain very valuable information for other entity processes. For instance, since Wikipedia is a topic-classified encyclopedia, when you do an entity search, you can get information about the kind of documents in which the entity can occur; in other words, which is the most usual topic for it to occur in. Besides, that classification category can be very useful for entity disambiguation too.

With all the improvements presented so far, we hope to get a stronger entity name translation system in the future.

References

Aduriz I., Alegria I., Arriola J.M., Ezeiza N., Urizar R. 1998. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. Proceedings of COLING-ACL'98.

Alegria I., Balza I., Ezeiza N., Fernandez I., Urizar R. 2003. *Named Entity Recognition and Classification for texts in Basque*. Proceedings of JOTRI II.

Al-Onaizan Y., Knight K. 2002. *Translating Named Entities Using Monolingual and Bilingual Resources*. Proceedings of ACL 2002.

Al-Onaizan Y., Knight K. 2002. *Machine Transliteration of Names in Arabic Text.* Proceedings of ACL 2002.

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

Chen H., Yang C., Lin Y. 2003. *Learning Formulation and Transformation Rules for Multilingual Named Entities*. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition.

Kukich K., 1992. *Techniques for automatically correcting word in text. ACM Computing Surveys* Vol. 24 No. 4 377-439

Moore R. C., 2003. *Learning Translations of Named-Entity Phrases from Parallel Corpora*. Proceedings of EACL 2003.