Translating Named Entities using Comparable Corpora

Iñaki Alegria, Nerea Ezeiza, Izaskun Fernandez

IXA NLP Group
University of the Basque Country
Donostia, Basque Country
i.alegria@ehu.es, n.ezeiza@ehu.es, izas.fernandez@gmail.com

Abstract

In this paper we present a system for translating named entities between different language pairs, using comparable corpora. We present the different experiments we have tried, where we have translated entities from Basque into Spanish, and from Spanish into English. The aim of this experiments is twofold: on the one hand, we want to validate the strategy we propose to translate Basque named entities into Spanish taking advantage of comparable corpora; on the other hand, we want to prove that this approach is applicable to different language pairs and that the performance is reasonable.

1. Introduction

Person, location and organization names, are the main types of named entities (NEs), and they are expressions commonly used in all kinds of written texts. Recently, these expressions have become indispensable units for many applications in the area of information extraction, as well as for many searching engines. We can find many tools dealing with the identification and classification of named entities (CoNLL[1]) for specific languages. But, there is less published research on NEs translation. Luckily the interest is increasing considerably in the last years as we will see in the following section.

Our main goal is to build a multilingual NE database, which can be very useful for translation systems, multilingual information extraction tools (i.e. Question Answering) or multilingual systems in general. Since getting the information for that multilingual NE database was a complex task, we decided to work in the field of NEs' translation; furthermore, we wanted too design a system for translating those expressions between different language pairs.

If we look at the works published about NE translation, we can distinguish 3 types of systems: systems based on parallel corpora, which are the most widely used; 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 corpora is not an easy task, and it becomes harder when one of the languages in the pair is a minority language, as it is the case of Basque. Nevertheless, we can use comparable corpora to solve the problem of lacking parallel corpora. Comparable corpora are those datasets which are written in different languages but are not translations of one another, thus, they cannot be aligned. But they are supposed to deal with similar subjects and to be written in similar styles. Compiling that kind of corpora is much easier than obtaining parallel ones, 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 for possible entity translations.

For this work, we obtained the comparable corpora with the NEs tagged from the Hermes project[2](news databases: cross-lingual information retrieval and semantic extraction). All the entities have been automatically identified and classified. Those datasets are newspaper articles borrowed from different newspapers of the same year but they are not translations of one another. Anyway, the articles from different newspapers deal

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

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

with similar topics and news: international news, sports, politics, economy, culture, local issues and opinion articles, but with different scopes.

The Basque corpus has 40,648 articles with 9,655,559 words and 142,464 NEs from *Euskaldunon Egunkaria*, a newspaper entirely written in Basque; the Spanish corpus has 16,914 articles with 5,192,567 words and 106,473 NEs from the news agency *EFE*[3]; and finally, the English dataset has also been borrowed from *EFE*, and has 16,942 articles 3,631,335 words and 49,768 NEs.

As we can see, there are much more articles in the Basque corpus than in the others. And, even the Spanish and English corpora have similar amount of articles, the Spanish set has twice the number of NEs in the English set. However, we assume that they share common NEs and they could be an interesting resource for the NE translation task.

For our experiments, we have used two comparable datasets, one for the Basque-Spanish language pair, and another for the Spanish-English pair.

Besides these two datasets, we have also used some other information sources in order to develop the language independent NEs translation system:

- A finite-state transducer based on edit distance (Kukich, 1992), simulating simple cognates and transliteration transformations (Al-Onaizan *et al.*, 2002b) in a language independent way;

- A bilingual dictionary for the corresponding language pair;

- An element rearrangement module for language pairs that follows different syntactic patterns.

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 the NE translation system using a limited amount of linguistic knowledge. In section 5, we present the results of the experiments, and finally, section 6 presents some conclusions and future work.

2. Related Works

Recently, considerable research effort has been focused on machine translation systems (MT) and their improvement. But most of the MT systems translate named entities without any specific treatment. That is the reason why most systems will translate the Spanish form *escuela de derecho de Harvard* into *school of the right of Harvard* instead of *Harvard Law School* which is the correct English form, as Reeder argues (Reeder, 2001). So besides being a good way to obtain multilingual NE information, NE translation can be considered a helpful task for MT improvement.

Concerning the resources, despite the difficulty to get bilingual parallel corpora for many languages, most NE translation systems work with parallel datasets. Furthermore, those bilingual corpora are aligned at paragraph or even at phrase level. For example, Moore's work (Moore, 2003) uses bilingual parallel English-French aligned corpora, and he obtains a French form for each English entity applying different statistical techniques.

Although comparable corpora has been less studied, there are some known systems designed to work with them as well; Such as the system that translates entity names from Arabic to English (Al-Onaizan *et al.*, 2002a), and the Chinese-English translation tool presented in ACL 2003 (Chen *et al.*, 2003).

The main goal of both systems is to obtain the equivalent English form, taking Chinese and Arabic respectively as source language. 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 are different. Frequency based methods are used in Chinese-English translations, while in the Arabic-English language pair, a more complex combination of techniques is applied.

Similar techniques are applied at (Sproat *et al.*, 2006) and (Tao *et al.*, 2006), in which transliterate English-Chinese named entities using comparable corpora. The former combines a supervised phonetic transliteration technique and a phonetic frequency correlation approach, while the latter combines those techniques, but applying the pho-

[3] EFE is a news agency with delegations in Madrid and Miami

netic approach in an unsupervised way, where the distance is determined by a combination of substitution, insertion and deletion of characters.

Finally, we also want to mention the work (Poliquen *et al.*, 2005) which is integrated at the news analysis system NewsExplorer[4]. This research tries to extract person names from multilingual news collections to match name variants referring to the same person, and to infer relationships between people based on the co-occurrence information in related news.

In this paper, we present the research carried out for translating entity names using comparable corpora. We consider this method language independent, even though a bilingual dictionary is required, because we don't use any language dependent linguistic rule for the translation process. We have applied our method to Basque-Spanish and Spanish-English language pairs. We have also compare our results to the ones obtained with a language dependent NE translation system (Alegria *et al.*, 2006).

3. Experimental settings

When we started working at the NE translation task, we designed a language dependent tool for translating NEs from Basque to Spanish using comparable corpora. That system used linguistic information for both transliteration and entity element rearrangement. We tested this system using a set of the most common entities, and we obtained interesting results, with about a 78.7% F-score.

Since our goal is to obtain not only bilingual, but also multilingual NE information, and bearing in mind that designing a system for each language pair in a language dependent way is very expensive, we decided to experiment designing a relatively language independent tool following a similar strategy, and using comparable corpora and bilingual dictionaries. Firstly, we tested this tool in the Basque-Spanish language pair, in order to validate the methodology, and we compared it to the language dependent tool. We saw that the performance was even better than we expected and, it obtained an F-score of 77.5%, which is quite close to the performance of the language dependent tool. For this reason, we wanted to see if the tool could be really applied to other language pairs, and hence be useful for extracting multilingual NE information without an exhaustive linguistic modelling of other languages. So we tried the same experiment in the Spanish-English language pair.

As we have mentioned before, we have used two main resources for our experiments: comparable corpora and bilingual dictionaries. We have already described the corpora in the introduction. Concerning the bilingual dictionaries, we have used a set of 74,331 Basque words with their corresponding Spanish translations for Basque-Spanish experiments, while for Spanish-English experiments this resource contains 73,784 entries.

For evaluation purposes, we have used similar corpora, but extracted from different years. For each language pair, Basque-Spanish and Spanish-English, we have extracted 200 most frequent NEs in the source language and we have translated them manually.

In order to carry out an evaluation based on correct NEs, since the NEs were automatically treated, we verified that all the entities were correctly identified, because if the original entity was not a correct expression, the translation system could not probably propose a correct translation.

4. System Description

As we have mentioned before, we have applied a similar strategy to that used in the language dependent system for the design of the language independent NE translation tool.

The system uses 4 main modules: a grammar for transliteration combined with a bilingual dictionary for those words that cannot be translated only applying transliteration but also need some translation; an element rearranging module for the construction of the whole entity from components, which will treat the possible different syntactic structures between both languages, as it happens in the Basque-Spanish pair; and finally a searching module to decide which candidate is the most suitable. This architecture is described in Figure 1. In the following subsections we will present each module in detail.

[4] http://press.jrc.it/NewsExplorer/entities/en/1.html

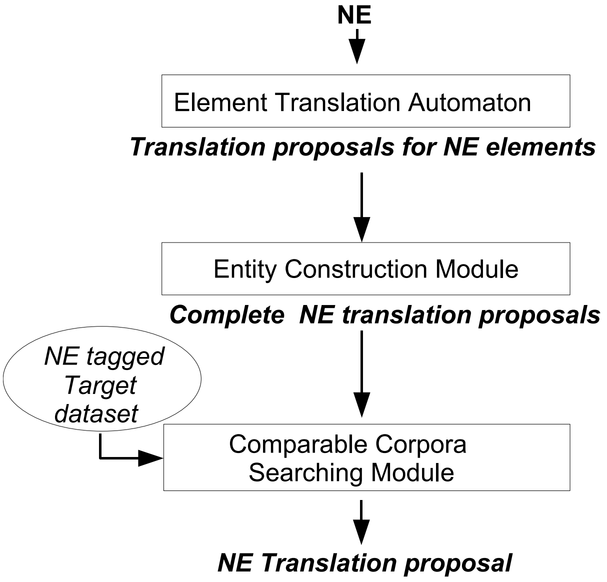
NE
Element Translation Automaton
Translation proposals for NE elements
Entity Construction Module
Complete NE translation proposals
NE tagged Target dataset
Comparable Corpora Searching Module
NE Translation proposal

Figure 1: System Architecture

4.1. Entity element translation module

The entity translation module has two main components: a transliteration finite-state automaton; and a bilingual lexicon.

We have used two main resources to automatically generate the transliteration rules: an edit distance (Kukich, 1992) based on a finite state grammar and a lexicon of the target language. Since this process is automatic it can be applied to any other language pair that uses similar alphabets.

The edit distance grammar uses the typical character based edit operations: insertion, deletion and replacement of a character in a word. Each operation is implemented as a rule in *XFST* (Beesley and Karttunen, 2001).

There is no specific rule in the grammar for switching adjacent characters, because that transformation can be simulated just combining the deletion and insertion operations mentioned above.

So this module will be able to obtain the translations of some of the NEs applying transliteration. For example, for the Basque-Spanish language pair, the system will transliterate *Kuba* into *Cuba*, replacing *K* with the *C* character; for the Spanish-English language pair, the system will transliterate *Constitución* into *Constitution*, replacing the second *c* with *t* and *ó* with *o*.

Since each rule can be applied *n* times for each word, the set of all translated words that we obtain after applying rules independently and combining them, is too extent. In order to reduce the output proposal-set, the system combines the grammar with a lexicon of the target language, and it restricts the transformation rules to at most two applications per word, avoiding the generation of words with more than two transformations (see Figure 2).

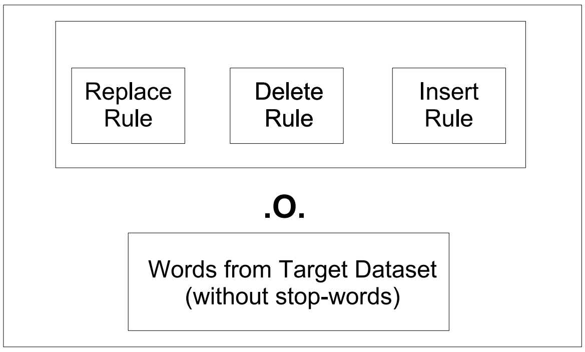

Figure 2: Transliteration automaton generation

We have generated three transliteration automata (TA) combining the mentioned resources:

- An automaton that copies the input word into the output (TA Max-transformations=0)
- An automaton generating words with at most one transformation (TA Max-transformations=1)
- An automaton generating words with at most two transformations (TA Max-transformations=2)

For the experiments, the target lexicons have been constructed using all the words from each target training set, excluding grammatical words such as prepositions, articles, etc., and using stop-lists[5].

However, there are some translations that cannot be obtained applying only transliteration rules. The system uses a source-target bilingual dictionary, converted into an automaton for those words. This automaton is combined with the three transliteration automata mentioned before. The application strategy is shown in Figure 3.

[5] http://www.lc.leidenuniv.nl/awcourse/oracle/text.920/a96518/astopsup.htm

The system firstly tries to obtain a translation proposal applying the zero-transformations TA to the input entity element. When the element is not found in the target lexicon, it applies the bilingual dictionary, and so forth.

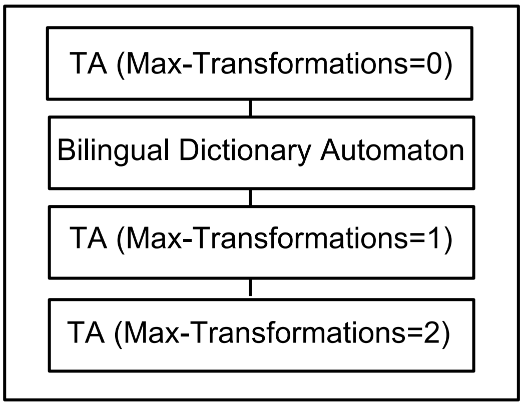
TA (Max-Transformations=0)
Bilingual Dictionary Automaton
TA (Max-Transformations=1)
TA (Max-Transformations=2)

Figure 3: Element Translation Strategy

So this module is able to translate not only the transliterated words in the comparable corpora, but also, the words that cannot be translated using transformation knowledge and that need information from a bilingual dictionary, such as 'Erakunde' vs. 'Organización'[6].

Since we have considered these datasets comparable, we assume that most of the source words would have their corresponding translation in the target dataset, in order to verify the correctness of the final translation automaton's output.

4.2. Entire Entity Construction

Since we want to build a language independent system that works just having two different language datasets, we don't want to use further linguistic information to combine syntactically the entity components. But we cannot ignore the possibility of having different syntactic patterns between languages, and this makes necessary to include some treatment for element rearrangement. This happens, for example in the Basque-Spanish language pair; Entity constituents may occur in different positions in both languages, so this module is applied before searching for translation candidates in the comparable corpora.

We might use many approaches to order 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 it will include the correct one, if the independent translation of all the elements has been done properly.

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

4.3. Comparable Corpus Search

Once the system has worked out all possible translation candidates for the whole entity, the following step consists on selecting the most suitable proposal. For that purpose, the system searches for them in the target language dataset, where entities are tagged.

Every translation proposal obtained from the previous step will be searched in the target dataset and each proposal will be positioned at a ranked list according to its frequency in the training corpus. Thus, the most repeated entities in the corpus will appear at the top of the list, being the most suitable translation proposals.

So briefly, the system takes a NE in source language as input, applies the translation module to each element, then it constructs the entire entity translation candidates, and finally it searches for them in a comparable corpora in order to obtain the most suitable ones, as described in Figure 4.

```
Input: NE composed of n element.
Output: An ordered list of NE  translations
Initialize: proposals$_i$ (i=1..n)
foreach element$_i$ ; i=1..n do
  proposals$_i$ = Apply element translation
                   automaton to element$_i$
end for
entity_translations = entity construction(proposals$_{1..n}$)
ranked_entity_translations = Searching
```

Figure 4: NE Translation Tool

[6] Organization

5. Experiments

As we have mentioned before, we have used a set of 200 most frequent NEs for each language pair, both Basque-Spanish and Spanish-English for evaluation.

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

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

When we compared the results for the Basque-Spanish pair of the language independent system with the ones obtained with the language dependent system (Alegria *et al.*, 2006), we saw that although the latter gets almost a 1.3% better performance, the performance of the language independent system could be considered a good approach with no need of exhaustive linguistic structure study.

However, we wanted to measure the performance of the Spanish-English language pair as well to verify if the results could be considered similar. The results of both experiments are shown in Table 1.

Lang. Pair	**Pr.**	**R.**	**Fs**
eu-es	82.02%	73%	77.5%
es-en	75.15%	62%	67.94%

Table 1: Language Independent System results

Observing these results, it seems that the system works considerably worse on the second language pair. In order to know the reason of that significant loss, we have reviewed all the supposed incorrect translations. We have observed that 26 of those translations were considered bad translations, because the frequency of the source NE form was higher than the one of the target form. This could be due to writing errors done by non-native speakers in the English EFE dataset. For example, when the system translates the Spanish form *Italia* into English, it creates a list of candidates where both *Italy* and *Italia* are generated. Then, as we have seen, it searches the candidate list at the comparable corpora and it ranks that list using frequency information on the corpus. Since in the English corpus *Italia* occurs more often than the correct form *Italy*, the former will be proposed as the most suitable translation, although the latter is the correct one. So when we evaluate this translation we see that an incorrect translation is proposed. Nevertheless, the error happens due to errors at the target corpus and not because of the bad performance of the language independent translation tool.

So, we can conclude that the system is very sensible to the target dataset correctness. And so, we guess that, if those 26 NE forms have their corresponding correct English form, the system would translate them correctly, and the results would be 5% better than the results for the Basque-Spanish pair.

6. Conclusions and Further Work

We have presented an approach for the design and development of a language independent NE translation system in order to obtain NE multilingual information, using comparable corpora, which seems to work well for different language pairs that have similar alphabets and writing habits.

To construct a new NE translation system, it is necessary to collect NE tagged comparable corpora for source and target languages, and also a bilingual source-target dictionary. The next step would be to extract the list of words (excluding stop-words) in the target dataset to generate the word translation automata using the general transliteration grammar already developed (as shown in Figure 2). Then the bilingual dictionary must be combined with the TAs obtained in the previous step (as shown in Figure 3). And finally, NEs in the target corpus must be extracted and stored along with their frequency, in order to select the most suitable translation among all the candidates.

Another way to select the most suitable NE translation is to use the web instead of the target dataset, as in (Moore, 2003). Nevertheless if we used the web, the system would be considerably slower due to the size of the resource, and consequently the answer time would be higher.

Another important issue is how to represent and link all this multilingual information to answer to a single language question in different language.

And finally, we want to improve the NE systems, including the translation system presented in this paper, and using the multilingual information we are collecting from all the comparable corpora.

7. Acknowledgement

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01), and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185), and partially supported by the Spanish Ministry of Education and Science (FIT-340000-2007-157 carried out on Tekniker). We would like to thank Itsaso Esparza for her help during the writing process of the paper.

8. References

Alegria I., Ezeiza N., Fernandez I. 2006. *Named Entities Translation Based on Comparable Corpora.* Proceedings of Multi-Word-Expressions in a Multilingual Context Workshop in EACL 2006.

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:Xerox Tools and Techniques.* 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.

Poliquen B., Steinberger R., Ignat C., Temnikova I., Widiger A., Zaghouani W., Žižka J. 2005. *Multilingual person name recognition and transliteration.* CORELA - COgnition, REpresentation, LAnguage, Poitiers, France, CERLICO. ISSN 1638-5748, 2005, vol. 3/3, no. 2, pp. 115-123.

Reeder F., 2001. *The Naming of Things and the Confusion of Tongues.* MT Evaluation: Who Did What To Whom Workshop on MT Summit VIII.

Sproat R., Tao T., Zhai C. 2006. *Named Entity Translation with Comparable Corpora.* Proceedings of the 21st International Conference on Computational Linguistic and 44th Annual Meeting of the ACL 2006.

Tao T., Yoon S., Fister A., Sproat R., Zhai C. 2006. *Unsupervised Named Entity Translation Using Temporal and Phonetic Correlation.* Proceedings of the 2006 EMNLP.