A Proposal For An Arabic Named Entity Tagger Leveraging a Parallel Corpus*

Doaa Samy
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
doaa@maria.lllf.uam.es

Antonio Moreno
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
sandoval@maria.lllf.uam.es

José Mª Guirao
Dpto. de Lenguajes y Sistemas
Universidad de Granada
jmguirao@ugr.es

Abstract

The term Named Entity (NE), first introduced in 1995 by the Message Understanding Conference (MUC-6), is widely used in the field of Natural Language Processing and Information Retrieval. Since 1995, a lot of studies have addressed NE recognition, tagging and classification. These studies reflected its efficient role in IE systems (Sekine, 2004; Grishman and Sundheim, 1996; Hasegawa et al., 2004) as well as its effectiveness when used as anchor points in alignment techniques (Melamed, 2001; Samy et al., 2004). In this paper, we cover three main aspects concerning Arabic NE recognition and tagging. First, we present an overview of the linguistic nature and the studies concerning NE in Arabic texts. Second, we highlight the methodology of developing tools leveraging parallel corpora and previously developed tools for other languages. Third, we present our proposal for an Arabic NE tagger; its different modules, its coverage scope and the methodology used for its implementation. However, it could also be considered a method for aligning NE in parallel corpora. Finally, we evaluate the results against a gold standard. At the end, we discuss the final conclusions and future work.

1 Introduction

In this section, we will introduce an overview of the research held in the field of NE in general and a historical review of studies addressing the transliteration of Arabic Names.

1.1 Named Entities

NE recognition has proved to be an outstanding factor in the improvement of IR, CLIR and QA systems. In this paper, we try to highlight its importance in parallel text processing and alignment of parallel corpora.

The early NE classifications considered two main classes: names and numeric expressions. Both classes covered a range of 7 to 10 categories. Names might include categories such as: person names, organizations, location names, while numeric expressions cover the scope of: time, date, money and percent expressions (Sekine, 2004). These categories have been extended aiming at a wider coverage. An example of such expansion is the "200 category extended named entity hierarchy" proposed by Sekine (2004).

Although the idea of such an extensive categorization seems so appealing, it is quite beyond the

* This research has been supported by the grant TIN2004-07588-C03-02 (Spanish Ministry of Education and Science).

scope of our Arabic NE tagger for the time being, as it is a very laborious task in terms of time and annotation effort. Besides, we believe that in cases where languages lack resources for NE, which is the case in Arabic, it is more effective to start with basic categories. Once these resources are available, research should proceed on with its respective expansion.

1.2 Named Entities and Arabic Transliteration

Proper names constitute an important building block in the basic NE classifications. However, Semitic languages, in general, and Arabic scripted languages, in particular, present a challenge to the automated approaches for Proper Names and/or NE recognition. This fact could be explained if we take into consideration that a wide range of automated detection of Names (in Roman scripted languages) is based on formal orthographic criteria. These systems make use of the initial capitalisation of names of persons, locations, job titles and organizations. Also, upper case letters are used to indicate acronyms. Arabic scripted languages, on the other hand, do not provide such orthographic distinction, as they do not distinguish between upper case and lower case. That is why systems dealing with Semitic or Arabic Proper Names have to adopt different techniques to overcome such challenges.

To our knowledge, early studies tackling this issue in a computational context date to the early nineties (Roochnik, 1993; Arbabi et al., 1994). Such studies focused mainly on developing techniques and algorithms for transliteration. In this aspect, we consider it interesting to point out the following observations.

Reviewing the previous literature helped us establish the following key stages in the development of research concerning Arabic names:

Early beginnings (1993-1995): Interest in NE and Arabic Name transliteration almost coincided chronologically, although transliteration was prior to the concept of NE (first introduced in 1995).

The nineties: Despite the strong connections between both research fields, these fields remained unrelated, and each followed its own course independently. This situation prevailed because the target of transliteration focused mainly on machine translation systems (Stalls and Knight, 1998) or security issues, for example, border controls or passport checking as mentioned by Arbabi (1994); hence Information Retrieval as an important application field was not targeted at that time.

2000 to present: research in both fields (NE and Arabic Name Transliteration) began to converge in some way, although they have been limited to Arabic names transliteration and they did not include other categories of Arabic Named Entities. Besides, these studies had as a main target: IR and CLIR systems (AbdulJaleel et al., 2003; Darweesh et al., 2001; Al-Onaizan and Knight, 2002; Larkey at al., 2003; Gey and Oard, 2001, Cowie and Abdelali). The only occasion, where transliteration was mentioned within the general framework of NE, was in the study of Al-Onaizan (2002) on "*Translating Named Entities using monolingual and bilingual resources"*, also designed and implemented from the perspective of IR/CLIR applications

After this review of previous work, it is clear that all approaches consider transliteration of Proper Names an indispensable step towards Arabic NE recognition. However, we would like to insist on the fact that transliteration covers only a subset of NE and that there is still a need for a comprehensive study that covers the rest of NE categories in Arabic scripted languages, in particular, without limiting the approaches to transliteration.

In this paper, we are trying to fill this gap by introducing a proposal for an Arabic NE recognition leveraging a Parallel Corpus (Spanish-Arabic) covering a wider scope of categories such as organization names, job titles and acronyms. Our approach is different in its resources and its main target application. Our main resource is an aligned parallel corpus and our final target is to identify the Arabic NE. In this way, the tagged NE would serve as anchor point for the alignment process.

2 Methodology

Developing a tagger is a task requiring the availability of either monolingual or bilingual resources. Almost all previous work in the field developed its techniques using data from bilingual dictionaries, lexicons or just simple lists of Proper and location names. The recent experiments, which try to adopt a totally statistical approach, depend mainly on lists of Proper Names and their corre-

sponding transliterations (Abduljaleel, 2003). Even the hybrid approaches combining linguistic and statistical methods validate their transliterations candidates against lists of proper names or against web counts (Al-Onaizan, 2002).

Our methodology, on the other hand, relies on two main types of resources; parallel corpora and previously developed tools for other languages.

2.1 Parallel Corpora

New approaches to develop NLP tools focus on the feasibility of using parallel corpora as resources. Such approach proved to be effective in terms of time and effort. Besides it provides the advantage of dealing with the different linguistic phenomena *in situ*, i.e. it offers an empirical data set for developing and testing the tools. Recent research on Word Sense Disambiguation makes use of parallel corpora (Diab and Resnik, 2002). Building Wordnets is another field which made use of parallel corpora (Diab, 2004).

For our tagger, we used an Arabic-Spanish parallel corpus aligned on the sentence level and tagged on the level of POS. The size of the subcorpus used for the experiment is not large (1200 sentence pairs), but due to its nature and its source, it contains a considerable number of NE. The corpus consists of UN documents published on the web. Since it was quite difficult to obtain parallel and reliable texts in this language pair (Spanish-Arabic), we opted for the UN documents as both Spanish and Arabic languages are official UN languages. The advantages of using this corpus can be summarized in the following points:

- *Reliability:* Considering the source, we could guarantee a *translation and transliteration* quality for the Named Entities.
- *Representativeness*: The corpus is a representation of Modern Standard Arabic on one hand, and of Standard Spanish on the other.

2.2 Previously developed tools for other languages

The second resource consists of previously developed tools for other languages. This resource used together with parallel corpora proved to give good results in many NLP applications.

Since we are using a Spanish-Arabic parallel corpus, the tools, which were mainly developed for processing the Spanish corpus, were used as a starting point for developing our Arabic tools. We, basically, relied on the output of the Spanish NE tagger. It is a rule-based tagger enriched with a monolingual Spanish lexicon. This tagger searches for patterns of Spanish NE and the patterns matched are tagged in xml with the tag:

```
<ne type = "" id = "">….</ne>
```

The Spanish NE tagger covered only two main NE categories: "np" (Nombre Propio /*Proper Noun*) and "date". However, for the purpose of our experiment, the first type was extended to include:

- Person names
- Location names (Geographical locations and toponyms)
- Organizations (Political of Administrative Entities)
- Position (job titles)
- Acronyms

Following the new classification criteria, we had to modify the values of the *type* attribute in the Spanish Corpus.

3 Implementation

3.1 Scope and Structure

The above categorization is a semantic categorization. However, the implementation modules do not correspond strictly to this semantic classification. Instead, the implementation was based on pattern matching, lexical, orthographic and phonetic criteria. There are three basic modules:

- A module for date expressions
- A module for names based on simple transliteration. This covers the categories of person names, location names and some acronyms when phonetically transliterated.
- A module based on a bilingual lexicon. This module covers the categories of organizations and positions (job titles)

The "date" Module: Arabic date tagging depends mainly on regular patterns and a small lookup lexicon of months and days. The bilingual

lexicon of months includes months in Spanish and their equivalent in Arabic according to the Gregorian calendar (January, February, …etc) and the Lebanese calendar, since both are of common use in Arabic UN documents.

Transliteration Module: By transliteration, we mean the process of formulating a representation of words in one language using the alphabet of another language (Arbabi, 1994). In other words, it consists of the representation of a word in the closest corresponding letters or characters of a different alphabet or language, so that the pronunciation is as close as possible to the original word (Abdul-Jaleel, 2003).

Our implementation is a simple, straightforward one, but it proved to be efficient as it succeeded in meeting our main goal of detecting the Arabic names in the corpus. The main advantage over other more sophisticated approaches is that the parallel corpus plays a double role as a resource and a target at the same time. In addition to this, the fact that the parallel corpus is aligned reduces significantly the context and scope of search for valid transliterations.

To avoid encoding schemes problems or unrecognized characters, we decided to implement the transliteration module by means of numerical codification using the *Unicode* value for each Arabic character. Another solution was to use the Buckwalter's transliteration scheme considered almost a classic standard in Arabic NLP. However we decided to use Unicode as it supposes more portability to other languages if different phonetic/orthographic criteria are applicable.

On the other hand, and in the transliteration mappings from Roman characters, each character was given all its corresponding possibilities in the Arabic alphabet and consequently it is given the numeric *Unicode* value referring to each of these characters.

Arabic Character	Roman Character	Code
ب	[Pp]|[Bb]	0628
ر	[Rr]|[Rr]r	0631
ج	[Gg]|	062C
غ	[Gg]|	063A

Table 1. Example of Arabic characters and their codes

Expansion and Omission: In the transliteration module, we tried to deal with two phenomena: expansion and omission. Expansion consists in the possibility that one Roman character might be transliterated into two or more Arabic characters. For example, the "*t*" might have two possible transliterations in Arabic, either "ت" (062A) or "ط" (0637). The mapping, in this case, would be as follows: when a letter *t* is found, it could be transliterated either by character code 062A or 0637.

Omissions are common in short vowels' transliterations. Arabic scripted languages do not transliterate the short vowels. Instead, it uses the diacritics. But, in Modern Standard Arabic texts, words rarely appear with diacritics. This creates ambiguity for computational systems on all levels, starting from the tokenization till the semantic levels. In this aspect, transliteration is not an exception. However, the most practical way to deal with such phenomena is to handle the omissions. To do that, we used the regular expression operator "?" to indicate that the preceding character code might occur zero or one time(s).

Tokenization: To our knowledge, this feature has never been addressed in previous literature concerning the transliteration because almost all approaches were aiming at finding the best transliterations for a given name independently of its context. That is why tackling the tokenization problem was not considered. In our case, since we deal with a corpus, NE appear in their real context and one important issue, in this respect, is that NE as other nouns in Arabic may appear preceded by clitics. These clitics might be a conjunction "و", a preposition "ب", "ل"or both "وب", "ول". To handle such feature, we had to expand the possibilities of matching by indicating that the string might be preceded by one or more pre-clitics.

Look-up module: In case of organizations and job titles, the Named Entity is either a one-word NE, such as `Embajador` *(Ambassador)*, `Presidente` *(President)*, or a compound NE; two or more tokens, such as `Naciones Unidas` (*United Nations*). Both types are looked up in the general lexicon used for POS tagging, since these words are originally common words, but they have passed from common words to NE through a semantic process to refer to a certain entity. This semantic phenomenon is reflected orthographically in the use of upper case. The look-up is easy and feasible, as it does not need especial effort for creating lists of NE referring to organizations or job titles.

3.2 Algorithm

This section explains how the tagging process takes place given that the Spanish NE have been previously annotated according to the above-mentioned classification. Our implementation relies on this basic assumption: "Given a pair of sentences where each is the translation of the other; and given that in one sentence one or more NE were detected, then the corresponding aligned sentence should contain the same NE either translated or transliterated".

This assumption is a simplistic one, as it doesn't take into consideration common phenomena in translation such as omission or addition. Despite this fact, NEs usually tend to be conserved in translations as they represent significant pieces of information. Such a semantic weight is reflected in the way translators deal with them. While a translator might have more flexibility in translating common nouns or expressions, when dealing with NE, the translator rather tries to keep the translation as close as possible to the source. Starting from this assumption, we follow this algorithm.

Input: The input consists of the file containing the aligned parallel corpus with Spanish NE tagged. The corpus is processed so that each pair of aligned sentences (*x, y*) is handled one at a time. We begin by processing the Spanish sentence in the following way:

- Previously tagged Spanish NEs are extracted from the Spanish sentence.
- Extracted NEs are classified in sub lists depending on their type.
- First, NEs of type date are passed to the date module.
- Given the list of tagged dates in Spanish in a sentence *x*. The system looks up the bilingual lexicon of months and numbers to find their equivalent in Arabic. Once found, the system searches the corresponding aligned Arabic sentence *y* for the pattern generated. If the generated pattern is found, it is tagged by the same tag as its Spanish equivalent and it is given the same ID number. If not, it exists this module.
- Second, NEs of type Person names, location names, toponyms and some acronyms[1] are passed to the transliteration module.
- For each Spanish NE and according to the mapping scheme, the system provides a combination of all possibilities of transliteration. The output consists of the Spanish NE together with a string with all transliteration possibilities. Different possible transliterations for each character are separated by "|". In case of vowels the specific numeric code is followed by "?" indicating that zero instances or one of the preceding character could occur. For example, given the proper name `Carl`, the transliteration module generates the following string

```
(0643|0633|062B|0642|062A0634)
(0629|0623|0639|0627|0647|0622
|0649|0621)? 0631 0644
```

- A list of all the Arabic words in the corresponding Arabic sentence is extracted. Each word is converted to a string of numeric codes, according to the codification scheme. In the example mentioned above, the Arabic word "كارل" receives the following codification:

```
0643 0627 0631 0644
```

Comparing the Arabic string "0643 0627 0631 0644" against the above transliteration returns true. Thus, "كارل" is the corresponding NE equivalent to "`Carl`".

- Finally, the valid candidate is automatically tagged by the same tag and is given the same ID number of its Spanish equivalent.
- Spanish Nes of type organization or job title are passed to the lookup module. The output of this stage is the looked-up Spanish NE, together with its Arabic translation obtained from the bilingual lexicon.
- Arabic translations are searched in the corresponding aligned sentence. If found, the

1 Acronyms are dealt with in the Arabic text by different ways. One possibility is to be transliterated phonetically. Another possibility is to use the name in its full form.

Arabic NE is tagged with the same tag and the same ID number of its corresponding Spanish NE.

Tagging Acronyms: Acronyms are handled in one of two ways. An acronym first is passed to the transliteration module. If found, then the Arabic translator has opted for a transliteration of the Acronym. Otherwise, the Acronym is returned to its full form, since usually the first occurrence of an acronym in a text is accompanied by its name in full form. We keep track of this name and if the transliteration module fails to find a candidate, it passes to the look up module where it searches for the equivalent translation. When found, it is tagged with the same tag and given the same ID number as its corresponding Spanish NE.

Unknown Named Entities: NE, which failed to be recognized through the previous stages, are names whose Arabic equivalents are totally different such as "Grecia" (*Greece*) "اليونان" or "Egipto" (Egypt) "مصر". This is explained in terms of the History of Language, which is far beyond our scope. The only way to tag such unknown words is either by human intervention, or by consulting a bilingual list of names if available.

Final Output: The final output consists of the same aligned corpus with the Arabic NE tagged indicating their type and given the same ID numbers of their corresponding Spanish ID.

4 Evaluation

The results of the NE tagger were evaluated against a gold standard set. From the 1200 pairs of sentences, 300 sentences from the Spanish corpus were selected randomly with their equivalent Arabic sentences. For each pair, the output of the NE tagger was compared to the manually annotated gold standard set.

The evaluation took place on the different tagging levels testing in that way the different tagging modules. The best results were achieved in the "date" module and the "look-up" module.

In the acronyms, sometimes due to the inconsistency in translating the acronyms to the Arabic, beside the extended length of the name, the tagger was not able to correctly identify all the Arabic corresponding NE. The acronyms were correctly identified only in 76% of the cases.

The transliteration module showed high coverage and accuracy in recognizing the transliterated NE. It correctly identified and tagged almost all transliterated NE (Recall 97.5%), even when the NE in Spanish and Arabic was not a precise transliteration; such as "Somalia" and its Arabic equivalent "الصومال". This is due to expanding the possibilities on one hand, and handling the vowels' omission and the tokenization, on the other hand. The only drawback of expansion is that the system in some cases wrongly identified words as NE (Precision 84%). To improve the precision, we applied a filter to the Arabic words, which omitted the Stop Words from the possible transliterated candidates. This increased the precision result significantly reaching (90%). Table 2 shows NE distribution in the evaluation and Table 3 shows the evaluation results.

	Arabic	Spanish
N. of sentences	307	300
Total N. of NE	721	743
Average NE/sent	2.41	2.54
Proper Names	39	40
Toponyms	164	167
Acronyms	11	27
Jobs	123	128
Organizations	275	277
Dates	109	104

Table 2. NE Distribution in the evaluation corpus

Recall	Precision	Improved Precision
97.5%	84%	90%

Table 3. Evaluation results

5 Conclusion and Future Work

NE recognition leveraging a parallel corpus and reusing previously developed tools for other languages proved to be an efficient methodology, as it supposes a feasible and cost effective solution to develop resources specially for languages with scarce resources.

Results obtained show that our basic assumption was practical and applicable. Although the transliteration module could be considered a shallow one, as it does not apply sophisticated statisti-

cal methods, but it was efficient for the task and it managed to meet the suggested goals.

Although the transliteration was implemented considering the Spanish-Arabic, we tried in the majority of cases to follow more general criteria, applicable on English-Arabic transliteration or French-Arabic transliteration. This is because the NEs tagged in the Spanish Corpus are not exclusively Spanish names. They are names proceeding from different languages; English, French, German, …etc.

For future work, we would consider applying statistical models for transliteration. Also a character bigram would be of great significance.

On the other hand, a phonological transcription tool for Spanish might be applied to the Spanish NE. The information concerning the syllables and their divisions might help us in improving the transliteration module.

Finally, the more trained the tagger, the more NE it would recognize, since in each training pass, the lexicon is enriched with the new NE. Such a resource would be very useful in working not only with parallel, but also with comparable corpora. Besides, such a list of NE extracted from real text would be a valuable resource for IR and/or CLIR applications.

Other applications might include Example Based Machine Translation, Translation Memories or Computer Assisted Language Learning since a parallel aligned corpus with both POS and NEs tagged, is considered a valuable resource especially for uncommon language pairs as Spanish and Arabic.

References

AbdulJaleel, N. and Larkey, L. 2003. English to Arabic Transliteration for Information Retrieval: A Statistical Approach, *CIIR Technical Report IR- 261*.

Al- Onaizan, Y. and Knight, K. 2002. Machine translation of names in Arabic text. *Proceedings of the ACL conference workshop on computational approaches to Semitic Languages*.

Al- Onaizan, Y. and Knight, K. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. . *Proceedings of 40th ACL Conference,* Philadelphia, pp. 400-408.

Arbabi, M. Fischthal, S. M. Cheng, V. C. and Bart, E. 1994. Algorithms for Arabic name transliteration. *IBM Journal of Research and Developme*nt, 38(2): 183– 193.

Cowie, J and Abdelali, A. Interactive Cross Language Information Retrieval Using Transliterated Names Resolution. *Memoranda in Computer and Cognitive ScienceMCCS-04-331*.

Darwish, K., Doermann, D. Jones, R., Oard, D. and Rautiainen, M. 2001. TREC-10 experiments at Maryland: CLIR and video. In *TREC 2001*. Gaithersburg: NIST.

Diab, M. 2004. The feasibility of bootstrapping an Arabic Wordnet leveraging parallel corpora and English WordNet. *Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR 2004),* Cairo, Egypt, pp.71-77.

Diab, M. and Resnik, P. 2002. Word sense tagging using parallel corpora. *Proceedings of 40th ACL Conference*, Pennsylvania, USA.

Gey, F. C. and Oard, D. W. 2001. The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French, or Arabic queries. In *TREC 2001*. Gaithersburg: NIST.

Grishman, R. and Sundheim, B. "Message Understanding Conference - 6: A Brief History", *COLING-96*.

Hasegawa, T. Sekine, S. Grishman, R. 2004. Discovering Relations among Named Entities from Large Corpora. *Proceedings of ACL 04*; Barcelona, Spain.

Larkey, L., AbdulJaleel, N. and Connell, M. 2003. What's in a Name?: Proper Names in Arabic. Cross Language Information Retrieval. *CIIR Technical Report, IR- 278*.

Melamed, I. D. 2001. *Empirical Methods for Exploiting Parallel Text*, Cambridge/London: MIT Press.

Roochnik, P. 1993. *Computer-Based Solutions to Certain Linguistic Problems Arising from the Romanization of Arabic Names,* Ph.D. Dissertation, Georgetown University, Washington, DC.

Samy, D., Moreno Sandoval, A. and Guirao, J.M.2004. An Alignment Experiment of a Spanish-Arabic Parallel Corpus. *Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR 2004),* Cairo, Egypt, pp.85-89.

Sekine, S. 2004. Named Entity: History and Future.

http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf

Stalls, B. and Knight, K. 1998. Translating Names and Technical Terms in Arabic Text. *COLING/ACL*

*Workshop on Computational Approaches to Semitic Language*s. Montreal, Quebéc.