An Alignment Experiment of a Spanish-Arabic Parallel Corpus

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

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

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

Abstract

In the last years, parallel corpora have been garnering attention as an important resource for the development of linguistic tools. However, the Arabic presence in the actual panorama of the Parallel Corpora and the Parallel Text Processing is so limited. In this paper, we try to fill up this space presenting a preliminary approach for an alignment experiment of a Spanish-Arabic Parallel Corpus at the sentence level. The experiment adopts a hybrid methodology, since it applies statistical models, together with lexical information (Named Entities as anchor points). The corpus, object of the alignment process, is a parallel test corpus Spanish-Arabic, which have passed through different pre-alignment monolingual stages including: simple tokenization, segmentation and *Named Entities* tagging. In these stages, the already available tools for Spanish (developed by the *Laboratorio de Lingüística Informática*, Autónoma University Madrid) were used as a starting point for a) developing the necessary tools for the Arabic tokenization and segmentation b) tagging the *Named Entities* in the Arabic text. Finally, the results are evaluated according to a golden standard, which consists of the same test corpus manually aligned. Carrying out such an experiment revealed certain facts about studying "uncommon" linguistic pairs, in this case, the Spanish-Arabic, besides the lack of tools and resources for processing the Arabic language in a parallel context.

1. Introduction: Arabic Language and The Parallel Corpora

In this introductory section, we would like to highlight the state of art in the field of Arabic corpora. Meanwhile, the linguistically annotated corpora, either monolingual or parallel corpora, are an indispensable resource for developing Natural Language Processing tools. In the case of the Arabic language, the actual state of art reveals an increasing interest for building Arabic corpora, although these efforts are still not sufficient to meet the Arabic NLP community's needs.

On the other hand and concerning the parallel bilingual or multilingual resources, the actual panorama reveals a serious lack of multilingual or bilingual corpora where Arabic is one of the languages in concern. This conclusion was reached after conducting a general survey of the state of art concerning the multilingual resources. It can be detailed in the following aspects:

- There are no available parallel corpora in which Arabic forms one of the linguistic pairs.
- Most of the computational and corpus-linguistic studies concerned with Arabic have studied this language either independently (Dichy, 2001), (Khoja, 2001), (Attiya, 2000), (Rezaei, 2001) or in comparison mainly with English (Darweesh, 2003), (Diab 2004a), (Goweder & De Roeck, 2001), (Elkatib & Black, 2001) and in fewer cases with French (Gudière, 2002) or both English and French (Lelubre, 2001).
- Studies focusing on Arabic in a parallel context created its own bitexts, in other words, they created an *artificial* parallel corpus, since they translated parts of English corpora into Arabic using available MT systems (Diab, 2004a).

These conclusions will lead us to the second section of our paper

2. The Spanish-Arabic Linguistic Pair

The above survey points out some aspects of the difficulty of our task "Construction and Alignment of a Spanish-Arabic Parallel Corpus". Thus it is important to answer the following questions:

1-Why a Spanish-Arabic corpus?

2- In what way is it difficult to build up a Spanish-Arabic Parallel Corpus?

The answer to the first question summarizes one of the main benefits and reasons behind the everyday increasing interest in exploiting parallel corpora for NLP applications. Two fundamental facts can perfectly describe this interest:

1) Linguistic resources and tools are not equally available for the different languages (especially in the case of European and non-European)

2) Previously developed and tested tools for some languages can be reused as a starting point, though with certain adaptations, to develop necessary tools for languages that are scarce of linguistic resources. In some cases, not only can it be used as model, but also it can be directly bootstrapped for other languages leveraging aligned parallel corpora. This methodology has been adopted recently by different researchers in the Arabic NLP field, especially by M.Diab who made use of the available tools for English language and applied them to the Arabic text at different levels of linguistic processing such as tokenization, POS tagging, Base Phrase chunks (Diab et al., 2004) and finally Word Sense Disambiguation (Diab, 2004). The experiments proved the efficiency and the adequacy of the approach.

The situation, in our case, is very similar. The *Laboratorio de Lingüística Informática* in the Autónoma University of Madrid, have a number of developed and tested tools for Spanish language which can be adapted to the Arabic text

leveraging a parallel corpus. This fact gives the complete answer to our question "Why a Spanish-Arabic parallel corpus?"
Answering the second question concerning the difficulties for this linguistic pair can be summarized in the following aspects:
- Despite the available wide literature about comparative linguistic studies between Arabic and Spanish realized by different Arabists and Hispanists, these studies have always been addressed in terms of traditional and theoretical linguistics. To date, there are no linguistic studies addressing this linguistic pair from a computational perspective.
- Resources and tools for Spanish language are available either in a monolingual context or in a multilingual context. In multilingual contexts, Spanish is studied in comparison with English language on the first place, and/or other European languages in the second place. The recent studies of Spanish Natural Language Processing at the University of Maryland (Cabezas et al., 2001) is just one of several examples in the field. On the other hand, Spanish language has always been present in several European projects, among which we mention EUROTRA, EURODICAUTOM, CRATER, MULTEXT (McEnery, 1997).
- All the above stated facts point out to the conclusion that we are facing a novel linguistic pair in the field of NLP, hence scarce of linguistic resources or any previous literature. Thus we have to start our task from the beginning by building up bilingual resources for this linguistic pair. In the next section, we discuss the steps to build up the corpus.

3. The Parallel Corpus

This section will address the main tasks involved in the construction and the assessment of the corpus.

3.1. Building the Corpus

When using the term "parallel" we adopt the definition of McEnery and Somers, "a set of L1 texts and an equivalent set of L2 translations of L1"(McEnery, 1997). In other words, "a text, which is available in two (or more) languages"(Somers, 2001).
From the beginning we decided to opt for locating documents in Spanish and Arabic through the World Wide Web avoiding in this way scanning documents, and thus saving time and effort, on one hand and avoiding noise resulting from OCR errors, on the other hand. Our main criterion is to build up a reliable Parallel Corpus in terms of quantity and quality. Locating texts, which meet these criteria, was some how troublesome at the beginning, but these criteria were met in the official texts of the United Nations, since Spanish and Arabic are, among other languages, official languages of the UN.

3.2. Corpus Characteristics

The parallel Spanish-Arabic corpus consists mainly of annual reports of different UN institutions such as the Security Council, the Economic and Social Council, …etc. All documents are in both languages Spanish and Arabic[1], with a total size of about 2 million words. The corpus reveals the following features:

a) Concerning the Arabic part, it is a real representation of modern Standard Arabic as used in formal official documents.
b) High quality translation is guaranteed.
c) High frequency of *Named Entities*: proper names, dates, countries, …etc. due to text typology.

The following feature is considered the major disadvantage:
- Compared with other text typologies, the UN documents are "clean well-formatted" texts, since they don't contain so much noise on the formal aspect. Besides the translations reveal a high degree of accuracy and consistency. These features may be criticized since they might be considered an "idealized" input for NLP applications, which hardly reflects the real problems of noisy texts. We cannot deny this point of view, but taking into consideration the difficulty of locating texts in this linguistic pair together with the novelty of the approach, this can be a good starting point, which can provide us with the basic necessary resources for further investigation on noisy texts.

3.3. Corpus Assessment

3.3.1. Monolingual tokenization and segmentation

Tokenization: For the tokenization, we used the modules we had already developed for the Arabic corpus following the model of the Spanish tools (Samy et.al., 2004).

Segmentation: The output of this stage is crucial for the alignment process that is why we dedicated especial attention to this task and we manually revised the segmentation of the sample to be aligned. In this stage we would like to highlight a main source of noise during the Arabic segmentation:
The use of numbers within the Arabic text caused much noise in the detection of the sentence boundaries. The Arabic text follows a right-to-left direction, while the numbers (either the Latin [0-9] or the Indian [٠١٢٣٤٥٦٧٨٩] systems) are written in a left-to-right direction. If the sentence boundary coincides with a number, the direction changes causing alterations in the position of the full stops, ending up as a source of noise during the segmentation process.
On the other hand, the results of the segmentation showed that both Spanish and Arabic documents are aligned on the paragraph leveled, since the number of paragraphs coincided with their corresponding translations. This fact helped in the posterior process of alignment.
The following table summarizes the results of the tokenization and segmentation conducted on a sample of the corpus.

Language	Spanish	Arabic
No. of Tokens	39,496	25,144
No. of Sentences	1,168	1,165
Average tokens/sentence	33.815	21.582

[1] Documents in English language were also downloaded for future multilingual researches.

Table 1: Tokenization and Segmentation Results (Samy et.al., 2004)

4. The Alignment Process

Once we have prepared the parallel corpus, we can now proceed with the alignment experiment. By alignment, we mean "Given parallel texts *U* and *V*, an alignment is a segmentation of *U* and *V* into *n* segments each, so that for each *i,* $1 \leq i \leq n$, u_i and v_i are mutual translations. An *aligned segment* a_i is an ordered pair (u_i, v_i). Thus, an alignment *A* can also be defined as a sequence of aligned segments: $A \equiv \langle a_1, \ldots, a_n \rangle$" (Melamed, 2001:9). Our alignment experiment adopts the methodology based on statistical and lexical factors.

Our program is based on a statistical model for aligning sentences depending on the correlation between sentence length in parallel texts *U* and *V*. The model we adopt is a combination between the approaches of Gale and Church and (1991) on one hand and the approaches of Brown (1991) and Chen (1993), on the other. The key idea is based on the observation that long sentences tend to be translated into longer sentences, while short sentences tend to be translated into shorter ones. This model proved its efficiency in cases of 1-1 alignments; however, it is very sensitive to any noise either in the form (noise at the segments boundaries level) or in the content (such as cases of omissions in translations). To avoid such problems, we make use of the tagged *Named Entities* as lexical anchor points.

4.1. Tagging the Named Entities

Tagging the Named Entities was carried out in two separate monolingual modules. We distinguish between two main classes of Named Entities:

- Dates
- Proper Names

The set of Proper Names include the following subsets:

- Names of Persons
- Names of Administrative Entities, Institutions and Authorities
- Acronyms
- Toponyms (Country Names, Regions, Cities, …etc)

The module for identifying and tagging dates is based on pattern recognition. On the other hand the module for Proper Names tagging is based mainly on the formal textual aspects, since it detects the use of the Upper case.

The next step in tagging the Named Entities consists of tagging the Arabic Named Entities. The same classification and the same tagset were adopted for the tagging task.

The dates were tagged through a module based on pattern recognition of Arabic dates

Proper Names tagging was done using a bilingual lexicon created from the list of Named Entities identified and extracted from the Spanish text. The Arabic equivalents were provided manually and thus a bilingual lexicon was created which is used in the following step for tagging the Arabic Named Entities conserving the same attributes (identification number and types) of those of their Spanish equivalents.

Regarding the Arabic Named Entities, we would like to point out these observations:

- Arabic language does not distinguish between Upper case and Lower case; hence, the only solution is to build a bilingual lexicon of Named Entities.
- The use of acronyms in Arabic is not common and in the majority of the cases where an acronym appeared in the Spanish text, it was translated by its full name in the Arabic Text. The exceptions to this rule are so few where the Arabic language adopts the transliteration of the acronym, such as in the case of KFOR an acronym referring to the Kosovo Forces. In this case, the Arabic translation adapted a transliteration (كفور).
- The occurrence of Named Entities in texts is usually quite unique and that is why it is used as anchor points. However due to the high frequency of Named Entities in this type of text (an average of 2.41 per sentence in the Arabic corpus), we reached the conclusion that tokens of Named Entities with the highest frequencies are not so significant when used as anchor points. In other words, the higher the frequency of the Named Entity, the less its significance as an anchor point. Named Entities such as "Council", "Security Council" and "President" are clear examples of this case. These Named Entities appears with a very high frequency in the different segments and thus they are less significant when used as anchor points.
- Finally, the occurrence frequency of Named Entities in the Spanish corpus is not identical to that in Arabic. The following cases were observed:

  - Sometimes, the Arabic text, for stylistic reasons, opts for the use of anaphors to avoid repetition, referring to the Named Entities in the form of pronouns.
  - In other cases due to syntactical reasons and questions of word order, the Arabic opts for using adjectives instead of Named Entities. For example, "la frontera entre Tayikistán y el Afganistán" (The borders between Tajikistan and Afghanistan). The Arabic translation did not use the two Named Entities *Tayikistán* and *Afganistan*, instead it used the adjective form to translate this phrase (الحدود الطاجيكية الأفغانية), which is equivalent to saying the *Tajik-Afghan border* instead of being translated into "*The border between Tajikistan and Afghanistan"*.

4.2 The Alignment Algorithm

The algorithm can be defined as follows:

- The parallel texts are preprocessed (tokenized and segmented). The output of this step is a segmented text with the structural units (paragraphs and sentences) tagged.
- The Named Entities are tagged in the Spanish corpus.
- Once the Named Entities are tagged in the Spanish corpus, we proceed with the Arabic Named Entities tagging as described above
- The alignment process is carried out in several stages:
- It searches the corpus for anchor points (Brown et. al, 1991; Chen, 1993). They are points with high probability of being mutual translations, since they have similar positions in the text and they are very short segments with one Named

Entity or maximum two. These criteria coincide with the headings and subtitles in the UN documents.

- The program, then, follows in three successive stages.

► First, it operates between the established anchor points in a way that it makes sure that in case of errors, these errors are not extended in the rest of the alignments.

► Second, it locates sentences whose Named Entities match and whose lengths are consistently correlated.

► Finally, sentences with no anchor points or sentences which failed to be aligned in the previous stages are passed again to be aligned, but this time depending exclusively on the statistical information. This is done adapting the model of Gale and Church (1991).

5. Results and Evaluation

The results provided by the alignment program are evaluated according to a golden standard, which consists of a sample corpus aligned manually. The golden standard consists of 307 Arabic sentences aligned to a total of 300 sentences in Spanish.

The results of the alignment program were satisfactory, considering it is a first approach to the subject.

According to the golden standard, there are a total of 301 alignments, with 13 cases of multiple alignments distributed as follows:

- Ten cases of 2-1 alignments (from Arabic to Spanish)
- Three cases of 1-2 alignments (from Arabic to Spanish).

From a total of 301 alignments (according to the golden standard), the system correctly aligned a total of 292 and failed to align 9 cases. Most of these errors were registered in cases of multiple alignments:

- 6 failures in cases of 2-1 alignments (Arabic-Spanish), thus reaching only 40 % accuracy in 2-1 alignments;
- All cases of 1-2 (Arabic Spanish).

However, the 1-1 sentence alignment reached a high accuracy rate 99,65 %, with an error rate of 0,35 %.

	Golden Standard	Automatic Alignment	Accuracy Rate in %	Error Rate in %
Total Alignments	301	292	97,01	2,99
1-1 Alignments	288	287	99,65	0,35
2-1 (Ar-Es) Alignments	10	4	40	60
1-2 (Ar-Es) Alignments	3	0	0	100

Table 2: Alignment Results

The final aligned text is passed to a module, which converts it into a XML file according to the TMX standard (Translation Memory Exchange). The following figure shows a snapshot of the final output

```
<tu tuid="205" datatype="Text">
  <tuv xml:lang="AR">
    <seg>
                                        وزاررئيسا إندونيسيا و البرتغال و
                                        رئيس وزراء أيرلندا الإقليم
                                        خلال تلك الفترة.
    </seg>
  </tuv>
  <tuv xml:lang="ES">
    <seg> Los Presidentes de Indonesia y
Portugal y el Primer Ministro de Irlanda
visitaron el Territorio en ese período.
    </seg>
  </tuv>
</tu>
<tu tuid="206" datatype="Text">
  <tuv xml:lang="AR">
    <seg>
                                        وأشير
                                        إلى أن الحالة الاقتصادية والاجتماعية
                                        كانت
                                        موضع قلق المتكلمين في تلك الجلسة.
    </seg>
  </tuv>
  <tuv xml:lang="ES">
    <seg> En la sesión se señaló que
causaba preocupación la situación económica
y social imperante.
    </seg>
  </tuv>
</tu>
```

Figure1: Output (XML file)

6. Conclusions and future work

The alignment experiment proved to be a successful one. This fact shows the feasibility of the approach and the utility of using previously developed tools for other languages as a starting point for developing new ones. In this way, it is obvious that we can perform the alignments with minimal linguistic resources and thus contrasting the opinion of Choueka et. al, which states that when dealing with Semitic languages, no statistical procedures, especially alignment is possible without some normalizing pre-processing; mainly lemmatization (2000).

The results of this work would serve as a starting point for developing alignment tools for more noisy texts, besides it can be tested on other language pairs such as Arabic-English or Arabic-French making it possible to establish a comparison between the different results. On the other hand, we will proceed with our study to analyze the linguistic nature of the Named Entities in both Arabic and Spanish.

The resulting output could provide valuable resources for a variety of applications including translation memories, cross language information retrieval and as a useful resource that can be incorporated in applications for Foreign Language Acquisition.

Acknowledgements

This work was funded by a scholarship from the Spanish Agency of International Cooperation AECI.

References

Attia, M. (2000). A Large-Scale Computational Processor of the Arabic Morphology, and Applications, Ph.D. Thesis. Cairo University.

Brown, P., Lai, J. and Mercer, R. (1991). "Aligning sentences in Parallel Corpora". In *Proceedings of the Association for Computational Linguistics*, Berkeley, pp.169-176.

Cabezas, C., Dorr, B. and Resnik, P. (2001). "Spanish language processing at University of Maryland: Building infrastructure for multilingual applications". *In Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies (SLPLT-2)*. Available at
ftp://ftp.umiacs.umd.edu/pub/bonnie/slplt-01.htm

Chen, S.F. (1993). "Aligning Sentences in Bilingual Corpora using Lexical Information". In *Proceedings of the Meeting of the Association for Computational Linguistic*s, pp. 9-16.

Choueka, Y.; Conley, E.S. and Dagan, I. (2000). A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English, Véronis, Jean (ed.): Parallel Text Processing. Alignment and Use of Translation Corpora, Kluwer Academic Publishers: Dordrecht / Boston / London, 2000, pp. 69- 96.

Darwish, K. (2002). "Building a Shallow Arabic Morphological Analyzer in One Day". In *Proceedings of the Association for Computational Linguistics (ACL-02), 40th Anniversary Meeting*, pp. 47-54.

Darwish, K. and Oard, D.W. (2003). "Evidence Combination for Arabic-English Retrieval". In *CLIR Experiments at Maryland for TREC-2002*, University of Maryland, College Park.

Diab, M. (2004a). "An Unsupervised Approach for bootstrapping Arabic Sense Tagging" In *Proceedings of Arabic Script Based Languages Workshop*, Coling.

Diab, M., Hacioglu, K. and Jurafsky, D. (2004b). "Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks" In Proceedings of HLT-NAACL, Boston.

Dichy, J. (2001). "On Lemmatization in Arabic - A Formal Definition of the Arabic Entries of Multilingual Lexical Databases". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects (ACL/EACL2001 workshop), 39th Annual Meeting and 10th Conference of the European Chapter, (ACL2001)*, Toulouse, France.

Elkatib, S. and Black, W. J. (2001). "Towards the Design of English-Arabic Terminological and Lexical Knowledge Base". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects (ACL/EACL2001 workshop), 39th Annual Meeting and 10th Conference of the European Chapter, (ACL2001)*, Toulouse, France.

Gale, W.A. and Church, K.W. (1991). "A Program for Aligning Sentences in Bilingual Corpora". In *Proceedings of 29th Annual Meeting of the ACL*, pp.177-184.

Goweder A. and De Roeck, A. (2001). "Assessment of a Significant Arabic Corpus". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects (ACL/EACL2001 workshop), 39th Annual Meeting and 10th Conference of the European Chapter, (ACL2001)*, Toulouse, France.

Grefenstette, G.(1999). Tokenization. In van Halteren (ed.) Syntactic Wordclass Tagging. Dordrecht, Kluwer.

Khoja, S. (2001): "APT: Arabic Part-of-speech Tagger". In *Proceedings of the Student Workshop at the "Second Meeting of the North American Chapter of the Association for Computational Linguistics" (NAACL2001),* Carnegie Mellon University, Pittsburgh, Pennsylvania.

Lelubre, X. (2001). "A Scientific Arabic Terms Data Base: Linguistic Approach for a Representation of Lexical and Terminological Features". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects (ACL/EACL2001 workshop), 39th Annual Meeting and 10th Conference of the European Chapter, (ACL2001)*, Toulouse, France.

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates*.

McEnery, T. (1997). "Multilingual Corpora-Current
Practice and Future Trends". In *13 th ASLLB Machine Translation Conferenc*e, London, pp. 75-86.

Melamed, I. D. (2001). Empirical Methods for Exploiting Parallel Text, Cambridge/London: MIT Press.
Practice and Future Trends". En 13th ASLLB Machine

Resnik, P.and Smith, N. 2003. "The Web as a Parallel Corpus". In *Computational Linguistics* 29(3).

Rezaei, S. (2001). "Tokenizing an Arabic Script Language". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects (ACL/EACL2001 workshop), 39th Annual Meeting and 10th Conference of the European Chapter, (ACL2001)*, Toulouse, France.

Samy,D., Moreno Sandoval, A. and Guirao, J.M. (2004). "Construction of a Bilingual Lexicon of Verbs Spanish-Arabic based on a Parallel Text". In *Proceedings of the Fourth International Conference of Language Resources and Evaluation (LREC 2004),* Lisbon, pp.1571-1574.

Somers, H. (2001). "Bilingual Parallel Corpora and Language Engineering". In *Anglo Indian Workshop "Language Engineering for South Asian Languages" LESAL*, Mumbai.