Improving the Accuracy of English-Arabic Statistical Sentence Alignment

ARTICLE *in* INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY · APRIL 2011

Impact Factor: 0.58 · Source: DBLP

3 AUTHORS:

Mohammad Salameh
University of Alberta
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Rached Zantout
Rafik Hariri University
31 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Nashat Mansour
Lebanese American University
79 PUBLICATIONS **805** CITATIONS
SEE PROFILE

Improving the Accuracy of English-Arabic Statistical Sentence Alignment

Mohammad Salameh[1], Rached Zantout[2], and Nashat Mansour[1]
[1]Department of Computer Science and Mathematics, Lebanese American University, Lebanon
[2]College of Computer and Information Sciences, Prince Sultan University, Saudi Arabia

Abstract: *Multilingual natural language processing systems are increasingly relying on parallel corpus to ameliorate their output. Parallel corpora constitute the basic block for training a statistical natural language processing system and creating translation and language models. Several systems have been devised that automatically align words of a pair of sentences, each in a language. Such systems have been used successfully with European languages. In this paper, one such system is used to align sentences in an English-Arabic corpus. The system works poorly given raw unaligned sentence English-Arabic sentence pairs. This prompted the development of a preprocessing step to be applied to the Arabic sentences. The same corpus was then preprocessed and a significant improvement is reported when alignment is attempted using the preprocessed unaligned sentences.*

Keywords: *Word alignment, sentence alignment, parallel corpora, and statistical natural language processing.*

Received January 15, 2009; accepted August 3, 2009

1. Introduction

Computing research and applications for the Arabic language has recently increased. This is due mainly to the increase in the number of Arab internet users who do not master other languages [7]. Another reason for the interest in Arabic is non-Arab (usually European and American) interest in Arab countries. Arabic is more difficult to treat compared to European languages mainly because of its rich morphology.

There are many difficulties for machine treatment of Arabic text compared to treating English text [5]. First, a single word in Arabic can have many variations depending on the morphological variations that it can undertake. Like English, Arabic adds prefixes and suffixes to a word to form other variants. However, unlike English, Arabic can also add infixes to words. This makes algorithms for English morphological analysis not applicable to Arabic. Second, the prefixes and suffixes of a word may not always be attached to the other letters in the same word. Third, in Arabic a whole English sentence can be represented using one single word (e.g., "Should we then force it upon you" is translated to "أفنلزمكموها"). Fourth, unlike English, the prepositions and pronouns are not separate words. Prepositions and pronouns in Arabic are usually attached to the word (for example, the 2-word English phrase "his book" is translated to a single word "كتابه" in Arabic). Fifth, a single English word can have a multiword Arabic translation (e.g., "غير مزود بالسلاح" is translated to "unarmed"). Sixth, spaces, in Arabic, might not separate two words from each other. For example, the conjunction و is usually written without a space between it and the next word. Seventh, the word order in an Arabic sentence is usually different from that in the English sentence. In fact, the word order might change in different Arabic translations of the same English sentence. For example, the sentence "The boy went to school" can be translated into " ذهب الولد الى المدرسة" or " الى المدرسة ذهب الولد" or even to " الولد ذهب الى المدرسة". Eighth, Arabic sentences might not contain any verbs.

Recently, Natural Language Processing (NLP) systems in general and Machine Translation (MT) systems in particular have improved their output quality by resorting to statistical techniques. A prerequisite for the use of any statistical technique is the availability of appropriate data. For example, Statistical Machine Translation (SMT) systems require the existence of a Language Model (LM) and a Translation Model (TM). Such models are usually built by automatic processing of multilingual corpora. Multilingual corpora are a large number of texts translated to all languages supported by the corpus. Before such corpora can be used, they have to be aligned so that a word in the text in one language is linked to the corresponding word in the text of the other language. Several systems exist that can do the alignment automatically by processing a large unaligned corpus. One such system is the GIZA++ system [13] that has been proven to work well for corpora of parallel European languages texts.

In this paper, we review related literature in section 2. The results of trying to use GIZA++ directly on an English-Arabic corpus are reported in Section 3. These results show that GIZA++ does not work well when

raw Arabic text is used. The complexity of the Arabic language is proven to be an obstacle for statistical automatic sentence aligners like GIZA++. In Section 4, a solution is detailed that improves the performance of GIZA++. In this solution, the Arabic and English texts are preprocessed before being handled by GIZA++. Guidelines for preprocessing the Arabic part of the multilingual corpus are presented and proven to ameliorate the automatic alignment process. In Section 5, the paper is concluded with a summary of the achievements and suggestions for future research.

2. Better Alignment in Literature

Alignment of multilingual corpora is usually divided up into sentence alignment and word alignment. Sentence alignment identifies correspondences between sentences in one language and sentences in another language. The basic units for segmenting parallel texts in sentence alignment are paragraphs and sentences. It is usually assumed that both parallel texts have the same number of paragraphs. Sentences inside paragraphs do not necessarily have a one-to-one alignment in a multilingual corpus. A sentence in a source language might be translated into 2 or more sentences in the target language. Similarly 2 or more source language sentences might be translated into a single target language sentences. Several clues in a text can aid in aligning parallel texts such as titles (including chapter, section, table and figure titles), cells inside a table, and items of an enumeration (items usually listed each on a separate line). Sentence alignment techniques vary from length based models to lexical based models or a combination. Length based techniques mainly rely on comparing the length of the sentences and therefore are considered knowledge poor [1]. Such approaches achieve high accuracy for many European languages [3]. Length based approaches that use the numbers of Characters in a sentence were proven to work better than those that use the number of words [2, 3]. Lexical based methods are considered knowledge-rich [10]. Anchor words are chosen from the pair and checked whether they correspond to each other. One of the methods relies on calculating the probability of word pairs by checking the frequency of words in the source sentence and trying to find how many words in the target sentence correspond to their translations. Another method is to find chains of corresponding vectors after mapping the text into a two dimensional space [10]. Combinations of length-based and lexical-based sentence alignment techniques exist [15]. String similarity measures and machine readable dictionaries can be used to enhance the sentence alignment. Word frequencies and occurrences are checked to achieve good results. K-vec is one of the approaches where sentence pairs are split into equal segments and parallel segments are checked if they have words with similar meanings [9].

Word alignment is the process of linking corresponding words and phrases in a parallel text. The aim is to extract the maximum number of corresponding sets from the parallel text that can be used in a desired application. Factors that affect word alignment are Structural, lexical, and grammatical differences between the languages; morphological differences between languages; and spelling errors inside the parallel text. Word alignment techniques are divided into two groups: the Association approaches and the Estimation approaches. Association approaches use heuristics to perform the word alignment [12] while Estimation approaches use statistical rules [9]. Preprocessing to ameliorate SMT quality has been proven to be effective for morphologically rich languages such as German [11], Spanish, Catalan, and Serbian [14], and Czech [4]. As far as Arabic is concerned [8] and [6] showed that morphological preprocessing helps only in the case of small corpora.Using GIZA++ without preprocessing.

3. Properties of the Corpus Used

The corpus used in this paper was collected from the UN documents and resolutions of the past ten years. It contains 2154 English-Arabic sentence pairs whose properties are shown in Table 1.

Most of the used terms are related to the political and geographical context of the countries that the resolutions were addressing (Palestine, Lebanon, Syria, Iraq, Afghanistan, and African countries). The documents contain many political terms and named entities (names of places, political figures, months, and chemicals). In addition, the sentences in the documents contain many abbreviations for names of UN projects, offices and other entities that are translated into meaningful complete Arabic phrases in the Arabic version of the texts.

Table 1. Propertries of used corpus.

Number of Pairs	2154
Maximum Number of Words in an English Sentence	132
Maximum Number of Words in an Arabic Sentence	175
Maximum Ratio of English to Arabic Sentence	1.5
Minimum Ratio of English to Arabic Words	0.291
Average Ratio of English to Arabic Words	0.706

For example: "UNMOVIC" in English is translated to "لجنة الامم المتحدة للرصد والتحقق و التفتيش" in Arabic. In the corpus, a small number of the English sentences were translated to more than one Arabic sentence. Those sentences were either split if it was proper to do so or ignored. The sentences in the corpus were prepared for GIZA++ by transforming them into English-Arabic parallel sentences in an XML file. The

XML file as shown in Figure 1 consists of <sentence> nodes. Each <sentence> node has 3 elements: <ID> the serial number of the English-Arabic sentence pair, <EN> the English sentence, <Ar> the Arabic translation of the English sentence.

```
<sentence>
  <ID>45</ID>
   <En>
  Welcomes the continued contribution of UNIFIL to operational demining
</En>
   <Ar>
     رحب بالمساهمة المستمرة لقوة الأمم المتحدة المؤقتة في لبنان في عمليات إزالة الألغام
</Ar>
</sentence>
```

Figure 1. Example of a sentence node.

The XML files were then processed by Giza++. In many sentence pairs, most of the English words were aligned to NULL, 36 out of 58 in the example in Figure 2. 44.72% of the English words were aligned to NULL. Only 22.32% of the English words were aligned correctly. Only 10.72% of the Arabic words were aligned to NULL and 36.04% of the Arabic words were aligned correctly with their English counterparts. Going through the words that were not aligned properly many problems were discovered. The first problem was that many English words have more than one Arabic translation. GIZA++ treated each translation as a separate word even if the meaning of the word was still the same. The second problem was that the same Arabic word appeared in different forms inside the corpus. GIZA++ treated those occurrences as separate words and not multiple occurrences of the same word. The third problem was the existence of large (more than 100 words) sentences in both the English and Arabic texts. GIZA++ sets a limit of 100 words per sentence to process it correctly. The fourth problem was that most of the English stopwords (words that are used to divide long sentences into smaller chunks) have more than one translation in Arabic. In some of the cases, some of the English stopwords did not have a corresponding independent word in the Arabic text.

Figure 3 shows that the stopword "which" has no corresponding independent word in the Arabic sentence.

4. Ameliorating GIZA++ Output

In order to ameliorate GIZA++'s results, preprocessing was done on the Arabic and English texts. The only preprocessing on the English text was done by transforming all letters into lowercase. The first preprocessing step for Arabic text was to remove diacritics from the Arabic text. This was done by filtering all the kashida's, transforming all forms of the Arabic Alif ("أ" and "إ") to one standard form, the "ا", separating the Arabic coordinating conjunction "و" from Arabic words by a space, separating the punctuation marks from the words around them and breaking up compound words in English and Arabic sentences into their constituent words. The second preprocessing step was to manually separate the prefixes and suffixes of each Arabic word from its body by spaces. This meant that now the prefix, the suffix and the body of one Arabic word are now three words instead of one. The previous preprocessing steps can be classified as character-level processing and word-level processing and were done in order to increase the cardinality of the Arabic root words in the corpus. The third preprocessing step can be classified as sentence-level which was to reduce the length of long sentences. This was done because GIZA++ produces better results with shorter sentences. Shorter sentences are produced by breaking up long sentences, in both Arabic and English texts, at the stopwords and punctuation marks.

```
# Sentence pair (340) source length 58 target length 47 alignment score : 8.04656e-73

وقد قررت المحكمة، بما لا يدع مجالا للشك أن إسرائيل ملزمة بوضع حد لانتهاكاتها للقانون الدولي، وبوقف تشييد الجدار الذي تبنيه في الأرض الفلسطينية المحتلة، بما فيها داخل القدس الشرقية وما حولها، وبتفكيك الهياكل المقامة في تلك المنطقة، وبإلغاء أو إبطال جميع القوانين التشريعية والتنظيمية المتصلة به

NULL ({ 22 36 }) The ({ 1 }) Court ({ 2 }) has ({ }) determined ({ }) beyond ({ }) any ({ }) doubt ({ 3 4 5 6 7 8 }) that ({ 9 }) Israel ({ 10 }) is ({ }) under ({ }) obligation ({ }) to ({ }) terminate ({ }) its ({ }) breaches ({ }) of ({ }) international ({ }) law, ({ }) to ({ }) cease ({ }) the ({ }) construction ({ }) of ({ }) the ({ }) wall ({ 18 19 }) being ({ 20 }) built ({ 21 }) in ({ }) the ({ }) Occupied ({ 23 }) Palestinian ({ 24 }) Territory, ({ 25 }) including ({ 26 27 }) in ({ }) and ({ }) around ({ 28 }) East ({ }) Jerusalem, ({ 29 30 }) to ({ }) dismantle ({ }) the ({ }) structure ({ }) therein ({ }) situated ({ 11 12 13 14 15 16 17 31 32 }) and ({ }) to ({ }) repeal ({ 33 34 35 37 38 39 }) or ({ 40 }) render ({ }) ineffective ({ 41 }) all ({ 42 }) legislative ({ 43 44 }) and ({ }) regulatory ({ 45 46 }) acts ({ }) relating ({ }) thereto ({ 47 })
```

Figure 2. Example of many words aligned to NULL.

```
<sentence>
  <ID>1901</ID>
  <En>The experts produced conflicting reports, however, which further compounded the stalemate</En>
<Ar> ازمة ال تعميق الى ت ادى ة متناقض تقارير وا اعد خبراء ال ان غير </Ar>
</sentence>
```

Figure 3. Missing stopwords.

The prefixes in Arabic words that were identified in this paper were:

- 1-letter Prefixes: "ل", "ب", "ف", "س", "و", "ي", "ت", "ن", "ا".
- 2-letter Prefixes: "كال", "بال".
- 3-letter Prefixes: "لل", "ال".

The suffixes in Arabic words that were identified in this paper were separated into two level suffixes:

- 1st level suffixes:
- 1-letter Suffixes: "ة", "ي", "ه", "ت", "ا", "ن", "كَ".
- 2-letter Suffixes: "تن", "كم", "هن", "نا", "يا", "ها", "هم", "ما", "ني", "كن", "تم".
- 2nd level suffixes: if the root of the word is a verb then separate the suffix (either "ان", "ون", "ين" or "ا" or "وا") from the verb by a space.
- If the root of the word is a noun then separate the suffix (either "ان", "ون", "ين" or "ات") from the noun by a space. As a special case, some nouns end in Alef Al-Tanween, in this case remove Alef Al-Tanween.
- If the root of the word is a relative adjective or a noun in their feminine form then separate the "ة" from the Adjective by a space.

As an example, after performing the split on the Arabic words, a sentence like: لكل إنسان الحق، على قدم المساواة التامة مع الآخرين، في أن تنظر قضيته أمام محكمة مستقلة نزيهة نظراً عادلاً علنياً للفصل في حقوقه والتزاماته وأية تهمة جنائية توجه إليه becomes ل كل إنسان ال حق ، على قدم ال مساواة ال تام ة مع ال آخر ين ، في ان تنظر قضية ه امام محكمة مستقل ة نزيه ة نظر عادل علني لل فصل في حقوق ه و التزام ات ه و اي ة تهمة جنائي ة توجه إلى ه

In the sentence level processing phase, three elements are added to each <sentence> node namely, <EnLen>, <ArLen>, <Ratio> as shown in Figure 4. The <EnLen> and <ArLen> fields contain the number of words in the English and Arabic sentences respectively. A "word" in this paper is the set of characters that are preceded and followed by a space. <Ratio> is the <EnLen> divided by <ArLen> and is used to decide whether the split position for a long sentence is good.

```
<sentence>
<ID>54</ID>
<En>It should be noted that , during the second terrorist attack near the Canal Hotel on 22 September , two UNMOVIC local staff were injured</En>
<Ar>ا اصيب قد ين محلي ال لجنة ال ين موظف من اثنين ان ملاحظة ال ب جدير ال من و ايلول 22 في قناة ال فندق قرب وقع الذي ثاني ال ارهابي ال هجوم ال خلال </Ar>
<EnLen>25</EnLen>
<ArLen>36</ArLen>
<Ratio>0.694444444444444</Ratio>
</sentence>
```

Figure 4. Updated sentence node.

The choice of the stop words in English and Arabic languages is important. Three criteria were manually used to decide whether an English-Arabic pair is a stopword. First, a stopword should appear frequently in the text. Second, an English stopword should have an Arabic translation that is also an Arabic stopword. Third, splitting English and Arabic sentence pairs at their stopwords should produce sentence pairs that have a high correspondence between their words.

After analyzing the commonly used English stopwords in the literature, several conclusions were attained. First, each English stop word has more than one translation in Arabic. For Example: "in order" is translated to "بغية", " حتى", "من اجل", "كي" and " ذلك ل". Second, many English stop words have the same translation in Arabic. For example, "within" and "during" are both translated to "خلال". Third, the same Arabic stopword (especially prepositions) can be the translation of many English stopwords. Fourth, some Arabic stopwords may not have any corresponding words in the English translation. Fifth, it is always easier to search for the stopwords in the English sentence first and then search for the corresponding stopword in the Arabic sentence. Sixth, possessive pronouns always precede the noun in English while they appear as suffixes that are attached to the noun in Arabic.

The preceding conclusions led to the exclusion of many English stopwords. Among the excluded stopwords are demonstrative pronouns (such as 'this', 'that', 'these', 'those'), indefinite Pronouns (such as 'anything', 'anybody', 'anyone', 'something', 'somebody', 'someone', 'nothing', 'nobody', 'none', 'no one'), possessive pronouns (such as 'mine', 'yours', 'his', 'hers', 'ours', 'theirs'), simple pronouns (such as 'I', 'you', 'he', 'she', 'it', 'we', 'they', 'me', 'him', 'her', 'us', 'them'), reflexive pronouns (such as 'self', 'myself', 'yourself', 'himself', 'herself', 'itself', 'oneself', 'ourselves', 'yourselves', 'themselves'), verb forms (such as 'am', 'are', 'is', 'was', 'were', 'be', 'being', 'been', 'has', 'have', 'had', 'having', 'can', 'could', 'want', 'wants', 'wanted', 'shall', 'should', 'will', 'would', 'may', 'might', 'must', 'ought', 'do', 'does', 'doing', 'did', 'done', 'make', 'makes', 'making'), coordinating conjunctions (such as 'and', 'but', 'or'), prepositions with very high frequency (such as 'under, 'from', 'in', 'by', 'for', 'upon', 'with', 'on' 'of', 'within', 'to' and 'at' which can be translated to the following arabic prepositions: 'من', 'في', 'الى', 'على', 'ب' and 'ل').

The stopwords that were used are the subordinating conjunctions shown in Table 2. Those stopwords achieve a fair split in English and Arabic sentences.

A split is considered successful if the ratio of the positions of the English stopword and its Arabic translation (Arabic stopword) is close to the ratio of word counts in both sentences (within a certain threshold value of 0.21 which was determined experimentally as a good value). Aside from using stopwords, punctuation marks (especially the comma) were also used to breakup long sentences. The split according to commas uses a threshold value of 0.12 (that also was determined experimentally).

Table 2. Stopwords used in the paper.

Particularly	لا سيما|خاصة	including	بما في
Between	بين	in order	كي|من اجل|حتى|بغية|ذلك ل
Before	قبل	which	مما|الذي|التي
Therefore	لذلك|اذن	although	بالرغم|على الرغم من
Though	رغم ان|و ان	when	حين|فيما
If	حتى|لو|اذا	while	في حين|بينما|في ما لا يزال
Because	نتيجة|نظر لان	unless	ما لم
Whenever	حيثما|كلما|متى	but	لكن
Within	خلال|داخل	during	خلال|اثناء
Across	عبر	under	تحت
According	طبق|وفق	prior	قبل|سابق
When	عندما|عند|بعدم	accordance	على النحو|وفق|عمل
Especially	لا سيما|خاصة	but also	بيد|لا بل|اذ
Rather	بدل		

Table 3 shows the results of successful splits. We can conclude from Table 3 that all English sentences have been transformed into ones that have less than 100 words. Table 4 shows that unsuccessful splits happened only with English sentences that have less than 100 words. This means that all English sentences that have more than 100 words have been split successfully.

Table 3. Successful splits.

Statistics about Sentences after Successful Split	
Maximum Number of Words in an English Sentence	91
Maximum Number of Words in an Arabic Sentence	147
Maximum Ratio of English to Arabic Words	2.5
Minimum Ratio of English to Arabic Words	0.26
Average Ratio of English to Arabic Words	0.738

The average ratio of English to Arabic words did not change drastically before and after the splits which means that the structure of the corpus was not changed by splitting. Around 13% of commas in English sentences either disappeared in Arabic translations or were replaced by conjunctions. After splitting, the 2154 pairs of English-Arabic sentences became 3636. GIZA++ was able to align 22.32% of the English words when the original text (without preprocessing) was used. With preprocessing, this percentage rose to about 45.72%. Before preprocessing 44.7% of the English words were not aligned to any Arabic words. This percentage dropped down to 20.9% after reprocessing. The words that were not correctly aligned even after preprocessing were the words that had a low frequency in our corpus. Table 5 gives an example of a sentence and its alignment before and after preprocessing. Before preprocessing, 11 English words were aligned correctly while after preprocessing, 23 English words were aligned correctly.

Table 4. Unsucsessful splits.

Statistics about Sentences that were not Split	
Maximum Number of Words in an English Sentence	92
Maximum Number of Words in an Arabic Sentence	155
Maximum Ratio of English to Arabic Words	1.5
Minimum Ratio of English to Arabic Words	0.29
Average Ratio of English to Arabic Words	0.709

5. Conclusions and Future Work

In this paper, the aim was to prepare the English-Arabic corpus for correct processing by the alignment software. This was done by increasing the frequency of Arabic words through character and word level processing. Another step was to decrease the length of English-Arabic sentence pairs by splitting them into smaller phrases using stopwords and commas. These preprocessing steps resulted in a 100% improvement in alignment and number of unaligned unaligned words. Many aspects of the work presented in this paper can be explored further. First, an automatic way for choosing the best stopwords from English and Arabic text should be devised. It is our belief that the stopwords will be function of the type of text and therefore stopwords used in literary text might be different from stopwords used in newspaper articles. Second, new ways for reducing the sizes of sentences should be developed. Currently, we rely on stopwords and commas. It remains to be seen whether including all punctuation marks and morphological information might ameliorate the splitting of long sentences. Third, Arabic particles and prepositions around the words should be used to improve the alignment rather than considering those particles and prepositions as independent entities. Fourth, dictionary entries and/or morphological information should be used to check the alignment produced by GIZA++ or to suggest alignment to GIZA++. A word pair that was already aligned in previous runs should be used to better the alignment of future runs. Fifth, automating the preprocessing steps described in this paper will enable the use of larger corpora that are commensurate with the corpora used for European languages.

Table 5. Improvements due to preprocessing.

Without Pre-processing		# Sentence pair (340) source length 58 target length 47 alignment score : 8.04656e-73 وقد قررت المحكمة، بما لا يدع مجالا للشك أن إسرائيل ملزمة بوضع حد لانتهاكاتها للقانون الدولي، وبوقف تشييد الجدار الذي تبنيه في الأرض الفلسطينية المحتلة، بما فيها داخل القدس الشرقية وما حولها، وبتفكيك الهياكل المقامة في تلك المنطقة، وبإلغاء أو إبطال جميع القوانين التشريعية والتنظيمية المتصلة به NULL ({ 22 36 }) The ({ 1 }) Court ({ 2 }) has ({ }) determined ({ }) beyond ({ }) any ({ }) doubt ({ 3 4 5 6 7 8 }) that ({ 9 }) Israel ({ 10 }) is ({ }) under ({ }) obligation ({ }) to ({ }) terminate ({ }) its ({ }) breaches ({ }) of ({ }) international ({ }) law, ({ }) to ({ }) cease ({ }) the ({ }) construction ({ }) of ({ }) the ({ }) wall ({ 18 19 }) being ({ 20 }) built ({ 21 }) in ({ }) the ({ }) Occupied ({ 23 }) Palestinian ({ 24 }) Territory, ({ 25 }) including ({ 26 27 }) in ({ }) and ({ }) around ({ 28 }) East ({ }) Jerusalem, ({ 29 30 }) to ({ }) dismantle ({ }) the ({ }) structure ({ }) therein ({ }) situated ({ 11 12 13 14 15 16 17 31 32 }) and ({ }) to ({ }) repeal ({ 33 34 35 37 38 39 }) or ({ 40 }) render ({ }) ineffective ({ 41 }) all ({ 42 }) legislative ({ 43 44 }) and ({ }) regulatory ({ 45 46 }) acts ({ }) relating ({ }) thereto ({ 47 })
With Preprocessing	1	# Sentence pair (746) source length 20 target length 27 alignment score : 7.59922e-42 قد قرر ت ال محكمة ، بما لا يدع مجالا لل شك ان اسرائيل ملزم ة بوضع حد ل انتهاك ات ها لل قانون ال دولي ، NULL ({ 4 16 22 25 }) the ({ }) court ({ 3 5 }) has ({ 1 }) determined ({ 2 }) beyond ({ 6 }) any ({ }) doubt ({ 7 8 9 10 11 12 }) that ({ 13 }) israel ({ 14 }) is ({ }) under ({ }) obligation ({ 15 }) to ({ }) terminate ({ 17 18 }) its ({ }) breaches ({ 19 20 21 23 }) of ({ }) international ({ 26 }) law ({ 24 }) , ({ 27 })
	2	# Sentence pair (749) source length 6 target length 12 alignment score : 3.14924e-18 ها داخل ال قدس ال شرقي ة و ما حول ها ، NULL ({ 3 5 7 8 }) in ({ }) and ({ }) around ({ 1 2 9 10 11 }) east ({ }) jerusalem ({ 4 6 }) , ({ 12 })
	3	# Sentence pair (747) source length 15 target length 18 alignment score : 8.35513e-23 و ب وقف تشييد ال جدار الذي تبنيه في ال ارض ال فلسطيني ة ال محتل ة ، NULL ({ 1 2 5 10 12 15 }) to ({ }) cease ({ 3 }) the ({ }) construction ({ 4 6 }) of ({ }) the ({ }) wall ({ }) being ({ }) built ({ 7 8 }) in ({ 9 }) the ({ }) occupied ({ 11 13 14 16 17 }) palestinian ({ }) territory ({ }) , ({ 18 })
	4	# Sentence pair (750) source length 19 target length 30 alignment score : 1.25909e-38 و بتفكيك ال هياكل ال مقامة في تلك ال منطقة ، و ب الغاء او ابطال جميع ال قوانين ال تشريعي ة و التنظيمي ة ال متصل ة ب ه NULL ({ 1 3 5 7 9 18 20 26 }) to ({ }) dismantle ({ }) the ({ }) structure ({ }) therein ({ 2 4 6 }) situated ({ 8 10 11 }) and ({ 12 }) to ({ }) repeal ({ 13 14 }) or ({ 15 }) render ({ }) ineffective ({ 16 19 }) all ({ 17 }) legislative ({ 21 22 }) and ({ 23 }) regulatory ({ }) acts ({ }) relating ({ }) thereto ({ 24 25 27 28 29 30 })

References

[1] Aswani N. and Gaizaukas R., "A Hybrid Approach to Align Sentence and Words in English Hindi Parallel Corpora," *in Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 67-64, 2005.

[2] Brown P. and Lai J., "Aligning Sentences in Parallel Corpora," *in the Proceedings of 29th Annual Meeting for ACL*, pp. 169-179, 1991.

[3] Gal W. and Church K., "A Program for Aligning Sentences in Bilingual Corpus," *in the Proceedings of 29th Annual Meeting of the ACL*, pp. 177-184, 1991.

[4] Goldwater S. and McClosky D., "Improving Statistical MT through Morphological Analysis," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP*, pp. 159-163, 2005.

[5] Guessoum A. and Zantout R., *Arabic Morphological Generation and its Impact on the Quality of Machine Translation to Arabic*, Antal Vanden Bosch and Guenter Neumann, 2007.

[6] Habash N. and Sadat F. "Arabic Preprocessing Schemes for Statistical Machine Translation," *in Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 49-52, 2006.

[7] Hamandi L., Damaj I., Zantout R., and Guessoum A., "Parallelizing Arabic Morphological Analysis: Towards Faster Arabic Natural Language Processing Systems," *in Proceedings of CIBITIC*, Lebanon, pp. 455-459, 2006.

[8] Lee Y., "Morphological Analysis for Statistical Machine Translation," in *Proceedings of* of *the North American Chapter of ACL*, pp. 232-237, 2004.

[9] Melamed D., "A Geometric Approach to Mapping Bitext Correspondence," *Conference of Empirical Methods in NLP*, Philadelphia, pp. 1-12, 1996.

[10] Melamed D., "A Portable Algorithm for Mapping Bitext Correspondence*," in Proceedings of 35th Conference of Association for Computing Linguistics*, Spain, pp. 305-312, 1997.

[11] Nießen S. and Ney H., "Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information," *Computer Journal of Computational Linguistics*, vol. 30, no. 2, pp. 55-59, 2004.

[12] Osh F. and Ney H., "A Systematic Comparison of Various Statistical Alignment Models," *Computer Journal of Computational Linguistics*, vol. 29, no. 2, pp. 19-51, 2003.

[13] Osh J., GIZA++: Training of Statistical translation models, www.fjoch.com/GIZA++ .html, Last Visited 2008.

[14] Popovi´c M. and Ney H. "Towards the Use of Word Stems and Suffixes for Statistical Machine Translation," *in Proceedings of the Conference on Language Resources and Evaluation*, pp. 159-163, 2004.

[15] Simard M. and Foster G., "Using Cognates to Align Sentences in Bilingual Corpora," *in the Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Canada, pp. 67-81, 1992.

Mohamad Salameh received his MS and BS degrees in computer science from the Lebanese American University. He is currently working at exceed it services as a consultant for Microsoft share point server 2007 and project server 2007. His research interests are in natural language processing for Arabic language and computer graphics.

Rached Zantout received his BE from the American University of Beirut, Lebanon in 1988, his MSc from the University of Florida in 1990, and PhD from the Ohio State University in 1994, all degrees being in electrical engineering. He was a research associate and teaching associate for most of his graduate studies. Directly after finishing his PhD, he joined Scriptel Corporation and worked on several R&D projects to develop a new generation of graphic input devices. Between 1995 and 2000, he was an assistant professor at King Saud University in Riyadh (Saudi Arabia). Then he moved to Lebanon and taught at reputed Lebanese universities like the University of Balamand, American University of Beirut, Lebanese American University, Beirut Arab University, Hariri Canadian University. He is currently an associate professor at the Prince Sultan University. His research interests cover robotics, artificial intelligence, and natural language processing. He currently works on developing components for machine translation and natural language processing with a special focus on tools related to the Arabic Language. He also has active research in the area of autonomous robot navigation, computer vision and digital image processing.

Nashat Mansour is a professor of computer science at the Lebanese American University. He received BE and M Eng Sc degrees in electrical engineering from the University of New South Wales, and MS in computer engineering and PhD in computer science from Syracuse University. His research interests include software testing, application of evolutionary algorithms to real-world problems, protein structure prediction, and Arabic-related computing.