Use of Stemming for Information Retrieval

Mohamed Abdel Fattah, Fuji Ren, Shingo Kuroiwa
Faculty of Engineering, the University of Tokushima
2-1 Minamijosanjima
Tokushima, Japan 770-8506
(mohafi, ren, kuroiwa) @is.tokushima-u.ac.jp

Abstract

Arabic is a morphologically rich language that presents significant challenges to many natural language processing applications because a word often conveys complex meanings decomposable into several morphemes (i.e. prefix, stem, suffix). By segmenting words into morphemes, we could improve the performance of English/ Arabic translation pair's extraction from parallel texts. This paper describes an algorithm to automatically extract an English/ Arabic bilingual dictionary from parallel texts that exist in the Internet archive after using an Arabic light stemmer as a preprocessing step. Before using the Arabic light stemmer, the total system performance was 64.68%, then the system accuracy increased to 70.35% after applying the Arabic light stemmer on the Arabic documents.

The algorithm has certain parameters which values could be changed to increase the system precision and the number of extracted translation pairs. Like most of the systems done, the accuracy of our system is directly proportional to the number of sentence pairs used. But our system is able to extract translation pairs from a very small parallel corpus. This new system can extract translations from only two sentences in one language and two sentences in the other language if the requirements of the system accomplished. Moreover, this system is able to extract word pairs that are translation of each other and the explanation of the word in the other language as well. By controlling the system parameters, we could achieve 100% precision for the output bilingual dictionary. But the size of the dictionary will be smaller.

1- Introduction

Recent years saw an increased interest in researches and experiments conducted for Arabic language. Some researches were related to stemming and segmentation of Arabic words [1, 2, 3, and 4]. Some others dealt with cross-language information retrieval for query translation using several tools like bilingual dictionaries and machine translation systems [5, 6, 7, and 8].

Some researchers tried to construct bilingual dictionary from the Internet Web documents. Kumiko TANAKA &Kyoji UMEMURA used a third language to construct a bilingual dictionary [9].

Craig tried to extract English/ Spanish translation pairs from English/ Spanish parallel texts and he achieved precision of 18% [10]. Craig's system dealt with word to word translation only, whereas our new system deals not only with word to word translation, but also word to phrase as well. Also the precision and the total number of extracted translation pairs are higher for our system.

Jinxi Xu and Ralph Weischedel used a statistical machine translation toolkit called WEAVER developed by John Lafferty at Carnegie Mellon University to automatically extract translation pairs from the parallel corpora. WEAVER has a component to automatically derive word translations based on sentence-aligned [11]. While Ashish Venugopal presented a technique that begins with improved IBM models to create phrase level knowledge sources that effectively represent local as well as global phrasal context from parallel corpus [12]. Unfortunately there is no English-Arabic parallel corpus, so we extract bitexts automatically form the Internet archive.

The proposed system consists of the following modules:

1- *Extracting English / Arabic parallel documents from the Internet archive.*
2- *Preprocessing step to remove the English and Arabic stop list words from the documents and make stemming and also to make sentence to sentence alignment.*
3- *Constructing the bilingual dictionary.*

2- Extracting parallel documents from the Internet archive

Three steps are required to find parallel documents:

1- Locating the pages that might contain parallel documents,
2- Generating the document pairs that might be translation of each other,
3- Filtering-out of non-translation candidate pairs.

We simply sent queries to the AltaVista search engine. These queries contained some words like: "Arabic version", "English version", "Arabic", "English", "To Arabic", and "To English", in order to download the pages that might contain English / Arabic parallel documents. Then we collected two types of pages:

a- A ***parent page*** is one that contains hypertext links to different-language versions of a document.

b- A ***sibling page*** is a page in one language that itself contains a link to a version of the same page in another language.

In each page, we download all the files in the English page and all the files in the Arabic page as well. According to the documents name (for example an Arabic file called **exp_a38**, is most probable to be translation of the English file that called **exp_e38**), we gathered the English- Arabic document pairs that are probably parallel. Then if the following relation:

(0.8*(English document size) <= (Arabic document size) <= 1.2*(English document size)),

Verified for the English- Arabic document pair, we expect that this pair is parallel and we remove the document pairs that are not verifying the above relation. We have collected 3467 document pairs that are expected to be translation of each other. We filtered manually 378 document pairs that are actually translation of each other. These 378 document pairs contain 15853 sentence pairs.

3- Preprocessing

A- Sentence alignment for the 15853 sentence pairs, based on the sentence length.

B- Removing the English and Arabic stop word lists from the English and Arabic documents respectively. The English stop list contains words like possessive pronoun, pronouns, prepositions and some other words that have no counterparts in the other language such as: (the, an, a). The Arabic stop list contains pronouns, prepositions, and words like (ولقد, إن, أن) that have no counterparts in English language.

C- Preprocessing step for English and Arabic documents to make them suitable for the next step, for instance delete some symbols such as (, . ! ’ ‘ ; “), remove diacritics, and also remove the new lines or double spaces in the same sentence if exist.

D- Convert all plural English words to singular.

E- Applying the “Aitao Chen and Fredric Gey” Arabic light stemmer on the Arabic documents [1]. The stemmer non-recursively removes the prefixes in the pre-defined set of prefixes, and recursively removes the suffixes in the pre-defined set of suffixes in the following sequence:

1. If the word is at least five-character long, remove the first three characters if they are one of the following:

وال, لال, سال, اال, مال, ولل, كال, فال, بال.

2. If the word is at least four-character long, remove the first two characters if they are one of the following:

وا, ال, فا, كا, ول, وى, وس, سى, لا, وب, وت, وم, لل, با.

3. If the word is at least four-character long and begins with و , remove the initial letter و.

4. If the word is at least four-character long and begins with either ب or ل, remove بor ل only if, after removing the initial character, the resultant word is present in the Arabic document collection.

5. Recursively strips the following two-character suffixes in the order of presentation if the word is at least four characters long before removing a suffix:

ون, ات, ان, ين, تن, تم, كن, كم, هن, يا, نى, وا, ما, نا, هم, ية, ها.

6. Recursively strips the following one-character suffixes in the order of presentation if the character is at least three-character long before removing a suffix:

ت, ى, ه, ة.

4- The System

The system consists of the following blocks:

1- Queries sent to the Internet archive to pick parent and sibling pages.
2- Searching for the document pairs that expected to be translation of each others.
3- Filtering to get the actual parallel documents.
4- Preprocessing and alignment to extract sentence pairs.
5- The algorithm to output the bilingual dictionary.

4.1- The Algorithm

Step1: For n=1, and m=n+1.

Step2: Compare the Arabic sentence number (n) and the English sentence number (n) with the Arabic sentence number (m) and the English sentence number (m) respectively. If there is only one Arabic word common between the two sentences, then extract that word and extract the associated English word or phrase and put them in a table.

Step3: Increment (m), then repeat (step2) until the end of the document pairs.

Step4: Increment (n), and m=n+1, then repeat (step2) and (step3) until the end of the document pairs.

Step5: In (step2) exchange Arabic by English, and then repeat (step1 to step4).

The result of the previous steps produced the following three data types:

a- One English word translated to one Arabic word.

b- One English word translated to an Arabic phrase.

c- One Arabic word translated to an English phrase.

Step6: For the data in (a), for each Arabic word in the table, extract all the English words associated with the same Arabic word, then calculate the repetition percentage of each English word. If this percentage exceeds a specific threshold (Th) value, then copy this Arabic word and the associated English word in a final file.

Step7: Repeat (step6) for data (b and c).

Step8: Remove the final extracted words (in the final file) from the original English and Arabic documents.

Step9: Repeat steps 1 to 8 Several times until no significant words are extracted.

Step10: Use an online dictionary to specify the correct and the incorrect translations in the final file.

4.2- Experimental Results

The precision and the number of translation pairs of the output dictionary resulted from applying the previous algorithm depend on the value (Th) and the repetition of each translation pair. The precision is directly proportional to (Th) and repetition, and the number of translation pairs is inversely proportional to (Th) and repetition. We have applied the algorithm on 700 sentence pairs only for one time and without stemming. Table 1 shows the results for different values of (Th), and any repetition.

Table1: the effect of (Th) on the precision and the number of extracted word pairs.

Th	0.2	0.4	0.6	0.8	0.99
Precision	69.23%	76%	83.82%	87%	88.7%
Number of extracted word pairs	195	167	136	131	126

Also for (Th) = 0.99 and the number of repetition more than 4, we could reach 100% precision for 22 extracted pairs. Applying the algorithm on the whole 15853 sentence pairs (once again without stemming), taking Th= 0.8, for several trials after removing the extracted translation pairs of the previous trail for the next trial; we could achieve the results in table 2. The results in table 2 are of type (a), (One English word translated to one Arabic word)

Table2: the effect of trial number on the precision and the number of extracted word pairs

Trail	First	Second	third	fourth	fifth	sixth	seventh	eighth
Precision	66.09%	71.41%	65.68%	45.8%	40.65%	43.87%	61.53%	34.88%
Number of extracted word pairs	643	2435	1256	524	214	98	65	43

The total system precision is 64.68% for 5278 extracted word pairs.

For the result of data (b), (One English word translated to an Arabic phrase) there were 53 correct translation pairs out of 195, so the precision was 27.17%.

For the result of data (c), (One Arabic word translated to an English phrase) there were 321 correct translation pairs out of 1124, so the precision was 28.55%.

Applying the algorithm on the whole 15853 sentence pairs (but this time after applying the stemming step), taking Th= 0.8, for several trials after removing the extracted translation pairs of the previous trail for the next trial; we could achieve the results in table 3. The results in table 3 are of type (a), (One English word translated to one Arabic word)

Table3: the effect of trial number on the precision and the number of extracted word pairs

Trail	First	Second	third	fourth	fifth	sixth	seventh	eighth
Precision	72.16%	77.5%	70.51%	51.21%	46.19%	50.76%	67.34%	44.44%
Number of extracted word pairs	564	1814	936	412	171	65	49	27

The total system precision is 70.35% for 4038 extracted word pairs.

For the result of data (b), (One English word translated to an Arabic phrase) there were 41 correct translation pairs out of 143, so the precision was 28.67%.

For the result of data (c), (One Arabic word translated to an English phrase) there were 248 correct translation pairs out of 917, so the precision was 27.04%.

After stemming, the system accuracy increased but the total number of extracted translation pairs decreased. The accuracy increased because of the decrease of the system confusion due to the increase of the translation pair frequency after stemming. The total number of extracted translation pairs decreased due to that many Arabic words have been reduced to one word after stemming.

On the other hand the accuracy did not increase too much after stemming because the formation of broken Arabic plurals is complex and often irregular. The following example is an incorrect translation pair that was obtained as a part of the output dictionary: (tool = ادو). The original English word was "tools"; the system before stemming translated "tools" to "ادوات" which is correct. But after changing the plural English word "tools" to singular "tool" and after stemming of the Arabic word "ادوات" to be "ادو", the translation become (tool = ادو) which is not correct since the singular word of "ادوات" is "اداة" not "ادو" because the Arabic plural word "ادوات" is irregular.

5- Conclusion

The objective of this cross language information retrieval (CLIR) research was to examine the effect of Arabic stemmer on the accuracy of the system that is able to extract English-Arabic bilingual dictionary from the Internet archive documents. In this paper we applied a new system on the English and Arabic documents extracted from the Internet archive. The system could achieve 100% precision for high frequency words. The algorithm based on statistical co-occurrence of the English and Arabic words in parallel documents. It is obvious that if we applied the algorithm on a significant parallel corpus, the precision and the number of translation pairs will be much better since the frequency of many words and phrases will be high. Stemming has been increased the system accuracy.

In the future work, we will build a more efficient Arabic stemmer and extract more English / Arabic parallel documents in order to construct a significant English- Arabic parallel corpus and run the algorithm on a large number of sentence pairs to construct a big bilingual dictionary and to improve the precision. Using the constructed bilingual dictionary and parallel corpus we can implement a good English / Arabic machine translation system.

Acknowledgment:

This work has been partly supported by the Education Ministry of Japan under Grant-in-Aid for Scientific Research B (No. 14380166).

References

[1] A. Chen, F. Gey, "Building an Arabic Stemmer for Information Retrieval", the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 19-22, 2002.

[2] J Xu, A Fraser1 and R Weischedel, "TREC 2001 Cross-lingual Retrieval at BBN", The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16 2001, pp 68-77. 2001.

[3] Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam and Hany Hassan, "Language Model Based Arabic Word Segmentation", Proceedings of ACL-2003, Sapporo, Japan, pp399-406, 2003.

[4] Monica Rogati, Scott McCarley and Yiming Yang, "Unsupervised Learning of Arabic Stemming using a Parallel Corpus", Proceedings of ACL-2003, Sapporo, Japan, pp391-398, 2003.

[5] A. Chen, F. Gey, Univ. of California at Berkeley, "Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval", The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16 2001, pp 529-533. 2001.

[6] Fredric C. Gey, Douglas W. Oard, "The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries", The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16 2001, pp 16-26. 2001.

[7] M. Franz, J.S. McCarley, "Arabic Information Retrieval at IBM", the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 19-22, 2002.

[8] K. Darwish, D.W. Oard, "Evidence Combination for Arabic-English Retrieval", the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 19-22, 2002.

[9] K Tanaka& K Umemura, "Construction of a bilingual dictionary intermediated by a third language," In Proceedings of the 15th International Conference on Computational Linguistics, August 1994.

[10] Craig J.A. McEwan1, Iadh Ounis1 and Ian Ruthven, "Building Bilingual Dictionaries from Parallel Web Documents," LNCS, Spring, pp 303- 323, 2002.

[11] J. Xu, R. Weischedel, "TREC-9 Cross-lingual Retrieval at BBN", The Ninth Text REtrieval Conference (TREC 9), Gaithersburg, Maryland, November 13-16, 2000, pp 106-116, 2000.

[12] Ashish Venugopal, Stephan Vogel and Alex Waibel, "Effective Phrase Translation Extraction from Alignment Models", Proceedings of ACL-2003, Sapporo, Japan, pp319-326, 2003.