

Building Bilingual Parallel Corpora based on Wikipedia

Mehdi Mohammadi

Department of Computer Engineering
Engineering Faculty, Sheikh Bahaie University
Isfahan, Iran
mehdi.mka@gmail.com

Naser QasemAghaee

Department of Computer Engineering
Engineering Faculty, Sheikh Bahaie University
Isfahan, Iran
aghaee@eng.ui.ac.ir

Abstract— Aligned parallel corpora are an important resource for a wide range of multilingual researches, specifically, corpus-based machine translation. In this paper we present a Persian-English sentence-aligned parallel corpus by mining Wikipedia. We propose a method of extracting sentence-level alignment by using an extended link-based bilingual lexicon method. Experimental results show that our method increase precision, while it reduce the total number of generated candidate pairs.

Keywords- Parallel corpora, Sentence alignment, Wikipedia

I. INTRODUCTION

Parallel corpora refer to bodies of text in parallel translation [1]. They are the main prerequisite of much multilingual research activities, such as developing multilingual lexicons and corpus based machine translation systems. But finding parallel text corpora is a difficult task except for a few dominant languages. *Comparable* corpora as a source of translation knowledge have attracted the attention of many researchers. Comparable corpora are composed of document pairs describing the same topic in different languages. They are not *parallel* (mostly sentence-to-sentence translated) corpora composed of good bilingual sentence pairs, but still contain various levels of parallelism, such as words, phrases, clauses, sentences, and discourses, depending on the corpora characteristics.

Since high-quality parallel corpora with Farsi and English as language pair do not exist or are not accessible for the research community because of copyright restrictions, the compilation of aligned parallel corpora for Farsi and English languages is one of the priorities of the Example based machine translation system.

In recent years, cooperative edited multilingual project, Wikipedia, is increasing multilingual researcher's interest to use it as a prime resource. Wikipedia is a multilingual encyclopedia that is online and freely available and it is developed for most of the world's languages. These encyclopedias are editable by anyone and freely-distributable. These attributes of Wikipedia encyclopedias with their content and structure make them an appropriate and useful resource for building multilingual parallel corpora in wide range of the world's languages. The text in Wikipedia is formatted using a special but simple markup, so it is easy to add and maintain Wikipedia contents. Each encyclopedia in a separate language has its own articles.

Many of these articles are shared between the different encyclopedias and conduct a same concept, but are not simply translations of one another. In other words, In Wikipedia, some pages may be translations of each other whereas the majority of the pages probably written independently of each other. Even if one page is not a translation of another, they certainly share some common information, so Wikipedia is a comparable corpus not a parallel corpus. The pages of different types of topics such as a biography of a person, description of a country or city and the definition and description of a concept etc., needs to include same subject matter in any language. This sameness is identifiable in the hypertext links and their anchor texts in Wikipedia [2].

The aim of this paper is to build a Farsi-English parallel corpus that aligned at sentence level. Existence of a parallel corpus without any noise is the precondition of many approaches to find similar sentences. However, there are some methods for finding similar sentences across multiple languages in non-parallel but comparable corpora that can be used in case of Wikipedia. In Wikipedia, a topic described in two different languages as two versions of a same entity can be considered for searching similar sentences [2].

Briefly, we download each Persian page with its English version if it is available. Then a bilingual dictionary is generated from titles of each page pair. In addition to each word pair, a unique Id of each pair is maintained. Each word or phrase in the sentence pairs that have occurred in the bilingual dictionary are represented by their Id. Then we can calculate similarity between sentences by some similarity measures used widely in Information Retrieval. By applying three similarity measures, Dice, Cosine and Jaccard coefficients to a ten closely translation pairs, we found that the Jaccard similarity measure is best suited for our purpose rather than Dice and Cosine Coefficients. Gao also tested it on his parallel texts. We show if our proposed method is preferable rather the same work.

The remainder of the paper is formed as follows. The next section describes some related works. In section 3, some detailed information of Wikipedia that are useful in this research is explained. We then explain our proposed method in detail. In section 5, experimental results and benchmarks are presented. The paper is finally concluded in section 6.

II. RELATED WORKS

The presented work in this paper is in the same concept as that by Adafre and Rijke [2]. They explain their

experiment in gathering similar sentences between two different language versions of Wikipedia. They proposed two approaches for identifying similar sentences. The first approach uses an online machine translation system to translate Dutch Wikipedia pages into English. In this method, similarity between sentences in original English and translated English pages is calculated based on number of shared words. Their second method uses the hyperlinks between documents and a bilingual dictionary created from Wikipedia titles. Although, comparing two methods is their research aim, by the first approach, “the chicken or the egg” problem springs to mind. In addition, as they noted, there is not translation system for every language pair in Wikipedia. The second method also generates a great number of unrelated sentences as translation pairs. To create parallel corpora, we apply their second method on Farsi and English Wikipedia by considering two restrictions: the correspondence between lengths of two chunks of text in different languages and the minimum of similarity measure that two sentence should meet. In this paper, we refer to Adafre’s Approach as *base* method.

Another method for building sentence-aligned corpus from Wikipedia is reported by Yasuda and Sumita [3]. They propose a Machine Translation bootstrapping framework that can simultaneously generate a statistical machine translation (SMT) and a sentence-aligned corpus. They align sentences using a bilingual lexicon and the lexicon automatically updated using the aligned sentences. Their method requires an initial sentence-aligned corpus. They showed that 10% of Japanese sentences have equivalent sentence in English.

Tyers and Pienaar [4] described a method for extracting bilingual dictionary entries from Wikipedia using the link structure between different languages (interwiki links) of Wikipedia. They found precision to be in the range 69–92% for four language pairs, so the quality of lexicons extracted from Wikipedia titles are generally reliable and as noted Adafre, conceptual mismatch between pages linked by interwiki links is rare [2].

Also, a number of techniques for aligning sentences in parallel corpora have been proposed. Gale and Church [5], Brown et al. [6] and Wu [7] used sentence length as the main feature to align sentences. In the case of the web, there are some researches to retrieve parallel texts. BITS [8], is a system to find and collect parallel text automatically. In fact it aligns texts in document level. STRAND [1] is another attempt to discover parallel pages across the web.

III. WIKIPEDIA

Wikipedia is a free multi-lingual online encyclopedia. It is collaboratively written by its users. Every one may add, edit, or remove content using a simple markup language. This paper uses the Persian and English versions of Wikipedia. The Persian Wikipedia contained over 80,000 articles as of October 2009 and the English version has over 3 million articles. Each article in Wikipedia is uniquely identified by its title – a sequence of words separated by underscores, with the first word always capitalized. Wikipedia articles include hyperlinks to direct users to additional information. Hyperlinks can be internal, pointing

to other Wikipedia articles, or they can be external, pointing to outside resources. Furthermore, there is several links to other languages on every article that users can access through them to other version of a subject in different languages. In other words, different versions of a page in different languages are also hyperlinked. For example, on the English article for *physics*, there is a link to the same article, *فیزیک* (fizik), on the Persian Wikipedia. These links correspond to same article titles in different languages and are called *interwiki* links. So, for some pages, equivalent of its subject in other languages are provided as hyperlinks. Using this property we can align the corpus at the document or article level. It is also useful to generate a bilingual dictionary consisting of the Wikipedia titles.

For every Persian Wikipedia article, if there is a link to its English version, we download both Persian and English articles; otherwise, this Persian article is discarded. So we obtained a collection of documents that are aligned at document level. This alignment is integrated into a specific crawler for Persian Wikipedia.

IV. GATHERING SENTENCE TRANSLATION PAIRS

Our approach to align sentences uses a bilingual dictionary generated from Wikipedia titles using the interwiki links. It is same in concept as Adafre's method [2]. We improved the Adafre's approach by observing Gale's research findings [5]. Gale's program makes use of a simple statistical model of character lengths to align sentences in parallel corpora. The model is based on the observation that longer sentences in one language tend to be translated into longer sentences in the other language while shorter ones tend to be translated into shorter ones. Gale used a normal distribution for character occurrences in two languages and specified the model by the mean and variance parameter of this distribution. The mean is defined as the expected number of characters in language L2 per number of characters in language L1. They empirically obtained this parameter around 1 for two language pairs. There for this parameter is language independent. So, we can induce that each Persian sentence length should be correlated with its English translation according to mean parameter of a normal distribution. To determine this parameter for Persian-English pair, we select 30 Wikipedia page pairs randomly. Then we passed them to Adafre’s approach to find similar sentences across them. About 750 sentence pairs generated; we evaluated the generated pairs by specifying sentence pairs that contain correct translations. More than 80 sentence pair identified as translation pairs. Evaluation showed that Persian sentences and their English equivalent are almost correlated by their character lengths. This correlation is often in the range 1 ± 0.5 shown in figure 1. In this practice we did not remove stop words. We supposed only 1-1 correspondences in text units: a sentence in one language normally matches exactly one sentence in the other language. The Gale's approach is applied on Hansards corpora that is in parallel not comparable. Obviously, the Gale's approach is not individually suited to apply on comparable corpora such as Wikipedia, but it is useful with a method specialized for comparable corpora. As noted, all candidate pairs do not

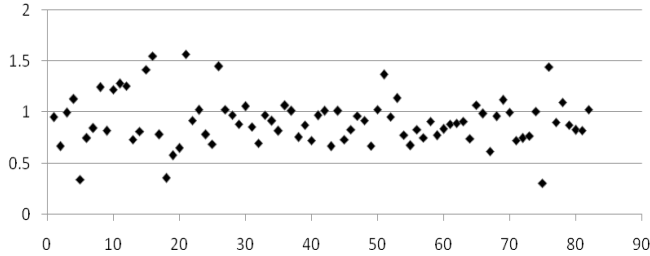


Figure 1- Distribution of character correlation for correct translation sentence pairs

contain equivalent translations. By analyzing the similarity of sentences in candidate pairs, it is revealed that for correct translations, the similarity is greater than a minimum bound. Figure 2 shows the similarity of all correct translation pairs in which minimum similarity is 0.016; we choose 0.02 as a threshold. Therefore we could limit the generated candidate pairs using these simple criteria. In other words, if calculated similarity for two sentences is lesser than the threshold, that pair should be ignored to add in candidate pairs.

We developed a specific crawler that starts crawling through Persian Wikipedia pages starting from home page of Persian Wikipedia¹ as seed url. Downloading action on every Persian page is done if a link to English version of that title exists. The English version is also downloaded and the document level alignment is formed. If there is no link to English version on a Persian page, that page is discarded. Recursively, the links in the downloaded Persian page are added in a queue to proceed later. By this procedure, we collect a number of document pairs.

In the first step of sentence alignment, we acquire the bilingual lexicon. For each Persian-English Wikipedia page pair, we take the title of the page in *title tag* of their html content. The titles may be a word or more than one word. A unique Id from English title is created by concatenating title words separated by an underscore. Persian-English titles are added to the lexicon with their unique Id.

In the next step, we split each Persian and English page to sentences, and then we associate each Persian sentence to each English sentence that their character length correlation is in the range 0.5 through 1.5. Unlike the Adafre's work, this does not lead to produce as many as Cartesian product of two collections of Persian and English sentences as candidate pairs. Then for each candidate pair, representing the sentences with generated bilingual dictionary is done. Representing the sentences is as follows: in each sentence, hyperlinks are substituted by their unique Id if it has entry in generated bilingual dictionary. There may be cases that word or phrases that have entries in Wikipedia, had been represented without hyperlink texts. So, we consider such cases to be held in similarity measurement. For each sentence, various N-Grams of that sentence for $N \leq 4$ are generated regarding the words order in the sentence. For each phrase generated by above procedure, if there is an entry for it in bilingual dictionary, we replace it with its Id.

¹ <http://www.fa.wikipedia.org>

Then we calculate similarity measure for two sentences of candidate pairs. We examined three similarity measures including Dice, Cosine and Jaccard coefficients in our application on a sample set of page pairs; but Jaccard similarity measure was best suited for our purpose rather than Dice and Cosine Coefficients. Gao also tested it on his parallel texts. In fact we calculate the number of dictionary Id's shared between sentences. In this step, if the calculated similarity is greater than the threshold, we add two sentences to candidate list. Then the list of candidate pairs are sorted descending based on their similarity. So the candidate pairs that have high similarity placed in top of the list and vice versa. Because all of the candidates are not translation pairs, the next step is filtering the list to remove non translation pairs. Filtering is as follows: the first top score pair in the list is taken; in the remainder of the list, all sentence pairs that contain one sentence of the first pair (English or Farsi) in which their score is lower than the first pair, are removed. This procedure is iterated for the second top score pair in the list and continued till there is no pair to remove. The remained list contains most translation sentence pairs.

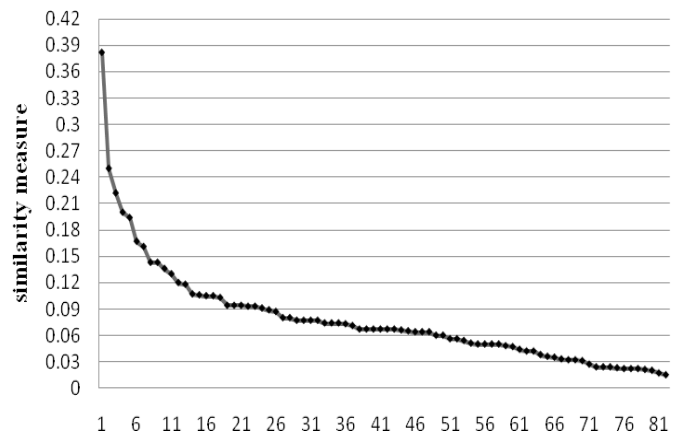


Figure 2- Similarity of correct translation pairs that are greater than a minimum bound

V. BENCHMARK

As noted in previous section, we apply the base method on 30 Wikipedia page pairs. By analyzing the correct results, we found some tips to improve the Adafre's method. So we developed our approach based on these findings. Next, We applied our extended method on identical 30 pairs. We added sentences length correlation and similarity threshold filtering to base method. The generated candidates are evaluated to find correct translations. Correct translation pairs are selected between all of candidate pairs if two sentences are exact translation of each other or one sentence contains in another sentence. The task of identification of correct translations is done manually. Table 1 shows the results of two approaches. In the table, Base method column refers to Adafre's link-based bilingual dictionary method and proposed method column show the results of our method.

Title		Base method		Proposed method	
English	Persian	Total Candidates	Correct Translations	Total Candidates	Correct Translations
Amedeo Modigliani	آماندو موديليانى	123	54	69	48
Germany	آلمان	108	1	64	3
United States	ايلالات متحده آمريكا	83	2	27	2
Avicenna	ابو علي سينا	42	0	9	0
Ethylene	اتيلن	22	3	15	4
Baron	بارون	4	0	0	0
Pope	پاپ	135	7	47	3
Bronze Age	عصر برنز	25	1	16	1
Submarine	زيردريايي	8	0	3	0
Big Lie	دروغ بزرگ	5	0	0	0
February 25	۲۵ فوریه	0	0	0	0
Nuclear astrophysics	اخترفيزيك هسته‌اي	1	0	0	0
North Pole	قطب شمال	9	1	1	0
Kepler's laws of planetary motion	قوانين كپلر	6	0	3	0
Volcano	آتشفشان	3	0	1	0
Equinox	اعتدال پاييزي	3	0	1	0
Economy	اقتصاد	16	0	9	0
Organism	اندامگان	1	0	0	0
Encyclopedia Britannica	دانشنامه بریتانیکا	21	0	4	0
Civilization	تمدن	23	1	13	1
Horoscope	زايچه	1	0	0	0
Biochemistry	بيوشيمي	4	0	3	0
United Nations	سازمان ملل متحد	59	9	18	4
Culture	فرهنگ	5	0	2	0
Management	مدیریت	12	0	7	0
Close-up	نماي نزديك	1	1	1	1
Ghetto	گتو	5	1	4	1
Database	پايگاه داده	14	1	5	1
Urban Planning	برنامه ريزي شهري	1	0	0	0
Qaumi Tarana	سرود ملي پاکستان	0	0	0	0
Sum		740	82	322	69
Average		15.6	2.7	10.7	2.3

Table 11- Applying two methods on 30 same topics (titles of English and Persian articles)

Both the base and proposed method have two columns as *total candidates* and *correct translations*. *Total candidates* show the number of total candidate pairs for a topic. *Correct translations* show the number of pairs that identified as correct translation.

By overviewing the page "*Amedeo Modigliani*" and its Persian counterpart that have more correct translation pairs revealed that they are parallel. The results also show some points of Adafre's notes like that the type of pages that have more correct translation pairs contains those that are about persons. We define the precision as the number of correct translation pairs over total number of candidates. In the proposed method, the precision is 21%. It is computed for the base method about 10%. In proposed method, the number of total generated candidate pairs reduced to approximately 43% of base method; whilst the precision is 2 times bigger than the base method. As a defect, by the

proposed method, some of correct translations in the base method, about 15% of them, are lost. Obviously it is because of sentences length limits and similarity threshold. Although removing such candidate pairs, unified the created corpus in term of sentence length and similarity. In the other hand, Wikipedia is a huge body of texts. Though we think this deflection can be ignored by covering more and more documents in a fast and more precise manner like the proposed method.

VI. CONCLUSION

In this paper we used Wikipedia as comparable corpora to generate a bilingual parallel corpus. We used links to similar entities in sentences to align sentences in bilingual corpus. We improved the base method by restricting it into considering length of sentences and defining a threshold for

similarity between two sentences. Results show that the proposed method is twice more precise than the base method. Also the number of total candidate pairs is concentrated to less than half of the base method. By the proposed approach, a bit of translation pairs are ignored; some of them because of low similarity measure between two sentences and others because of low degree of length correlation between two sentences.

We ran the proposed method on 1600 page pairs. It yielded about 12530 sentence pairs. The result obtained from this running and sample of Wikipedia page pairs, revealed that the similarity measure computed for each translation pairs is quite a small number. This causes to increase the noise in the corpus. The main reason refers to the fact that in Persian, some alphabet letters have various spellings like ‘ي’ and ‘ى’ (are pronounced as /i/). One of our feature works is to support various spellings of a word in representing sentences.

REFERENCES

- [1] Philip Resnik and Noah A. Smith, "The Web as a Parallel Corpus," *Computational Linguistics*, vol. 29, no. 3, 2003.
- [2] Sisay Fissaha Adafre and Maarten de Rijke, "Finding Similar Sentences across Multiple Languages in Wikipedia," , 2006.
- [3] Keiji Yasuda and Eiichiro Sumita, "Method for Building Sentence-Aligned Corpus from Wikipedia," AAAI'08, 2008.
- [4] Francis M. Tyers ,Jacques A. Pienaar, "Extracting bilingual word pairs from Wikipedia," in *LREC2008*, 2008.
- [5] W. A Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *In Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, California, 1991.
- [6] P. Brown, J. Lai and R. Mercer, "Aligning Sentences in Parallel Corpora," in *ACL Conference*, Berkeley, California, 1991.
- [7] D. Wu, "Aligning a parallel English-Chinese corpus statistically with lexical criterias," in *In proceedings of 32nd ACL conference*, 1994.
- [8] Xiaoyi Ma and Mark Y. Liberman, "Bits:A method for bilingual text search over theWeb," in *MachineTranslation Summit VII*, 1999.