International Journal of Intelligent Information Systems
2016; 5(3): 42-47
http://www.sciencepublishinggroup.com/j/ijiis
doi: 10.11648/j.ijiis.20160503.12
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Presenting an Optimal Method for Constructing an English-Persian Comparable Corpus

Seyede Roya Mohammadi, Noushin Riahi

Computer Engineering Department, Alzahra University, Tehran, Iran

Email address:
roya_mohammadi1@yahoo.com (S. R. Mohammadi), nriahi@alzahra.ac.ir (N. Riahi)

To cite this article:
Seyede Roya Mohammadi, Noushin Riahi. Presenting an Optimal Method for Constructing an English-Persian Comparable Corpus. *International Journal of Intelligent Information Systems*. Vol. 5, No. 3, 2016, pp. 42-47. doi: 10.11648/j.ijiis.20160503.12

Received: March 23, 2016; **Accepted**: June 7, 2016; **Published**: June 18, 2016

Abstract: Multilingual corpora are the main sources in language information retrieval fields. The quality of many researches such as machine translation strongly depends on the quality of these corpora. One of these corpora's is comparable corpus. Considering their quality, these corpora contain broad range of information but constructing them has its special problems which lead to a few numbers of pairs in comparable corpus unlike its large dataset. In this paper we present a new method for increasing the quality and quantity of comparable corpus. We built a Persian-English comparable corpus from two independent news collections: BBC news in English and Hamshahri news in Persian.

Keywords: Comparable Corpus, Corpus Quality, Hamshahri Corpus, Query, RATF Factor

1. Introduction

Cross Language Information Retrieval is a growing research field and is not limited to a particular topic. Some of its applications are information retrieval, natural language processing, speech processing and machine translation. Most of these applications require huge amounts of data that is named corpora. Such corpora are used to extract the desired information for languages and making efficient resources for these languages.

There are two kind of corpus: parallel and comparative. Parallel corpus is a collection of texts containing origin language sentences and their translation in the target language. Parallel texts are important sources of training data for statistical machine translation. They are used for other applications such as questions answering systems, language information retrieval and tag extraction, too.

However, parallel corpora have small volume and are strictly limited to the language and scope. In addition, there is a few number of Persian-English parallel corpora. Although some attempts were done to extract the parallel corpus in Farsi language using the Internet, but there is still the small volume and limited scope problem.

Comparable corpus is one of the most challenging fields in machine translation and natural language processing applications. Comparable corpus consists of texts in more than one language, which are in the same field of subject and same period of time which speak about similar information.

Comparing with parallel corpus, comparable corpus has a lot of benefits from which we can name its accessibility, availability and variety in subjects. On the other hand, parallel corpora are based on this assumption that varying in size of texts is not very much. We can provide texts needed for comparable corpus easily by using news feeds, books, magazines, advertisements and etc. so in the recent years studies on using comparable corpora for NLP application in the most of the languages has been increased significantly.

Research has shown that the volume of training data has a significant impact on the performance of machine translation systems and comparative corpus can compensate the small volume of parallel corpus. In fact, the addition of parallel phrases and sentences from a non-parallel corpus to the training data can increase the efficiency of machine translation systems. Moreover, the researchers have demonstrated that the effect of comparative corpus on the performance of machine translation systems is same as the human translation data with a similar size and scope [1].

According to this, the aim of this research is to create Persian-English comparative corpus with the use of information from the press.

The rest of the paper is organized as follows. In the second part, the related works are presented. The third part is devoted to an explanation on how to make a comparable corpus. In part 4, we explain the algorithm presented for increasing quality of comparable corpus and in part 5 we analyze the achieved results from experimental results.

2. Related Works

Unlike parallel corpora that are very limited, comparable or semi parallel corpora are found in many subjects and language pairs. Fung and Rapp in years 1998 and 1999 proved that to find the sentence alignments, we don't need the texts to be exactly parallel and instead of parallel text it is possible to use texts which have common phrases and concepts. After those results, researchers tried to construct bilingual corpora [2-9].

Fung and Cheung [10] aligned any source language documents to some target language documents and examined all possible sentence pairs, but there is no fixed list of documents pairs. Hence after a round of statements extraction, the list might be completed with more documents and the operation should be repeated.

In the reference [11], comparative corpus is created base on this assumption: the words in the corpus that are translated or that are on shared issues, usually have similar occurence time periods.

Talvensaari and et al. in [12] offered a method based on the cross language information retrieval documents to align the two different languages. In their approach, making comparable corpus, includes two steps, acquisition and alignment. In the first step, texts of two languages (especially the news published in the two languages) are collected using web crawling. In the second step for the alignment of two documents, first for each document of the source language, the key words (which are indicative of the general content of the document), and then words are translated to the target language. So a query is made to the target language that represents the source document and the target language documents are extracted. Finally, the alignment between the source document and the documents retrieved from the target language is done based on the similarity of them and their time occurrence.

Talvensaari and et al. in their next research [13] offered an automatic method to create a comparable corpus in a specific domain. In that study, the authors made their experiments by crowling on the field of genomics. For this purpose, first a number of specific words of that field would be given to search engine and some documents are extracted. Then the documents paragraphs would be extracted and similarity between the query words and paragraphs would be computed. If each similarity exceeded of a threshold, then that paragraph would be selected and aligned to the source language paragraph and they were transfer to the corpus. In this research a particular area was considered. They didn't use news to create the corpus, so the occurrences time wasn't important. The alignment was in the paragraph level not in the document level, because two documents may be aligned only in a section of the documents.

Constructing comparable corpora is commonly done in 3 steps. The first step is constructing text victors. Shao and NJ in 2004 used language models for this purpose [14]. In this step Utsuro in 2007 presented a new binary similarity criterion that had been extracted from a pre training parallel corpus [15]. Xu, et al. [7] used dependency parsers. Also there were researchers like Fung [10] and Shezaf [16] who worked on using different relation criteria and weighting formula.

The second step is translating victors. In this step Kuhen and Night in 2002 produced an automatic dictionary based on unique words. Saralegi and et in 2008 used cognates in this step [17].

The purpose of [18] is to improve the quality of comparative corpus for extracting the translating knowledge. They defined a criterion for the corpus adaptation and tried to improve the corpus quality by an iterative process.

In Persian language we tried to make a large bilingual comparable corpus. In this corpus we used weighted extracted keywords from text, name entities and main words in title, differently. At last we used the sum of the weights as a criterion for calculating the similarity of source and target language. In addition, the use of weighting method could be increased the quality of corpus, too.

3. Constructing the Comparable Corpus

The process steps for constructing a comparable corpus are: primitive text selection, text preprocessing, document translation and the similarity value calculation between each two documents. The steps will be explained in this section.

Primitive text selection

The first step in the automatic creation of a bilingual comparable corpus is the text extraction via Internet. We used Hamshahri news from 2002 to 2006 for Persian text and BBC news in the same time period for English text. We deleted the local news from them and then converted them to a standard text format.

Text preprocessing

In parallel corpus, the source and translated texts usually placed by an author on the site, but in comparative corpus, generally texts from different sources are collected, so they can have much punctuation differences with each other. It is better in the first step to remove the punctuation differences between the texts of two languages as much as possible. It is done in text pre-processing.

One of the pre-processing operation is remove stop words. In English Inquery stop word list and in Persian the list of Noshatel university was used for this purpose. Next operation is homogenization of specific symptoms such as $ and %. The time and history in the texts should be converted to the standard format, too. Beside these, punctuation in texts other than the quotes were removed.

Finally, we removed the acronyms in English and replace them with full expression. For this purpose, a list of English common words were extracted and then replaced.

Document translation

We used a bilingual English-Persian dictionary with more than 100,000 distinct word in the C # language for the initial translation of English texts, and attached it to our program. Words in English texts were extracted and translated using the dictionary.

There were a lot of untranslated words after the initial translation, the reason is the inflected and derived words. To solve the problem we used Porter, a stemmer, to stem the words.

After the previous translation step, there were still many untranslated words. These words were translated by Google translation systems. This system is very comprehensive and can almost translate all of the words. But translation via internet is slow. For this reason, we tried to translate the most of words by the internal dictionary.

Names could not be translated, in the process of translating. Named Entities (people, places and organisations) should be converted to phonetic translation by Google Translate. We used name entities weighting to enhance the efficiency. Name entities were recognized by LingPipe and converted to Persian by Google transliteration system.

Similarity values computing

The final similarity between two documents was obtained by a linear combination of name entities similarities, document title words similarities and keyword similarities.

To compute the documents similarities, first the documents passed from a time filter. Our documents are news, and the possibility that an agency produces news after a few days is low. We compared only the news that occur in four days interval, with each other.

Then the documents keywords were selected. The Relative Average Term Frequency (RATF) value was computed for each document words and the keyword were selected based on the RATF value.

$$RATF(k) = (cf_k/df_k) \times 10^3/\ln(df_k + SP)^p \quad (1)$$

Where cf_k is the collection frequency means of the number of appearance of words set k in the total documents, and df_k is document frequency means of the documents number containing words set k. SP and p are two constant for reducing the weights of rare words. We set SP=1800 and p=3 based on [20].

In order to build a representative of source documents query, first each document words were sorted in descending order, based on the appearance frequency. Then keywords with the same frequency were sorted in descending order, one time based on their appearance in document and another time based on their RATF values. Now, m keywords with the highest ranking is selected for each document. We set m=30 base on [20].

After the keyword selection, they should be translated to the target language using dictionary, google translate and transliteration, as before said. After making a target language query, we used a retrieval model to rank the target documents based on similarity to the query. These keywords similarity values were used in the final step.

Then name entities were identified and transliterated to target language and the name entity similarity value was computed between each two documents in the source and the target langauges.

The document title words similarities are important, too. These similarities are more important than the document words similarities. Because document title is usually chosen so that to represent the general purpose of the document, in fact it can be considered as news summary. For this reason, some researchers have only used document title words similarities and time occurrence of news for the documents aligning.

To compute the similarities, titles were separated from the news before the Keywords selection step. Then stop words were removed from the titles and document title words similarities were computed by an Information retrieval system.

Finally, we calculated the similarity between each two documents by a linear combination of keywords similarity, name entities similarity and document title words similarity.

$$SC = \left(w_{pn} \times N_{pn} + w_t \times N_t + w_{kw} \times N_{kw}\right) \times NF \quad (2)$$

w_{pn} and N_{pn} are the weight and similarity value of name entities, w_t and N_t are the weight and similarity value of titles, and w_{kw} and N_{kw} are the weight and similarity value of keywords. NF is a factor to normalize the obtained values in the range of 0 to 100. Some experiences showed that w_{pn}=5, w_t =3 and w_{kw}=1 are almost optimized.

4. Improving the Quality of Comparable Corpus

In the presented method for constructing the comparable corpus, one of the factors used in calculating the final weights, was the words which were selected based on document frequency and RATF factor. As we said, the words extracted from documents in that section were used as candidate words in queries. After selecting the words, translation system tried to translate them to target language. At first, system used a dictionary for translating. If the dictionary system couldn't translate the words, we used online translation with Google translator. If the word couldn't be translated in this section again, we had to use Google transliteration system for transliterating word from the source language to the target language.

The problem we faced in this section was that using Google translation and transliteration system needed connecting to internet and unlike the dictionary system it needed a long time for translating given words. So we had to limit the words we selected as much as it was possible and for this reason we had two problems. First the number of words shouldn't go beyond a special number and the second one was that we had to select adequate number of words to be a good query for the document. So we used document frequency and RATF factor and we considered a constant number and tried to optimize this constant with experiments. But selecting a constant number for all the documents can’t be the best way as it was obvious that the length of documents in our data set is varied a lot, in special boundary. We saw in our sample dataset that we had a document with just 6 valuable words and beside it we saw a document with more than 1300 valuable words. So selecting an equal number of words from both of these documents was not a good idea. So we used an algorithm for selecting the number of words from each documents base on the number of its unrepeated words. The main equation for selecting words is equation 1:

$$\sum_{i=1}^{num\ of\ Docs} n_i \leq \frac{T}{t} \tag{3}$$

In which n_i is the number of words selected for document number i, t is the time consumed for translating one word and T is a constant which is selected in a way that T is an agreeable value for time period. We can rewrite equation 1 in another way. As we described at first, we tried to consider a constant value k for each document and we chose k so that the time consumed for translation was logical. So we can adapt this idea with equation 1 and rewrite it as equation 2:

$$\sum_{i=1}^{N} n_i \leq kN \tag{4}$$

In this equation k is constant amount and N is the number of documents. Now we have to find the amount of n_i s. For finding n_i s we first tried to calculate mode of the numbers gained from number of words in each document. After calculating the mode amount, we deleted the numbers which were too different with mode, because the numbers which have great different with other numbers can decrease the performance of algorithm significantly. After eliminating amounts out of mode range, we have to calculate the number of groups using equation 5:

$$G = \frac{n_{max} - n_{min}}{\beta} \tag{5}$$

In this equation nmax is the number of words in document with the maximum number of words and n_{min} is the number of words in document with minimum number of words and β is calculated by experiment. After specifying the number of groups we should calculate the number of documents in each group. In the next step we calculate the primary amount of n_i s by using equation 6:

$$n_i = \frac{cnt_i}{N} L_i \tag{6}$$

In this equation cnt_i is the number of words in the document i and L_i is the number of documents in a group containing the document i. After calculating these numbers, sum of the words selected primarily may be more than the sum we considered at first. So we should calculate the difference between these two amounts.

$$diff = \frac{T}{t} - \sum_{i=1}^{N} n_i \tag{7}$$

In the next step we calculate the primary amount of elimination by using equation 8:

$$nod = \frac{diff}{f_1 + f_2} \tag{8}$$

In this equation f1 and f2 are amount of the first and the second maximum group according number of documents in range. If calculated nod is less than λ then we can minus calculated nod from n_i s assumed for the documents in groups f1 and f2, but if nod is greater than λ then we should spread minus to all of the documents in dataset. In this situation the final minus amounts are calculated using equation 9:

$$\begin{cases} m_i = \frac{diff - \lambda N_{f1+f2}}{N}\ if\ d_i \in \{f_1, f_2\} \\ m_i = \frac{diff - \lambda N_{f1+f2}}{N} + \lambda\ if\ d_i \notin \{f_1, f_2\} \end{cases} \tag{9}$$

In this equation N_{f1+f2} is the total number of documents in group with the first and the second Rank according to the number of documents in that range.

At last, we should minus m_is from calculated n_i s.

5. Experimental Results

The experiments are done on the results of Persian English documents. For Persian documents we used Hamshahri corpus. This corpus contains 191440 documents (news) of years 2002 to 2006. As English dataset, we crawled BBC news from its site. It contains the same period of time news and it consists of 53697 documents. Our main criterion for analyzing precision of comparable corpus is 5 level criterions which is defined as below:

1. Equal subjects: two documents totally speak about one subject.
2. Related subject: two documents are about one subject but from different point of views.
3. Similar appearance: two documents are about related events.
4. Similar terms: similarity between events is less but there are a lot of common words between them.
5. Unrelated: there is no obvious similarity between two documents.

For analyzing the results by hand we used a month period of these documents and put achieved document pairs by algorithm in its related classes. The results are compared in a month period of January 2002.

Table 1. *Effect of using the title similarity in addition to keywords similarity.*

	Keywords		**Keywords and titles**	
	Alignment number	**Alignment number**	**Alignment number**	**Alignment percent**
Class 1	13	13	15	15.46
Class 2	52	52	55	56.7
Class 3	20	20	22	22.68
Class 4	4	4	5	5.15
Class 5	0	0	0	0
Sum	89	89	97	100

In the first experiments, we wanted to see the effect of using the title similarity in addition to keywords similarity. Table 2 show addition of title similarity causes to increase the number of alignments in classes 1 to 4, as well as corpus quantity.

In experiments done for constructing English-Persian comparable corpus, the constant amount we considered was 30 words per each document. This amount was selected to compare the results with the results of another method for constructing corpus. So for calculating the constant amount needed here, we used equation 4 instead of equation 3. After calculating the constant number, we assumed number of β equal to 100.

Considering the results, total translated words were equal to previous method but we had good increments in performance due to qualified balancing of number of words. Achieved results from this experiment can be seen in table 2. In this table we compared the results with the result of constant number method.

6. Conclusions

In this article, we discussed how to constract an English-Persian comparative corpus. We used RATF value to select the source language texts keywords. After pre-processing and translating, the target language texts keywords were determined. Then the name entities were extracted. The similarity value would be equal to the linear combination of name entity similarity, title similarity and keywords similarity.

Table 2. *Comparing results of experiments with constant number selection via dynamic number selection in a month period.*

	Dynamic number method		Constant number method	
	Pairs Num	Pairs percent	Pairs Num	Pairs percent
Class 1	21	18.42	19	18.09
Class 2	64	56.14	60	57.14
Class 3	23	20.17	22	20.95
Class 4	6	5.26	4	3.80
Class 5	0	0	0	0
Sum	114	100	105	100

The results show that using this method compared to using only similarity of keywords, cause to increase the number of first, second and third classes of alignment. This means the quality and quantity of the comparable corpus are increased.

In addition, we presented a method to choose each document keywords number based on the document length. The method caused to increase the quality and quantity of the comparable corpus, too.

References

[1] A. Blets, E. kow, “Extracting Parallel Fragments from Comparable Corpora for Date-to-Text Generation”, Proceeding INLG’10 Procedeeing of the 6^{th} International Natural Language Generation Conference, 2007, pp. 167-171.

[2] P. Fung, “Finding terminology translations from nonparallel corpora”, Proceedings of the Fifth Workshop on Very Large Corpora, pages 192–202, 1997.

[3] R. Rapp, Automatic identification of word translations fromunrelated english and german corpora. In Proceedings of the 37th annual meeting of the association for Computational Linguistics on Computational Linguistics, pages 519–526, Morristown.

[4] D. Herv´e, E. Gaussier, and F. Sadat, An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In Proceedings of the 19th International Conference on Computational Linguistics, COLING, pages 1–7, Taipei, Taiwan.

[5] R. Xavier, Y. Sasaki, M. Tonoike, S. Sato, and T. Utsuro, Compiling French-Japanese terminologies from the web. In proceedings of the 11st EACL, 2006, pages 225–232, Trento, Italy.

[6] E. Morin, D. B´eatrice, T. Koichi and K. Kyo, Bilingual terminology mining - using brain, not brawn comparable corpora. In Proceedings of the 45th ACL, 2007, pages 664–671, Prague, Czech Republic.

[7] J. Xu, W. Croft, “Query expansion using local and global document analysis”, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 18–22 August 1996, pages 4–11.

[8] R. Xiao and X. Hu, Corpus-Based Studies of Translational Chinese in English-Chinese Translation, Springer Heidelberg New York Dordrecht London, 2015, ISSN 2197-8689, ISSN 2197-8697 (electronic), New Frontiers in Translation Studies, ISBN 978-3-642-41362-9, ISBN 978-3-642-41363-6 (eBook), DOI 10.1007/978-3-642-41363-6.

[9] K. Benjamin Tsou, Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared, Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 1–2, Beijing, China, July 30, 2015.

[10] P. Fung and P. Cheung, “Mining very Non-parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM”, In EMNLP 2004, pages 57-63.

[11] T. Tao, C. X. Zhai, “Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration,” in SIGKDD, 2005, pp. 691-696.

[12] T. Talvensaari, J. Laurikkala, K. Jarvelin, M. Juhola, H. Keskustalo, “Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval”, ACM Trans. Inf. Syst., Vol. 25, No. 1, 2007, pp. 4.

[13] T. Talvensaari, “Effects of Aligned Corpus Quality and Size in Corpus-Based CLIR,” Advances in Information Retrieval, 2008, pp. 114-125.

[14] L. Shao and H. T. Ng, “Mining New Word Translations from Comparable Corpora”, In: COLING 2004.

[15] M. Tonoike, T. Utsuro, and S. Sato, “Compositional Translation Estimation of Technical Terms using a Domain/Topic-Specific Corpus collected from the Web”, Journal of Natural Language Processing, Vol. 14, No. 2, pp. 33-68, April 2007.

[16] D. Shezaf and A. Rappoport,. Bilingual Lexicon Generation Using Non-Aligned Signatures. In Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, 2010, pp. 98–07.

[17] X. Saralegi, I. San Vicente and A. Gurrutxaga, =Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In Proc. of the 1st Workshop on Building and Using Comparable Corpora (BUCC) at LREC 2008.

[18] B. Li, E. Gaussier, “Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora,” in Proceeding of the 23rd International Conference on Computational Linguistics, Beijing, China: Coling Organizing Committee, 2010, pp. 644-652.

[19] NJ, USA. Association for Computational Linguistics. Ghayoomi, Momtazi, Bijankhan, A study of corpus development for Persian, International Journal of Asian Language Processing 20(1), 2010.

[20] H. Hashemi, A. Shakery, H. Faili, Creating Persian English Comparable Corpus, CLEF, 2010.