International Journal of Information Science and Management
Vol. 14, No. 2, 2016, 97-106

A Hybrid Accurate Alignment method for large Persian-English corpus construction based on statistical analysis and Lexicon/Persian Word net

Mohammad Bagher Dastgheib
Ph.D. Candidate Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
Corresponding Author:
hdastgheib@gmail.com

Seyed Mostafa Fakhrahmad
Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran

Mansour Zolghadri Jahromi
Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran

Abstract

A bilingual corpus is considered as a very important knowledge source and an inevitable requirement for many natural language processing (NLP) applications in which two languages are involved. For some languages such as Persian, lack of such resources is much more significant. Several applications, including statistical and example-based machine translation needs bilingual corpora, in which large amounts of texts from two different languages have been aligned at the sentence or phrase levels. In order to meet this requirement, this paper aims to propose an accurate and hybrid sentence alignment method for construction of an English-Persian parallel corpus. As the first step, the proposed method uses statistical length based analysis for filtering of candidates. Punctuation marks are used as a directing feature to reduce the complexity and increase the accuracy. Finally, the proposed method makes use of some lexical knowledge in order to produce the final output. . In the phase of lexical analysis, a bilingual dictionary as well as a Persian semantic net (denoted as FarsNet) is used to calculate the extended semantic similarity. Experiments showed the positive effect of expansion on synonym words by extended semantic similarity on the accuracy of the sentence alignment process. In the proposed matching scheme, a semantic load based approach (which considers the verb as the pivot and the main part of a sentence) was also used in order for increasing the accuracy. The results obtained from the experiments were promising and the generated parallel corpus can be used as an effective knowledge source by researchers who work on Persian language.

Keywords: Parallel corpora, Hybrid sentence alignment, English-Persian corpus, Extended semantic similarity

Introduction

In recent years, due to the popularity of the Internet, the volume of online texts in distinct languages is increasing tremendously. To find the information needed in any language, a big challenge is the language in which the query is presented. Of course, automatic query

translation to other languages is a good solution to this problem. Further, Statistical Machine Translation (SMT) is an effective approach to build up an automatic machine translation since it can be learned by examples from a parallel corpus. A parallel corpus has many other applications in Natural Language Processing (NLP) including: part-of-speech tagging, word sense disambiguation, language teaching, phrase recognition and information retrieval. Lack of resources for rarely studied languages like Persian is a big challenge to all NLP researchers (Pilevar M., Feili H., Pilevar A., 2011).

Parallel corpora refer to bodies of text in parallel translation (Mohammadi M., GhasemAghaee N., 2010; Chen S.F., 1993; Gale W.A., Church K.W., 1993). Parallel corpora encompass texts in one language with their corresponding translations in some other language or languages (MosaviMiangah T., 2009). In some multilingual countries, formal parallel texts – i.e. parliamentary texts – are also constructed automatically. For dominant languages like Persian, parallel corpora must be aligned automatically or manually. Manual construction of parallel corpora is costly and accordingly not a good candidate for construction of large scale corpora.

To our knowledge, currently, there is no high quality parallel corpus that contains Persian as a language pair, or, if there is, it is not accessible due to copyright restrictions, or the size and quality of the corpora are not suitable for researchers (Pilevar et. Al., 2011; Mohammadi M. and GhasemAghaee N., 2010; Steinberger R., Bruno P., Widiger A., Ignat C., Erjavec T., Tifis D. and Varga D., 2006; Rasooli M.S., Kashefi O., Minaei-Bidgoli B., 2011). Thus, the goal of the present study is to meet this requirement by automatically generating a parallel English-Persian corpus, aligned at the sentence level.

The rest of the paper is organized as follows: Section 2 introduces the alignment and parallel corpus construction methods existing in the literature. The proposed scheme will be presented in details, in Section 3. Section 4 is devoted to the experimental results. Finally, Section 5 concludes the paper.

Literature review

In recent years, many sentence alignment models have come in handy for researchers to construct bilingual corpora of various types. These models can be categorized into two major classes, namely, statistical and non-statistical methods (Gale W. and Church K.W., 1993) the former being further divided into three sub-classes including length-based, lexical-matching, and the finally third one is hybrid approache (Pilevar et. Al., 2011; Mohammadi M. and GhasemAghaee N., 2010; Gale W. and Church K.W., 1993; MosaviMiangah T., 2009; Rasooli et. Al., 2011; Feili H. and Ghassem-Sani G., 2004; Moore R. C., 2002).

Early studies on sentence alignment methods were based on length-based approaches which rely on the theory that the original sentence and its translation have a similar length (Gale W. and Church K.W., 1993). This approach works best for highly correlated language pairs like English-French or English-German; since such languages have very similar alphabets (Gale W. and Church K.W., 1993; Moore R.C., 2002). For example, the English-German parallel corpus – from the economic reports of Union Bank of Switzerland (UBS) – is highly correlated (r=0.991) (Biçici E., 2008). The most important works on length-based sentence alignment were undertaken by Gale and Dunning in two distinct works (Gale W. and

Church K.W., 1993; Dunning T., 1993). This simple statistical model is language independent and can achieve global optimum as well (Rasooli et. al., 2011). Despite its advantages, length-based model suffers a problem, that is, small insertion or deletion can decrease accuracy drastically (Rasooli et. al., 2011; Moore R.C., 2002) besides error propagation (Rasooli et. al., 2011). Another disadvantage refers to when the model tries to align sentences that do not have the same language pair. Under such circumstances, the accuracy decreases and the correlation of the two languages is not high enough to cover the problem.

Lexical-matching approaches rely on bilingual lexicons (Biçici E., 2008). Recently, many studies on lexical- matching have been undertaken (Pilevar et. al., 2011; MosaviMiangah T., 2009; Rasooli et. al., 2011; Gautam M. and Sinha R.M., 2007; Braune F. and Fraser A., 2010; Sarikaya R., Maskey S., Zhang R., Jan E. E., Wang D., Ramabhadran B., and Roukos, S., 2009). In these works, a bilingual lexicon is used to translate a word from the source language into the target language. Then, a similarity measure, i.e. Jaccard or Dice, is used to calculate the similarity of the sentence pairs based on shared words. The main problem with lexical-matching approach is off-the-list vocabulary items – words that are present in the texts but are absent from the lexicon (Biçici E., 2008). In most studies undertaken based on the lexical-matching model, only nouns and their denotative meanings in the lexicon are used to calculate the similarity metric (Pilevar et. al.,2011; Mohammadi M. and GhasemAghaee N., 2010). As some features of the model, it needs more resources like huge bilingual lexicon; further, it suffers inefficient matching of all candidate pairs yet it can overcome the accuracy problem in a satisfactory way (Biçici E., 2008; Rasooli et. al., 2011).

The third model, the hybrid approach, utilizes a combination of statistical and linguistic features to accomplish the alignment task. In recent years, this model has been applied to many language pairs to overcome the accuracy problem (Deng Y., Kumar S. and Byrne W., 2007; Rasooli et. al., 2011). Most of the recent studies on sentence alignment are currently based on hybrid models. Moore (2002) used three phase model to align sentences. In first phase, a length-based filtering was undertaken and some sentences were extracted. In the second phase, IBM model one was applied on extracted parallel sentences from the first phase to build a bilingual lexicon. Finally, in the third phase, lexical and length-based information were used to finalize the alignment process. Deng et. al. (2007) used dynamic programming (DP) and divisive clustering to refine Moore's (2002) model. In another work, Simard et. al. (Simard M., Foster G. F., and Isabelle P., 1993) used cognate's similarity, relying on transliteration as a measure to calculate similarity of sentence pairs. Fattah et. al. (Fattah M. A., Bracewell D. B., Ren F., and Kuroiwa S., 2007) used a combination of punctuation similarity, cognate similarity and length similarity to model the hybrid approach. To combine the features, probabilistic neural network (P-NNT) and Gaussian mixture model (GMM) were drawn on. Sarikaya et. al. (2009) used cosine similarity based on TF-IDF model as a measure to align the sentences.

In what follows literature on parallel corpora, with Persian as a language pair, is presented. Pilevar et. al. (2011) used movie subtitle time overlap to align sentences. This caused a problem: first punctuation marks are not used much in movie subtitles, and hence sentence boundary is not available. This problem is more important for Persian since a Persian sentence does not commence with an upper-case letter. Another problem in informal texts like

movie subtitles is related to multi-shape words (words that are given in many shapes in informal texts). Rule-base correction methods were used to solve such problems. In another work, Mohammadi (Mohammadi M. and GhasemAghaee N., 2010) used Wikipedia pages as a source. First, a bilingual dictionary was constructed from Wikipedia titles using inter-wiki links. Then, a hybrid approach was used to combine length-based and lexical-based similarity metrics to align sentences. In this work, Jaccard was used as the metric that best fits calculation of lexical similarity of sentences. Rasooli et, al. (2011) used hybrid approach was as the aligning method. First, paragraphs were aligned by three similarity measures namely length, punctuation and lexical. After aligning the paragraphs, sentences were aligned by the above three measures. To combine the similarity measures, a linear model like mathematical union was used to compare similarity in sentence pairs. To calculate semantic similarity, a simple dictionary of nouns is used and only the first denotative meaning of each noun is considered to obtain the semantic similarity.

However, researchers in NLP domain suffer from lack of parallel corpora that contain Persian, and are of course freely accessible, as a language pair. Although there are some publicly accessed corpora for Persian like TEP (Pilevar et. al., 2011), as Bijankhan (Bijankhan M., Sheykhzadegan J., Bahrani M., and Ghayoomi, M., 2011) posits all of them have some problems, e.g. informal translation, small size of corpus, specific domain of sentences and some suffering a lot of noise. In a bilingual corpus, noise means insertion or deletion in translation from the source language to the target one. This paper aims to produce a sentence aligned Persian-English corpus automatically with a desired size and formal translation. The alignment process is done by refining the hybrid model for non-uniform language pairs with different alphabets like Persian and English. This model is applied to Persian-English language pairs to calculate the performance and accuracy of the proposed model. A manually aligned corpus from Persian articles that have English abstract is produced for evaluation.

Persian language (or Farsi) is an Indo-Iranian branch of the Indo-European languages and is the official language of Iran, Tajikistan and one of the two main languages used in Afghanistan. Persian has borrowed many words from other languages such as Arabic, but the structure of this language has never been changed, during the past centuries (Pilevar et. al., 2011; Haruno, M. and Yamazaki T., 1996). The Persian language has some free-word-order structures, for example adverbs could appear in the beginning, at the middle or at the end of the sentences.

Written style of Persian is RTL (right to left) and it uses modified Arabic alphabet with four extra alphabets (چ،گ،ژ، پ), and short vowels of the Persian language in the written style often all has been lost and it's very common in writing. This may cause some problems for pronunciation of words. Based on Wikipedia information, about 1% of the world's populations speak Persian as their native language and about 134 million people speak Persian as their first or their second language. All this reiterates the importance of undertaking research on Persian.

The alignment process

As mentioned in the previous sections, one of the main bottlenecks of many NLP applications in Persian language is the lack of a suitable Persian-English corpus which has

been aligned at paragraph, sentence or word level. The aim of this work is to generate a parallel corpus aligned at the sentence level automatically.

The proposed scheme primarily uses length-based methods for pre-processing tasks and then lexical methods are employed to improve the model. The alignment process is divided into three steps. In the first step, the paragraphs are aligned according to their lengths and other lexical features. In the second step, sentence level alignment is carried out. For this purpose, several pairs of sentences from the aligned paragraphs are chosen as candidates for alignment. The candidate sentences must be similar as their lengths by considering a length similarity threshold. As the final step, lexical methods are used in order to select the correctly aligned pairs from the whole set of candidates. Moreover, a new metric named extended semantic similarity (ESS) is also proposed as a semantic feature in this work in order to improve the overall accuracy of the model. This metric is defined in the following sections.

Paragraphs level alignment

In the first step of the alignment process, paragraphs are aligned using a hybrid approach which considers both paragraph lengths and lexical information. In this step, firstly paragraphs are chosen by their length similarity metric and finally semantic metrics are used to finalize the paragraph alignment process. Previous studies in this field show that the pure length based methods do not lead to an acceptable accuracy for the corpora including Persian texts (Pilavar et. al., 2011; Mohammadi M. and GhasemAghaee N., 2010; MosaviMiangah T., 2009; Rasooli et. al., 2011). The alignment method proposed in this paper considers various factors including length similarity (Gale W.A. and Church K.W., 1993), punctuation similarity (Rasooli et. al., 2011) and semantic similarity (Feili et. al., 2011; Gautam M. and Sinha R.M., 2007; Rasooli et. al., 2011) in order to overcome the accuracy problem. The semantic similarity metric makes use of the list synonym words for each word included in the paragraph. The average length of paragraphs in our corpus is about 500 characters or 100~120 words. There are about two or three paragraphs in each abstract we perform 1-1 paragraph aligning, i.e., one paragraph in Persian is aligned to one paragraph in English (Gale W.A. and Church K.W., 1993). The details of the similarity measure will be given in next section.

3.2. Sentence-level alignment

In this step, first of all, a method is employed which aligns and proposes a set of candidate pairs. The sentence alignment method uses dynamic programming (DP) approach which has been proposed by Gale et al. (1993) as a search strategy to pre-align candidate pairs. After generating candidate pairs, a length-based similarity measure is used in order to filter the candidates. The Pearson correlation coefficient for the Persian-English manually aligned training dataset is 0.798, so the test set filtering is performed by a threshold calculated according to this correlation. By this step, a probabilistic score is assigned to each proposed pair of sentences based on the difference between lengths of the sentences (Gale W.A. and Church K.W., 1993; Moore R., 2002). The scored sentences are then filtered with respect to the correlation and the assigned scores. For the length based scoring, the Poison distribution is used, since it only needs one parameter and is simpler than the Gaussian distribution (Gale W.A. and Church K.W., 1993) and has good performance (Rasooli et. al., 2011). The Poison length similarity metric is shown in (eq. 1).

$$P_{length}(s,t) = \frac{e^{-l.s^r}.(l_s.r)^{l_t}}{l_t!} \quad (1)$$

As shown in (eq. 1), lt and ls represent length of target and source sentences in terms of character and r is the sentence length rate. This metric uses Poison distribution (Gale W.A. and Church K.W., 1993). In Poison distribution, only the length rate information between the source and the target texts is required.

3.3-lexical and semantic similarity score

In the final step, two lexical similarity measures are used; the first one is punctuation similarity and the second is extended semantic similarity. For the punctuation similarity just the same as the method used by Rasooli et. al. (2011), we choose 11 punctuations and their equivalents in Persian. For each pair of paragraphs/sentences, the number of distinct punctuations is counted. Equation 2 calculates the punctuation similarity (denoted as Ppunc) for a pair of contexts, namely s and t (Rasooli et. al., 2011).

$$\mathrm{p_{punc}(s,t)} = \frac{\sum_{i=1}^{np} \mathrm{p(punc_i)}}{\mathrm{np}} \quad (2)$$

As shown in (eq. 2) $\mathrm{punc_i}$ is the mark of i-th punctuation which is defined as the proportion of the occurrence number of the punctuation in source and target texts. If the number of i-th punctuation in source and target texts are equal, then the $\mathrm{p(punc_i)}$ equals to one, otherwise it will be a number between zero and one. In (eq. 2), np is the whole number of distinct punctuations.

Extended semantic similarity is the last similarity measure that is used in scoring aligned pairs. This similarity measure uses semantic information of both source and target texts based on a bilingual lexicon. In the latest previous studies in this field (Chen S.F., 1993; Chuang T. C., Wu J. C., Lin T., Shei W. C. and Chang, J. S. ,2005; Feili et. al., 2011; Gautam M. and Sinha R.M., 2007; Rasooli et. al., 2011) only the word in the source language and its first meaning in target language lexicon are used in order to calculate the semantic similarity metric. In the present study, we proposed a new metric that is named extended semantic similarity (ESS). ESS uses the set of word synonyms and their meanings based on Farsnet which is a semantic net structure and lexicon for the Persian language (Shamsfard M., 2008). As an example the words "ساده" or "آسان" in Persian are equivalent and can be translated to “simple” and “easy” in English. These equivalent words can be replaced in the sentences but some equivalent words may not be presented in bilingual lexicon.

Verb is carried out a high semantic meaning of the sentence. So in this work, we propose that an extra bonus can be assigned to matching verbs in source and target sentence. The background of this idea is that deleting distinct part of sentence does not have same effect as deleting verb. So matching verbs in aligning sentences is more important and can affect the accuracy of alignment process. To calculate the similarity metric for sentence pair, Jaccard metric is used with an extra point when verbs are matched in source and target sentences. This method is also used to calculate extended semantic similarity.

After calculating the similarities, the calculated metrics were combined to classify the sentence pairs as relevant (aligned) or non-relevant (non-aligned). In order to train the

classifier, a set of MSRT's article titles were also used. The aligned sentences of the abstracts were then used as the test set in a cross validation process to overcome the lack of data volume. The three similarity measures were calculated and the trained a Naïve-Bayes classifier was used to combine the similarity scores and classify the sentences to relevant or non-relevant.

Experiments and Results

The effectiveness of the proposed alignment method was evaluated through a set of experiments. For stemming the used English words the well-known open source stemmer, Porter, was used. A similar algorithm was also used for Persian words. The data set used in this project was taken from Iran MSRT which includes more than 40K articles in English as well as their abstracts in both English and Persian languages. To evaluate the alignment method, 30 parallel articles from different categories were selected from Iran MSRT article repository. A manually sentence aligned dataset was produced randomly in size about 10K sentence pairs, and article abstracts were chosen with normal distribution over all subject categories like: humanities, engineering, medical science and etc. Sentence of abstracts were used in a DP framework to propose candidate sentences, then the sentence pairs were checked manually.

The types of alignment of sentence pairs are categorized in three classes. The simplest alignment type is one-to-one (1-1), which one sentence from the source language is aligned to one sentence in the target language. The experiments showed that more than 80% of alignments in a corpus are one-to-one. The second type of alignment is to align two (or more) sentences from source language to one sentence in the target language. This type of alignment is named 2-1. Similarly, 1-2 shows an alignment that one sentence in the source language is aligned to two (or more) sentences in the target language. The last type of alignment is that one sentence in source language is deleted from the translated target language (1-0); this sentence cannot be aligned to any sentences in the targeted language. Similarly the (0-1) type of alignment shows that one sentence added to the targeted language which cannot be found in the source language. The manually aligned dataset used in this work has more that 10K sentences. The types of alignments available for this dataset are shown in table 1.

Table 1
The distribution over the alignment types

Type of alignment	Percentage of total
1-1	91.5%
2-1 and 1-2	7.5%
1-0 and 0-1	1%

As shown in table 1, the manually aligned dataset has a minimum noise (1%). The text translation strategies for articles are formal translation, so the text of this manually aligned corpus has high quality. The Pearson correction coefficient for this dataset is r=0.79. This shows that length of Persian and English sentences are not highly correlated. The three similarity metrics that are described before are computed for each sentence pairs. Then the

aligned sentence pairs were checked manually and labelled as relevant (correctly aligned) or non-relevant (incorrectly aligned or not-aligned). In order to evaluate the proposed methods, this manually aligned dataset was used for further evaluations. As the first experiment, the base line pure length based method was evaluated on this dataset and the alignment precision was not more than 35.5%. This evaluation showed that the pure length based method cannot be accurate enough for pairs of languages that have different alphabet set like Persian-English. This experiment also showed that the correlation length of Persian and English sentences is not high enough to overcome the accuracy problem. In this case, small noises can drop the accuracy drastically. In the second set of experiments, semantic similarity, punctuation similarity and extended semantic similarity each in two modes (i.e., with or without considering extra semantic load for verbs) were evaluated in turn. The precision, recall and finally F1-measure were calculated for all sets of the experiments. The results of these experiments are given in Table 2.

Because of the baseline accuracy based on length method is low (precision=0.355), the hybrid approach applied in four sets of experiments. In the first set, the traditional semantic similarity and punctuation similarity are used for alignment task. The result of this experiment showed that precision increased but recall is low. This means that some correct link between sentences cannot be found. In the second set of experiment, extended semantic similarity added to the two traditional metrics but no extra point is calculated for matching verbs. By this way precision and recall and finally F1-measure are increased. This experiment showed that ESS is effective and can improve the performance of the alignment. In the third set of experiments, an extra bonus on matching simple verbs is considered. In this experiment, the complex verbs that contain more than one word are not supported. The results are not so desirable, so in the fourth set of experiments, complex verb matching is enabled. The result of final experiment showed that semantic load on verbs with supporting complex verbs are quite effective and can increase precision, recall and finally F1-measure.

Table2
The Results of sentence alignment methods

Method	Precision	Recall	F1-measure
1.Semantic similarity + Punctuation similarity	0.8152	0.25	0.3827
2.Semantic similarity + Punctuation similarity + Extended Semantic similarity (without semantic load on verbs)	0.9022	0.56	0.6929
3.Extended semantic + Semantic similarity+ punctuation similarity (with semantic load on verbs without considering complex verbs)	0.8478	0.5	0.6290
4.Extended semantic + Semantic similarity+ punctuation similarity (with semantic load on verbs with checking complex verbs)	0.9201	0.58	0.7115

In all experiments, to combine the similarity measures achieved from hybrid method used, a Naïve-Bayes classifier is trained by extracted metrics. Then for evaluation purpose, test dataset used in a 5 fold cross validation process. The experiments showed that synonym

expansion and a load bias on verbs can improve the efficiency of the alignment process at the sentence level. The semantic load is considered in two cases, i.e., with and without detecting complex verbs. As shown in the results, ESS and semantic load on verbs are effective, but these metrics are more complex to be calculated. So these metrics are suitable for language pairs that other methods like traditional length based method or semantic similarity method cannot achieve good results.

Conclusion

Lack of enough resources on Persian is one of the challenges of research on Persian NLP. Some problems Persian NLP studies suffer from include: lack of publicly accessed bilingual dictionary, availability of stemmer, and large scale bilingual parallel corpora.

In this work, it was intended to automatically produce a large parallel corpus for Persian-English pair, aligned at sentence level. To do so, use of a reliable and accurate method for alignment is inevitable. The hybrid model with a newly proposed extended semantic similarity metric was used to align about 40K text of article abstracts in Persian and English and the results will be accessible for future studies online[1].

As shown in the experiments, the base line pure length-based method could not achieve good accuracy. Hence, the hybrid method was used in order to improve the accuracy level.

The experiments showed that lexical semantic methods can improve the accuracy more significantly. The results of the present study revealed that exploiting on the verb's semantic load with supporting complex form of verbs can further improve the accuracy level. Due to the nature of the Persian Language, and abundant usage of synonymous words in this language, expansion of words based on a semantic net can be very effective in improving Precision and Recall measures of the system. This new metric based on semantic net is named extended semantic similarity (ESS) and can improve the overall performance of the alignment.

Endnote

1. This dataset will be available online freely at http://nlp.ricest.ac.ir

References

Biçici, E. (2008). Context-based sentence alignment in parallel corpora. Lecture Notes in Computer Science, 4919, 434-444.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. Language Resources and Evaluation, 45(2), 143-164.

Braune, F., & Fraser, A. (2010, August). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 81-89). Association for Computational Linguistics.

Chen, S. F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio (1993) 9–16.

Chuang, T. C., Wu, J. C., Lin, T., Shei, W. C., & Chang, J. S. (2005). Bilingual sentence alignment based on punctuation statistics and lexicon. In Natural Language Processing–IJCNLP 2004 (pp. 224-232). Springer Berlin Heidelberg.

Deng, Y., Kumar, S., & Byrne, W. (2007). Segmentation and alignment of parallel text for statistical machine translation. Natural Language Engineering, 13 (3), 235-260.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61-74.

Fattah, M. A., Bracewell, D. B., Ren, F., & Kuroiwa, S. (2007). Sentence alignment using P-NNT and GMM. Computer Speech & Language, 21(4), 594-608.

Feili, H., & Ghassem-Sani, G. (2004, August). An application of lexicalized grammars in English-Persian translation. In ECAI (Vol. 16, p. 596).

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 75-102.

Gautam, M., & Sinha, R. M. K. (2007, March). A hybrid approach to sentence alignment using genetic algorithm. In Computing: Theory and Applications, 2007. ICCTA'07. International Conference on (pp. 480-484). IEEE.

Haruno, M., & Yamazaki, T. (1997). High-performance bilingual text alignment using statistical and dictionary information. Natural Language Engineering, 3(1), 1-14.

Mosavi Miangah, T. (2009). Constructing a large-scale english-persian parallel corpus. Meta: Journal des traducteursMeta:/Translators' Journal, 54(1), 181-188.

Pilevar, M. T., Faili, H., & Pilevar, A. H. (2011). Tep: Tehran english-persian parallel corpus. In Computational Linguistics and Intelligent Text Processing (pp. 68-79). Springer Berlin Heidelberg.

Mohammadi, M., & GhasemAghaee, N. (2010, March). Building bilingual parallel corpora based on wikipedia. In Computer Engineering and Applications (ICCEA), 2010 Second International Conference on (Vol. 2, pp. 264-268). IEEE.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora (pp. 135-144). Springer Berlin Heidelberg.

Rasooli, M. S., Kashefi, O., & Minaei-Bidgoli, B. (2011). Extracting parallel paragraphs and sentences from english-persian translated documents. InInformation Retrieval Technology (pp. 574-583). Springer Berlin Heidelberg.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E. E., Wang, D., Ramabhadran, B., & Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In INTERSPEECH (pp. 432-435).

Shamsfard, M. (2008). Developing FarsNet: A lexical ontology for Persian. In4th Global WordNet Conference, Szeged, Hungary.

Simard, M., Foster, G. F., & Isabelle, P. (1993, October). Using cognates to align sentences in bilingual corpora. In Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2 (pp. 1071-1082). IBM Press.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. arXiv preprint cs/0609058.