Extracting Parallel Paragraphs and Sentences from English-Persian Translated Documents*

Mohammad Sadegh Rasooli, Omid Kashefi, and Behrouz Minaei-Bidgoli

Department of Computer Engineering, Iran University of Science and Technology
rasooli@comp.iust.ac.ir, kashefi@{ieee.org,iust.ac.ir},
b_minaei@iust.ac.ir

Abstract. The task of sentence and paragraph alignment is essential for preparing parallel texts that are needed in applications such as machine translation. The lack of sufficient linguistic data for under-resourced languages like Persian is a challenging issue. In this paper, we proposed a hybrid sentence and paragraph alignment model on Persian-English parallel documents based on simple linguistic features as well as length similarity between sentences and paragraphs of source and target languages. We apply a small bilingual dictionary of Persian-English nouns, punctuation marks, and length similarity as alignment metrics. We combine these features in a linear model and use genetic algorithm to learn the linear equation weights. Evaluation results show that the extracted features improve the baseline model which is only a length-based one.

Keywords: Sentence Alignment, Paragraph Alignment, Parallel Corpus, Bilingual Corpus, Persian, English, Machine Translation.

1 Introduction

A parallel corpus is made up of sentences of two languages that each sentence from one language is a translation of the sentence in the other one [1]. Sentence aligned corpora are necessary for the task of statistical machine translation. Furthermore, there are many other applications of parallel corpora such as lexicography and language analysis [2]. The importance of parallel corpus increased since the IBM translation models [3] has been introduced. One of the main steps of building a parallel corpus is sentence alignment [4], so in order to obtain a sentence aligned corpus, it is necessary to perform sentence alignment on bitext documents. Sentence alignment is the task of mapping each sentence in the source language to its corresponding sentence (or sentences) in the target language [5]. In addition, if the parallel data are not aligned in the paragraph level, it is essential to align paragraphs before the sentence alignment task, because knowing paragraph boundaries help to reduce candidate space for sentence alignment. There are some ways to collect bilingual texts. Using widely translated books such as holy religious books [6], multilingual catalogues [7], or getting parallel texts from web [8] are some of the ways for

* This paper is funded by Computer Research Center of Islamic Sciences (CRCIS).

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 574–583, 2011.

collecting bitext data. Therefore, building appropriate parallel text is a hardship for under-resourced languages similar to Persian language.

Persian is a variation of Arabic-script language that is mostly spoken in Iran, Afghanistan, Tajikistan and some parts of India and Pakistan. One of the most challenging issues of Persian language processing and information retrieval (IR) is the lack of feasible corpus. For example, to our knowledge, there is no syntactically and/or semantically tagged corpus in Persian. There are some bilingual Persian-English corpora. One of them is introduced in [9]. Another Persian-English parallel corpus is introduced in [10] that is made of Persian-English aligned movie subtitles wherein the Persian parts are in colloquial shape instead of official Persian language that is used in academic and governmental organizations and medias. Therefore, it is required to have a corpus of official Persian language. The process of manually aligning corpora is a time consuming task, where intelligent computer programs may help reduce the time, so automatically building parallel Persian-English corpora made sentence is a hot topic.

In this paper, considering the need for a feasible parallel Persian-English corpus, we propose a hybrid approach to extract aligned paragraphs and sentences from translated documents. We used some clues such as paragraph or sentence length, punctuation marks and a small bilingual lexicon of simple one-word nouns. In the following sections, after reviewing some related works on sentence alignment in Section 2, we describe our proposed method in Section 3. We have done two experiments on Persian-English data, one for paragraph alignment and one for sentence alignment that are described in Section 4. Finally, Section 5 concludes the paper.

2 Related Works

There have been many sentence alignment models in recent years that can be categorized into three model types: 1) length-based approaches, 2) Lexical matching approaches, and 3) Hybrid approaches. Furthermore, there are some other approaches such as measuring cognate similarity between sentences. Most of the works on sentence alignment models are based on the assumption that paragraph anchors are aligned [11-15]. Even though, in some works, candidate sentences are chosen based on a window size of adjacent sentences in the text without any information about paragraph anchors [16].

The first attempts on sentence alignment were done based on length-based models. In the length based approach, it is assumed that the sentence pairs of source-target languages are similar in their length. For example, the sentences in German-English parallel corpus have a correlation of 0.91 [17]. The first attempts on length-based approaches were in [11, 12, 17]. Not only this model is very simple and language independent, but also it can gain global optimum [18]. On the other hand, small deletions and insertions decrease the accuracy drastically [14] and error propagation may happen [18]. The second approach is based on lexical matching. In this approach, bilingual lexicons are used as guides to alignment. The early works on this approach were done by [14, 15, 19, 20].

In the third approach, the combination of statistical and linguistic features (such as bilingual lexical matching and simple linguistic clues) is used as a guide to the alignment task. The main reason to use this approach is that for many languages, simple statistical approaches do not gain enough accuracy and there is a need for new approaches to overcome this problem [21]. Most of the recent works on sentence alignment is based on hybrid models. In [13], three phases were used in order to extract aligned sentences. In the first phase, some aligned sentences were extracted via length based models. In this phase, a threshold was considered on length similarity to select only reliable parallel sentences in the corpora. In the second phase, IBM model 1 was used on the extracted aligned corpora and a bilingual lexicon was built. In the last phase, the program used both the gained lexical information and length similarity to find aligned sentences. This work became state of the art in its time and many other researchers improved this model. In [22], a combination of dynamic programming (DP) and divisive clustering was used to improve Moore's model[13].In this work, DP allows many-to-many alignments and divisive clustering refines those alignments with iterative binary splitting. In [23], a two step clustering approach was used to improve both accuracy and efficiency of Moore's model. In the first step the program finds a model-optimal alignment made up of possible 0/1 to 0/1 alignments and in the second step, it merges those alignments into larger ones. That method was 550 times faster than the work in [22].

In [24] cognates similarity (similarity based on transliteration) was used as a measure of similarity. In [25], the order of punctuation marks in bitext and lexical information were combined to achieve aligned sentences. In [26], the combination of punctuation marks, cognates and length similarity was used to find better alignments. In addition, probabilistic neural network (P-NNT) and Gaussian mixture models (GMM) were used to combine those features. In [5], a modification of the Champollion in [27] was proposed. This approach was based on a hybrid model that optimized the process of splitting the bilingual texts into small parts for alignment. In [28], an iterative model was used to improve alignment accuracy. That work was an extension of [29] that was used for aligning OCR generated texts. In that work, a length based approach was used in the initialization phase. In the next iterations, a statistical machine translation (SMT) model was built and based on that model; bleu measure was used in the next iterations to compare the translated text by the translation model to the candidate sentences. In the last iteration, the final SMT model was built from the sentence aligned corpora. In [30], a bootstrapping algorithm was done on bitext based on cosine similarity measure to measure similarity of the documents based on TF-IDF. In [31], Wikipedia was considered as a good source of multilingual data with many noises such as sentences without translations. Some features such as date matching, same pictures in the pages and a little manually aligned train data for building bilingual lexicon was used as guides to alignment. In [32], based on the assumption that parallel web pages have similar page structures and translators respect the original structure of the document, an HTML tree alignment model was proposed using dynamic programming. In [33], a language independent context model based on Zipfian word vector was proposed to improve sentence alignment problem. The Zipfian word vector is a vector of values in the sentences based on the logarithmic division of the word frequency in the sentence context and a threshold. In that work, dynamic programming was used to align sentences. There are also some alignment works on movie subtitles based on simple clues such as sentence length and time overlap that some of these works are proposed in [34, 35].

3 Extracting Parallel Paragraphs and Sentences

One of the main problems we faced in Persian-English parallel corpus extraction is the lack of paragraph aligned corpora. Indeed, in this paper we focus on introducing a new model that can be used both for sentence alignment and paragraph alignment. In the study we have done on Persian-English bitext, we found that the most of the paragraph lengths are in the similar length order and using pure length-based models results in unreliable alignment. On the other hand, the methods used in [13] is employed IBM model 1 [3] which is not efficient for long paragraphs. The length of paragraphs in Persian is about 100 words (based on our test bed). The IBM model performance is proportional to the sentence length, where for long sentences the space of candidates and processing time increases. We chose three most efficient ones as: 1) length similarity, 2) punctuation mark similarity, and 3) semantic similarity between words of source and target paragraphs or sentences that is calculated exploiting a bilingual dictionary of nouns.

3.1 Feature Similarities

For the length based similarity we used Poisson distribution that is employed in [13, 26, 31]. This distribution has only one parameter and is simpler than Gaussian distribution used in [11]. The Poisson distribution only needs the length rate between the source and target sentences as shown in (1) where l_t and l_s are sentence length of the target and source languages and r is the sentence length rate.

$$p_{\mathrm{length(S,T)}} = \frac{e^{-l_s r} . (l_{s.r}{}^{l_t})}{l_t\,!} \tag{1}$$

For the punctuation similarity, we chose 11 different punctuation types that are comparable in Persian and English[1]. Equation (2) calculates the punctuation similarity score of each punctuation mark (punc_i), where min(s,t) and max(s,t) are the minimum and maximum number of occurrence of punctuation punc_i in the source and target sentence or paragraph respectively.

[1] We replace brackets with their open form; as an example, close parenthesis mark is replaced with open parenthesis mark. Therefore, punctuation marks in the text are mapped into the (, ; ? ! . - { [" : set. Some Persian punctuation marks are different but corresponding with English ones such as "،" in Persian that corresponds to "," in English, or "؟" in Persian corresponds "?" in English.

$$p(punc_i)=\frac{\min(s,t)}{\max(s,t)},\max(s,t)>0 \tag{2}$$

So, the overall punctuation probability will be as equation (3), where np is the number of distinct punctuations.

$$p_{punc}(S,T)=\frac{\sum_{i}^{np} p(punc_i)}{np} \tag{3}$$

We used a small dictionary of Persian one-word nouns and their English translations. In order to calculate the semantic similarity of sentences or paragraphs, we count the co-occurrence of translated nouns (either in word form or stem form) in source and target sentences or paragraphs as shown in (4) where $dic(s_i)$ and $dic(t_i)$ are the number of translated word occurrences. In addition, *count(S)* and *count(T)* are the number of distinct words in the source and target languages respectively. In other words, for each noun in the source sentence or paragraph, if the corresponding translated word exists in the target sentence or paragraph, we count a co-occurrence score for that word.

$$p_{dic}(S,T)=\frac{\sum_{i,j}\frac{\min(dic(s_i),dic(t_i))}{\max(dic(s_i),dic(t_i))}}{\max(count(S),count(T))},\max(dic(s_i),dic(t_i))>0 \tag{4}$$

3.2 Combining Similarities

In order to combine similarity scores, we used a linear model similar to the mathematical union, in which each part of the equation is weighted by a coefficient as in (5) where w_1 and w_2 are in [0, 1] and the other coefficients are in [-1, 1]. The reason that we used this form is that it is simple to use and with this type of equation all types of linear combinations of variables is considered, so the learner adapts appropriate weight to each coefficient.

$$\begin{aligned}
&\mathrm{p}_{\mathrm{align}}(\mathrm{S,T})=\mathrm{w}_1\,\mathrm{p}_{\mathrm{punc}}(\mathrm{S,T})+\mathrm{w}_2\,\mathrm{p}_{\mathrm{length}}(\mathrm{S,T})+\mathrm{w}_3\,\mathrm{p}_{\mathrm{dic}}(\mathrm{S,T})\\
&+\mathrm{w}_4\,\mathrm{p}_{\mathrm{punc}}(\mathrm{S,T}).\mathrm{p}_{\mathrm{length}}(\mathrm{S,T})+\mathrm{w}_5\,\mathrm{p}_{\mathrm{punc}}(\mathrm{S,T}).\mathrm{p}_{\mathrm{dic}}(\mathrm{S,T})\\
&+\mathrm{w}_6\,\mathrm{p}_{\mathrm{dic}}(\mathrm{S,T}).\mathrm{p}_{\mathrm{length}}(\mathrm{S,T})+\mathrm{w}_7\,\mathrm{p}_{\mathrm{punc}}(\mathrm{S,T}).\mathrm{p}_{\mathrm{dic}}(\mathrm{S,T}).\mathrm{p}_{\mathrm{length}}(\mathrm{S,T})
\end{aligned} \tag{5}$$

With this assumption, it is needed to find 18 unknown weight coefficients in (5).

3.3 Using Genetic Algorithm for Weight Learning

We used genetic algorithm to find the unknown weight coefficients. It is worth to say that in[36], genetic algorithm was also used for sentence alignment. Elitism is used in the genetic algorithm in order to keep good chromosome in each generation. The fitness function is shown in (6). The precision of the weights of the chromosome in the training data is used in the equation to find the fitness.

$$\text{Fitness}(\text{cromosome}) = \frac{e^{\frac{\text{precision}(\text{cromosome})}{10}}}{\sum e^{\frac{\text{precision}(\text{cromosome})}{10}}} \tag{6}$$

4 Experiments and Results

We evaluate the effectiveness of our proposed alignment method for both sentence and paragraph alignment. In the *experiment 1*, we evaluated our method through paragraph alignment. The length rate for Poisson distribution has been chosen as 1. Uniform distribution was used to choose each chromosome gene for crossover in learning part using genetic algorithm. The uniform probability is chosen as 0.5 and the crossover rate as 0.9. The mutation rate is chosen as 0.01. In the *experiment 2*, we used our method for sentence alignment. To stem English words, we used an open-source code of Porter stemmer [37] and for Persian words, we used lemmatization code proposed in [38][2].

4.1 Experiment 1: Paragraph Alignment

In order to evaluate the effectiveness of semantic similarity and exploiting a bilingual dictionary, we relaxed (5) to (7) and compare the results.

$$\begin{aligned} p_{\text{align}}(S,T) &= w_1\, p_{\text{punc}}(S,T) + w_2\, p_{\text{length}}(S,T) \\ &+ w_3\, p_{\text{punc}}(S,T).p_{\text{length}}(S,T) \end{aligned} \tag{7}$$

In this experiment, we extracted about 400 lecture paragraphs from the Iran supreme leader official website[3] that provides both English and Persian edition of lectures. The translation mode in this website is free and there are many free deletions and insertions in the translations. There are also some Arabic sentences (like holy Quran verses) during lectures. In order to find parallel documents, we matched the dates of the lecture pages and aligned the paragraphs manually. Finally, we prepared a paragraph level aligned corpus with about 300 Persian paragraphs and about 350 corresponding English

[2] The Persian lemmatizer code is available in the Virastyar software pack in: `http://sourceforge.net/projects/virastyar/`.

[3] `http://www.khamenei.ir`

Paragraphs. The reason of difference between the numbers of Persian and English paragraphs in the manually aligned corpus is a source paragraph may be translated into two or more paragraphs in target language as we have one-to-one, one-to-two, one-to-three, and one-to-four alignment in our corpus. To overcome the small training data issue, we used 6-fold cross-validation. The results of paragraph alignment using our proposed method are shown in Table 1. In order to compare the results with the baseline length-based model, we tested the baseline on this data and did not get alignment precision more than 35% considering only the length of paragraphs, which means that the pure length-based model is not appropriate enough for sentence alignment.

Table 1. Results of paragraph alignment experiment

Fold #	*Precision with Semantic Similarity*	*Precision without Semantic Similarity*
1	100	71.93
2	94.12	79.41
3	100	44.12
4	90	70
5	89.01	47.26
6	81.25	59.37
Average	**92.40**	**62.01**

4.2 Experiment 2: Sentence Alignment

After experiment 1, we developed a visual user-friendly software for manually aligning sentences. We used translated novels as our initial data. In order to build a binary classifier, all bilingual sentences pairs are labeled by our method as “are-aligned” or “not-aligned”. In this way, we are able to calculate both precision and recall. We also evaluate the effectiveness of punctuation similarity relaxing (5) to (8).

$$\begin{aligned} p_{\text{align}}(S,T) &= w_1\, p_{\text{dic}}(S,T) + w_2\, p_{\text{length}}(S,T) \\ &+ w_3\, p_{\text{dic}}(S,T).p_{\text{length}}(S,T) \end{aligned} \tag{8}$$

The dataset of our test is composed of 26,108 aligned sentences (13,054 sentences in each language). The results are shown in Table 2, the punctuation similarity improved the model performance, but the improvement is not significant. The reason is that for sentences, numbers of punctuation marks are very small and it does not guide the model very well. On the other hand, in the paragraph level, there are much more punctuation marks which help the model predict better alignments.

Table 2. Results of sentence alignment experiment

Method	*Precision*	*Recall*	*F_1-Measure*
Precision with punctuation similarity	96.63	79.42	86.48
Precision without punctuation similarity	92.25	79.45	86.03

5 Conclusion

We faced some problems in using parallel data; as an instance, we tried to use texts extracted by OCR but due to poor performance of Persian OCR software, rate of erroneous recognized words was very high. The problem of copyright in translations, un-uniform Persian characters, the lack of bilingual electronic texts and colloquial language typography in many Persian texts made the task in Persian harder than languages like English.

As seen in the experiments, the performance of the model depends on the task. If it used to paragraph alignment, the punctuation mark and bilingual dictionary significantly improve the accuracy of the pure length-based model result about 35%. But in the sentence alignment, the length and number of words are small enough to suppress other features. Furthermore, the cognate similarity is not applicable to Persian-English texts, because the alphabet types of these two languages are different.

Acknowledgement. This paper is funded by Computer Research Center of Islamic Sciences (CRCIS). I would also like to thank Dr. Shahram Khadivi, Dr. Heshaam Faili, Maryam Aminian, Sina Iravaninan, Dr. Morteza Analoui, Mehrdad Senobari and other people who help me on this work.

References

1. Frankenberg-Garcia, A.: Compiling and using a Parallel Corpus for Research in Translation. International Journal of Translation (2009)
2. Tiedemann, J.: Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. In: Faculty of Languages, Department of Linguistics. Uppsala University, Uppsala (2003)
3. Brown, P., et al.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)
4. Simões, A., Almeida, J.J.: Parallel Corpora based Translation Resources Extraction. Procesamiento del Lenguaje Natural, 265–272 (2007)
5. Li, P., Sun, M., Xue, P.: Fast-Champollion: a fast and robust sentence alignment algorithm. In: 23rd International Conference on Computational Linguistics: Posters, pp. 710–718. Association for Computational Linguistics, Beijing (2010)
6. Resnik, P., Olsen, M.B., Diab, M.: The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues". Computers and the Humanities 33, 129–153 (1999)
7. Menezes, A., Richardson, S.D.: A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation, Toulouse, France, pp. 39–46 (2001)
8. Resnik, P., Smith, N.A.: The Web as a Parallel Corpus. Computational Linguistics 29, 349–380 (2003)
9. Miangah, T.M.: Constructing a large-scale english-persian parallel corpus. Meta: Translators' Journal 54(1), 181–188 (2009)
10. Pilevar, M.T., Faili, H., Pilevar, A.H.: TEP: Tehran English-Persian Parallel Corpus. In: Gelbukh, A. (ed.) CICLing 2011, Part II. LNCS, vol. 6609, pp. 68–79. Springer, Heidelberg (2011)

11. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning Sentences in Parallel Corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, pp. 169–176 (1991)
12. Gale, W.A., Church, K.W.: A program for Aligning Sentences in Bilingual Corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, pp. 177–184 (1991)
13. Moore, R.C.: Fast and Accurate Sentence Alignment of Bilingual Corpora. In: Richardson, S.D. (ed.) AMTA 2002. LNCS (LNAI), vol. 2499, pp. 135–144. Springer, Heidelberg (2002)
14. Chen, S.F.: Aligning Sentences in Bilingual Corpora Using Lexical Information. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp. 9–16 (1993)
15. Wu, D.: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp. 80–87 (1994)
16. Meyers, A., Kosaka, M., Grishman, R.: A Multilingual Procedure for Dictionary-Based Sentence Alignment. In: Farwell, D., Gerber, L., Hovy, E. (eds.) AMTA 1998. LNCS (LNAI), vol. 1529, pp. 187–198. Springer, Heidelberg (1998)
17. Gale, W.A., Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics 19(1), 75–102 (1993)
18. Tiedemann, J.: Bitext Alignment. Synthesis Lectures on Human Language Technologies 4, 1–165 (2011)
19. Kay, M., Röscheisen, M.: Text-translation alignment. Computational Linguistics 19(1), 121–142 (1993)
20. Melamed, I.D.: A geometric approach to mapping bitext correspondence. In: Conference on Empirical Methods in Natural Language Processing, EMNLP (1996)
21. Haruno, M., Yamazaki, T.: High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. Natural Language Engineering 3(1), 131–138 (1997)
22. Deng, Y., Kumar, S., Byrne, W.: Segmentation and alignment of parallel text for statistical machine translation. Natural Language Engineering 13(03), 235–260 (2007)
23. Braune, F., Fraser, A.: 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, Beijing (2010)
24. Simard, M., Foster, G.F., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In: Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992), Montreal, Canada, pp. 67–81 (1992)
25. Chuang, T.C., Wu, J.-C., Lin, T., Shei, W.-C., Chang, J.S.: Bilingual Sentence Alignment Based on Punctuation Statistics and Lexicon. In: Su, K.-Y., Tsujii, J., Lee, J.-H., Kwong, O.Y. (eds.) IJCNLP 2004. LNCS (LNAI), vol. 3248, pp. 224–232. Springer, Heidelberg (2005)
26. Fattah, M.A., et al.: Sentence alignment using P-NNT and GMM. Computer Speech & Language 21(4), 594–608 (2007)
27. Ma, X.: Champollion: A robust parallel text sentence aligner. In: LREC 2006: Fifth International Conference on Language Resources and Evaluation, pp. 489–492 (2006)
28. Sennrich, R., Volk, M.: Iterative, MT-based Sentence Alignment of Parallel Texts. In: 18th Nordic Conference of Computational Linguistics, NODALIDA 2011 (2011)

29. Sennrich, R., Volk, M.: MT-based Sentence Alignment for OCR-generated Parallel Texts. In: The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado (2010)
30. Sarikaya, R., et al.: Iterative Sentence-Pair Extraction from Quasi-Parallel Corpora for Machine Translation. In: 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), Brighton, United Kingdom (2009)
31. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 403–411. Association for Computational Linguistics, Los Angeles (2010)
32. Shi, L., Zhou, M.: Improved sentence alignment on parallel web pages using a stochastic tree alignment model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 505–513. Association for Computational Linguistics, Honolulu (2008)
33. Biçici, E.: Context-Based Sentence Alignment in Parallel Corpora. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 434–444. Springer, Heidelberg (2008)
34. Tiedemann, J.: Improved sentence alignment for movie subtitles. In: Conference on Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria, pp. 582–588 (2007)
35. Tiedemann, J.: Synchronizing translated movie subtitles. In: 6th International Conference on Language Resources and Evaluation, LREC 2008 (2008)
36. Gautam, M., Sinha, R.M.K.: A Hybrid Approach to Sentence Alignment Using Genetic Algorithm. In: International Conference on Computing: Theory and Applications (ICCTA 2007), pp. 480–484. IEEE Computer Society (2007)
37. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1980)
38. Kashefi, O., Nasri, M., Kanani, K.: Automatic Spell Checking in Persian Language. Supreme Council of Information and Communication Technology (SCICT), Tehran (2010)