ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 5, No.11 , September 2014
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Extracting Persian-English Parallel Sentences from Document Level Aligned Comparable Corpus using Bi-Directional Translation

Ebrahim Ansari[1], Mohammad Hadi Sadreddini[2], Alireza Tabebordbar[3] and Richard WALLACE[4]

[1] Department of Computer Science and Engineering, Shiraz University
Shiraz, Fars, Iran
ansari@cse.shirazu.ac.ir

[2] Department of Computer Science and Engineering, Shiraz University
Shiraz, Fars, Iran
sadredin@ shirazu.ac.ir

[3] Department of Computer Science and Engineering, Shiraz University
Shiraz, Fars, Iran
tabebordbar@tuv.ac.ir

[4] Distributed Systems Architecture Research Group, Complutense University
Madrid, Spain
wallacerim@aol.com

Abstract

Bilingual parallel corpora are very important in various filed of natural language processing (NLP). The quality of a Statistical Machine Translation (SMT) system strongly dependent upon the amount of training data. For low resource language pairs such as Persian-English, there are not enough parallel sentences to build an accurate SMT system. This paper describes a new approach to use the Wikipedia as a comparable corpus to extract Persian-English parallel sentences and eventually improve SMT system performance. This new approach is also applicable to other low resource language pairs. In order to calculate the similarity score between two sentences, a novel bi-directional translation-based information retrieval system is proposed. A length penalty score is introduced to increase the accuracy of extracted corpus. Using extracted parallel sentences, the performance of existing Persian-English SMT is improved drastically.

Keywords: *comparable corpus, bi-directional translation, statistical machine translation, Wikipedia, information retrieval*

1. Introduction

In recent years, machine translation (MT) systems have obtained reasonable results when applied to some popular language pairs such as English-French and English-Chinese. However, studies on statistical MT for low-resourced languages are always faced with the challenge of getting enough data to support any particular approach. Statistical machine translation (SMT) uses statistical methods based on large parallel bilingual corpora of source and target languages to build a statistical translation model. SMT also uses target language texts to build a statistical language model. These two models and a search (decoding) module are used to decode and find the best translation for each source language sentence [1]-[2].

The performance of a statistical machine Translation System depends significantly on the amount of training data. For some language pairs, such as Persian-English, there aren't enough parallel corpora for the training phase that eventually builds a statistical machine translation system. In recent years, variety of methods are proposed methods to extract parallel sentences from non-parallel (comparable) bilingual corpora.

Wikipedia is an online and multilingual encyclopedia which contains different articles in a variety of domains. Each article is linked to the article with the same topic in another language by inter-language link structure. Considering the number of articles, Persian is the 20th language in Wikipedia. The containing of more than 300,000 articles, shows Persian Wikipedia documents are ready for parallel sentence extraction and other multilingual research.

In the first step of our approach, all Wikipedia Persian and English documents are extracted. Afterwards our algorithm is applied for each Persian document with an existed English correspondence and potential parallel sentences are extracted. Then final filtering phase is applied on top score sentences and parallel sentences are extracted. We proposed a Bi-directional information Retrieval approach to find sentences in the target language that are the most probable translation of the source language.

This paper is structured as follows: In Section 2, the literature and related works are reviewed. Afterwards, the procedure of Wikipedia document extraction and our proposed Information Retrieval system and parallel sentence extraction are described in Section 3. Then, in section 4 we describe the results and evaluate our approach and the paper ends with our conclusion.

2. Related Work

Comparable corpus as a source of translation knowledge has attracted the attention of many researchers. Unlike parallel corpora, which are clearly defined as translated text, there is a wide variation of non-parallelism in comparable texts. Non-parallelism is manifested in terms of differences in author, domain, topics, time period and language. Most common text corpora have non-parallelism in all these dimensions. The higher the degree of non-parallelism, the more challenging is the extraction of bilingual information. Wikipedia is a multi-lingual comparable resource with a high comparability degree. Considerable amount of research has done for extracting parallel sentences or creating comparable corpora [3-17].

Moreover and in recent years, Wikipedia is a well-known and rich comparable corpus in various fields of natural language processing and machine translation systems [7], [13-14] and [16-17].

Zhao and Vogel [3] propose a maximum likelihood criterion which combines sentence length model and a statistical translation lexicon model extracted from an already existing aligned parallel corpus. An iterative process was applied to retrain the translation lexicon model with the extracted data. Their selected languages were Chinese and English.

Utiyama and Isahara [4] use cross language information retrieval techniques and dynamic programming to extract parallel sentences from an English–Japanese news corpus. The authors first try to find similar article pairs, and then, they treat these pairs as parallel texts, align their sentences on a sentence pair similarity score. Subsequently, they use dynamic programming to find the minimum-cost alignment over each document pair. They use the BM25 similarity measure for their algorithm.

Resnik and Smith [5] use their structural filtering system STRAND which filters candidate parallel pairs by determining a set of pair-specific structural values from the underlying HTML page. They report a precision of 98% and a recall of 61% on their developed English-Chinese parallel corpus.

Fung and Cheung [6] present a method to extract parallel sentences from very non-parallel corpora by exploiting bootstrapping on top of IBM Model 4. They claim that their "find-one-get-more" strategy principle allows them to add more parallel sentences from dissimilar documents, to the baseline set. Primary steps of their method is alike the former approaches while they uses similarity metric same other approaches. Then they used an iterative bootstrapping framework based on the principle of "find-one-get-more", which claims that documents found to contain one pair of parallel sentences must contain others even if the documents are judged to be of low similarity. They rematch documents with using extracted sentence pairs, and refine the mining process iteratively until convergence (Fung and Cheung 2003). Adafre and Rijke [7] extract similar sentences from Wikipedia article pairs by considering that Wikipedia consists of documents with several languages. They investigated two approaches. First approach uses a machine translation to translate Wikipedia pages from source language to target language. The second approach, a bilingual lexicon is used to extract parallel sentences from Wikipedia aligned documents. Finally, word overlap between sentences is used as a similarity measure.

Munteanu and Marcu [8], first use a dictionary to translate some of the words of the source sentences, and then use these translations to query a database for finding matching translation candidates and extracting final parallel sentences. In other work, Munteanu and Marcu [9] train a maximum entropy classifier to extract parallel corpus in Arabic, English and French languages. They show that a good-quality MT system can be built from scratch by starting with a very small parallel corpus (100,000 words) and exploiting a large non-parallel corpus. Abdul-Rauf and Schwenk [10]-[11] present another technique similar to [8]'s method and use a statistical machine translation system instead of the bilingual dictionary. In their approach they used an IR system to find the best candidates from translated sentences. Moreover, they used well-known evaluation metrics *WER* (Word Error rate), *TER* (Translation Error Rate) and *TERp* (Translation Error Rate plus) to decide the degree of parallelism between candidate sentences. Diep et al. [12] present an unsupervised method and use statistical translation system to detect parallel French-Vietnamese parallel sentences with mining comparable corpora. An iterative process was implemented to increase the number of extracted parallel sentence pairs which improved the overall quality of the translation.

Authors of [13] investigate potentiality of Wikipedia as comparable corpus and propose a ranking model to extract parallel sentences from Wikipedia. They used features in different categories, such as features related to word alignments, features related to distortions, 3 features which peculiar to Wikipedia markup and word level induced lexicon features. They compared their proposed model with prior model and could improve the process of sentence pair extraction.

Patry and Langlais [14] introduced a system named PARADOCS for mining Wikipedia parallel documents

using several content based features. PARADOCS has three components: 1- Extracting Wikipedia target documents that are more similar to the a source document, 2- To classify the extracted documents as parallel or non-parallel, 3-filtered out wrongly selected parallel pairs. Stefanescu et al. [15] use information retrieval system in their approach in order to reduce the search space and memory. After creating an index structure for target language sentences, for each sentence in the source language the content words are selected and translated to the target language using an existing dictionary. Finally, top N similar sentences in target language are selected as the translation candidates.

Otero and Lopez [16] introduce an automatic method to build comparable corpora from Wikipedia using Categories as topic restrictions. Their strategy relies of the fact Wikipedia is a multilingual encyclopedia containing semi structured information.

Authors of [17] propose an unsupervised approach to automatically synthesize Wikipedia articles in multiple languages. Taking an existing high-quality version of any entry as content guideline, we extract keywords from it and use the translated keywords to query the monolingual web of the target language.

3. Extracting Persian-English Parallel Sentences

In this paper, the goal is to implement an effective method to extract Persian-English parallel sentences from Wikipedia documents. All original Wikipedia pages are filtered because they contain non textual elements like: images, hyperlinks and etc. Then all documents in Persian part are aligned to English part documents. With considering Doc_P and Doc_E, two aligned documents in Persian and English languages respectively, the main goal is to find a potential parallel sentence in Doc_E, for any selected sentence in Doc_p. Therefore, we need to calculate the similarity scores between each Persian sentence in Doc_p and all English sentences in Doc_E. Our proposed similarity score consists of these three parameters: Two bi-directional weights using Persian-English and English-Persian SMTs, and a penalty computed based on sentence lengths. Our overall score is represented in Equation 1:

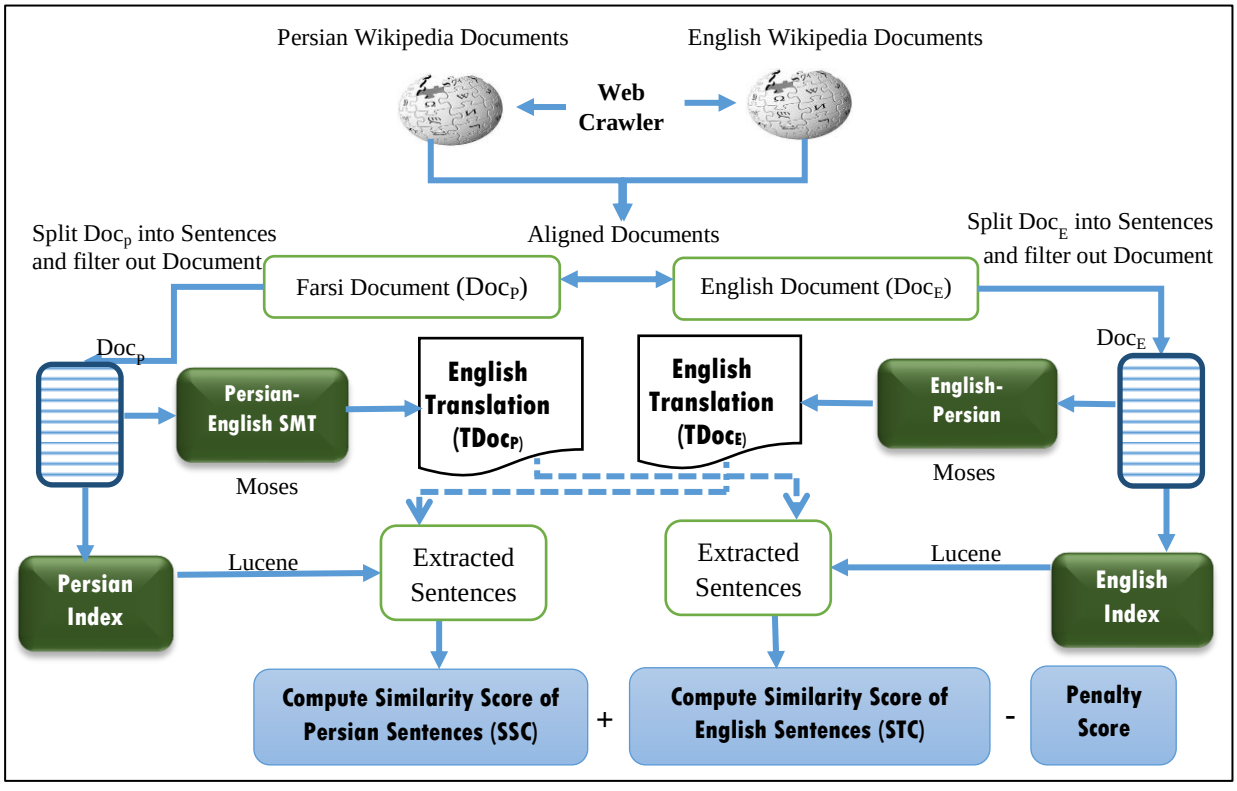

Figure 1: The process of Aligning and weighting sentences. In this figure, Doc_P and Doc_E are Persian and English documents and $Tdoc_P$ and TDoc$_E$ are Translations of documents Doc_P and Doc_E respectively

For any two sentences S_p and S_e selected from documents Doc_p in source language and Doc_p in target language respectively, we have defined our similarity metric as follows:

$$BiSim(S_P, S_E) = Sim(S_P, S_E) + Sim(S_E, S_P) - Penalty(S_P, S_E) \quad (1)$$

Where $Sim(S_S, S_T)$ is our similarity metric to calculate the similarity between S_S and S_T using a Source-Target SMT.

While two information retrieval systems are used in the similarity calculation phase (i.e. a Persian IR system and an English IR system), we need to translate our documents in both sides.

Figure 1 shows a demonstration of our approach. In next sections, our work is explained more precisely. In that figure and in the rest of this paper, for two aligned documents Persian and English, the following notations are used:

S_P A selected Persian sentence from Doc_p
TS_p Translation of sentence S_p
S_E A selected English sentence from Doc_E
TS_E Translation of S_E

$|S_P|$ and $|S_E|$ are the lengths of S_P and S_E respectively.

Full description of Figure 1 (i.e. our approach) is presented in the following sections.

3.1 Aligning Wikipedia Documents

We designed a web crawler to download Wikipedia documents in both Persian and English languages. Then all similar documents are aligned via inter-language link structure provided by Wikipedia. Aligned documents with major difference in number of sentences are discarded. We used the threshold 30% in our work. E.g. If the number of sentences in an English document Doc_E (i.e. $|Doc_E|$) and the number of sentences in Persian document Doc_P which is aligned to Doc_E (i.e. $|Doc_P|$) and $|Doc_E| > |Doc_P|$, we discard two corresponding documents Doc_E and Doc_P when

$$\frac{|Doc_E|}{|Doc_P|} < 30\%$$

For any two aligned documents which are passed by above filtering, the procedures of next sub-sections are applied to extract potential parallel sentences in order to create our new Persian-English parallel corpus.

3.2 Translation Phase

A Bi-directional Information Retrieval approach (i.e. a Persian IR and an English IR) is proposed to calculate the similarity scores and consequently to extract similar sentences. Considering two aligned documents Doc_P and Doc_E, a Persian document and its English aligned document, all sentences in both sides are translated: One translation to Persian-English translation and one English-Persian translation. Therefore, we need two different statistical machine translation systems. A Persian-English Machine Translation system and An English-Persian MT systems. The translation modules in this work are built using Moses toolkit [18] with the default setting and as follows:

- GIZA++ [19] was used for word alignments, the "-alignment" option for phrase extraction was "grow-diag-final-and"
- Fourteen features in total were used in the log-linear model: distortion probabilities (six features), one 3-gram language model probability, bidirectional translation probabilities (two features) and lexicon weights (two features), a phrase penalty, a word penalty and a distortion distance penalty.
- Two 3-gram models were created for both English and Persian languages. We built both language models using the SRILM toolkit [20] and our monolingual corpus.

Now, we have two new translated documents: $TDoc_p$, the translation of Persian document and $TDoc_E$, the translation of aligned English document to Persian. These translations will send to next section for further processing.

3.3 Bi-directional Score calculation using Information Retrieval systems

In this part the introduced bi-directional scores for two sentences S_P and S_E, sentences selected from Persian and English aligned documents Doc_P and Doc_E respectively, are calculated. This score is a summation of Persian-English score $Sim(S_P, S_E)$ and English-Persian score $Sim(S_E, S_P)$. In order to calculate these scores we use two implementations of Lucene IR system; An English and a Persian one:

1- Using sentences of $TDoc_p$ as the query database on indexed document Doc_E

In first step, Our IR machine creates an index database from Doc_E. Then each sentence in translated document $TDoc_p$ considers as an independent query, and the IR system calculates scores between it query and all sentences in index database (i.e. Doc_E).

2- Using sentences of $TDoc_E$ as the query database on index document Doc_P

The process of this IR system is similar to previous IR system.

Now, regarding above method, for any two sentences S_P and S_E, we have two similarity scores $Sim(S_s, S_t)$ and $Sim(S_s, S_t)$ and consequently a cumulative bi-directional score. Using this assumption, we could have the bi-directional cumulative score between any two sentences in Doc_P and Doc_E. Another issue should be considered in our final scoring system is the difference in the length of selected sentences. So we have applied a new penalty score which is increased when the difference in sentences' length is more than expected. This penalty is proposed in next sub section.

3.4 Length penalty

In order to compute the length penalty between any two sentences S_P of Doc_P and S_E from Doc_E, we use below Equation 2.

$$Penalty = \frac{\beta - \rho}{\lambda} \quad (2)$$

Where $\beta = \frac{|S_P|}{|S_E|}$ and $\rho = \frac{Avg_{Eng}}{Avg_{Per}}$

Avg_{Eng} is the average English sentence length and Avg_{Per} is the average for all Persian sentence.

The parameter λ is our tuning factor. In our experiments this factor is set to 2 which is gained experimentally based on our statistics. For each sentence in Source Document Doc_p and sentences in English document Doc_E which is aligned to Doc_p, in order to calculate the final score, this computed penalty will be added to our cumulative bi-directional score gained from previous part.

3.5 Final selection

Finally, for any Persian sentence S_P from document Doc_p, considering all sentences {S_{E1}, S_{E2}, S_{E3} … and S_{En}} from aligned document Doc_E, the potentially parallel translation is selected using Equation 3.

$$\text{argmax}_i(BiSim(S_p, S_{Ei})) \quad (3)$$

Then we use a selection threshold T in order to ignore those selected sentences with very low similarity score. I.e. based on Equation 4 all pairs with similarity score lower than T are filtered out:

$$BiSim(S_p, S_E) > T \quad (4)$$

4. Experiments and Results:

In this part, the results of our parallel extraction system are evaluated. In section 4.1 our dataset and primary SMT systems are introduced. In section 4.2 we evaluate the accuracy of our extracted parallel sentences and then in section 4.3 we evaluated the effect of applying our extracted data in four different Persian-English SMT systems.

4.1 Input datasets

We used Persian and English Wikipedia documents to create our initial comparable corpus. These documents are aligned based on our defined criteria introduced in Section 3-1. After this phase, we have 1200 aligned documents to extract potentially parallel sentences. As introduced in Section 3, two machine translation systems are needed. These translation systems are created using our setting which introduced in Section 3.2. We have used an available Persian-English parallel corpus to create our translation models and two monolingual Persian and English corpora to build our language models. The size of our Persian-English parallel corpus is about 105K Sentences. Using these systems, our filtered Wikipedia documents are translated to opposite language (i.e. English documents are translated to Persian and Persian documents are translated to English).

In order to create our test dataset we annotated 10 Wikipedia article pairs. We have considered two sentences as parallel, if they are in one of these three levels of parallelism: parallel sentence pairs, quasi parallel sentences and strong comparable sentences. Our implementation is applied on these 10 annotated documents.

4.2 Extracting parallel sentences

In this work, we have used different selection thresholds to determine whether two sentences are parallel or not. In each experiment, all sentence pairs with similarity score more than this threshold are considered as parallel sentence. The precision and recall metrics are used to evaluate our extraction method with different thresholds. Table 1 shows precision, recall and the number of extracted sentences using our test set (10 annotated documents introduced in Section 4-1) by applying different selection thresholds, 0.1, 0.2, 0.4, 0.6 and 0.8.

Table 1: Precision, Recall and the number of extracted sentences using different selection threshold

Threshold (0, 1)	Extracted sentences	Precision	Recall
0.1	289	49.12	96.5
0.2	241	62.65	91.51
0.4	142	78.16	67.27
0.6	103	90.29	56.36
0.8	73	94.29	42.81

Figure 2, shows a graphical demonstration of this experiment to compare precisions and recalls for threshold introduced in Table 1. As we expected, by decreasing the similarity score, i.e. stricter selection criteria, the recall value decreased and the accuracy of selected our selection becomes higher.

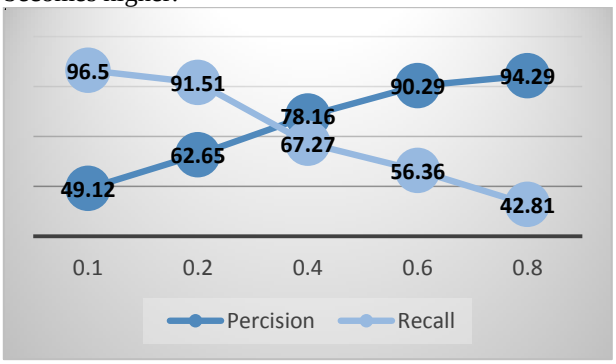

Figure 2: Precision, Recall and number of extracted sentences using different similarity threshold.

Figure 3 shows the compression of the accuracy of our bi-directional IR approach with two other strategies:
- A simple binary classifier introduced in Section 2.
- Using one direction translation and consequently a usual information retrieval system, instead of our proposed bi-directional system

For this experiment, the precision and recall values are shown for top *K* similar sentences selected using different approaches. The value of K, is 100, 200, 300, 400 and 500. Based on the results, for all values of K, our approach has more accuracy in comparison with classic binary classifier and the figure shows using penalty score and English-Persian translation could improve the results too.

4.3 Machine Translation Evaluation

To evaluate the effect of adding the extracted sentences to an existing machine translation system, we used a partition of TEP parallel corpus (Tehran university English Persian corpus) as our seed parallel data. Our selected corpus contains 100K parallel sentences. We create our baseline SMT system based on this corpus, besides; we created a test set by extracting parallel sentences from Wikipedia documents. To create a high quality test set we manually select 200 sentences. In addition of our baseline SMT system, we created five other SMT systems using extracted sentences and examine each SMT system with our test set individually.

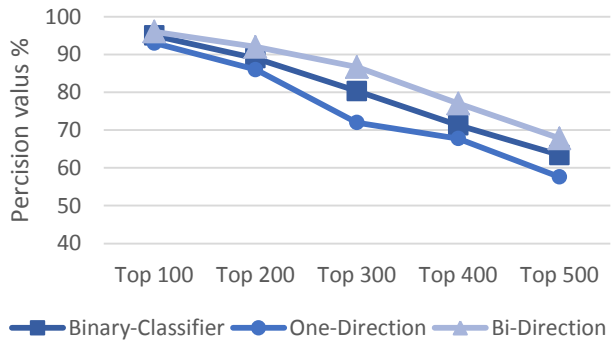

Figure 3: Precision of top 500 sentences in Bi-directional method (our proposed method), One-direction approach and classic method from 10 annotated Document.

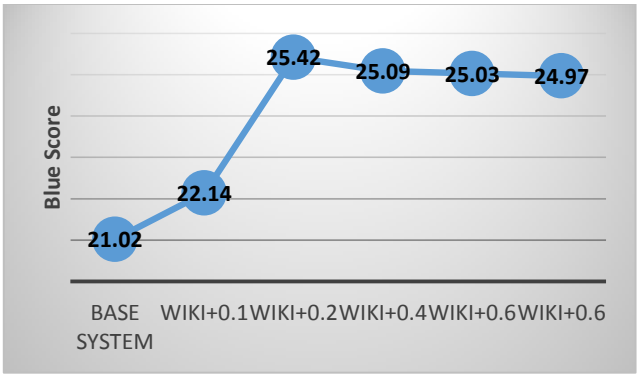

Figure 4: The BLEU score of SMT systems with adding extracted sentences with different selection threshold to baseline system

In the primary phase, we evaluated our test set with the baseline system. As it shown in Figure 4, the BLEU score of this system is 21.02. Afterwards we add Wikipedia extracted sentence pairs with different selection thresholds 0.1, 0.2, 0.4, 0.6 and 0.8. The BLEU scores are shown in Figure 4. This experiment shows adding extracted data to baseline system has a significant improvement in BLEU score of existing machine translation system. Based on results in Figure 4, the best improvement is when we added extracted data with selection threshold 0.4. Increasing in notice amount is the main reason of efficiency decrement when extracted sentences with higher thresholds (i.e. more than 0.2) are added to our trained SMT.

5. Conclusion

This paper focuses on extracting parallel Persian-English sentences from Wikipedia as a rich document level aligned comparable corpus. We introduce a bi-directional information retrieval and a translation based weighting system to extract Persian–English parallel sentences from document aligned level articles of Wikipedia. First, we train two Persian-English and English-Persian SMT systems to translate Persian and English Wikipedia documents respectively. Then, we performed our bi-directional information retrieval approach on the sets of translated bi-lingual comparable sentences to extract the candidate list of parallel sentences. Moreover a length penalty score is applied to our similarity scores. Using a ranking system, the sentence pairs with most cumulative score is selected as the parallel sentences.

The experiment results show that Wikipedia is a useful resource for extracting parallel data, even for low resource languages pairs. Applying the extracted sentences on the baseline statistical machine translation system has a large effect on translation accuracy and improves our Persian-English SMT. Based on the procedure of our approach, it also could be applied on other language pairs.

Acknowledgments

The authors gratefully acknowledge the contributions and helps of *Prof. M. Makhfif* to this work.

References

[1] Brown, P.F., S.A.D. Pietra, V.J.D. Pietra and R.L. Mercer, "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics. Vol. 19, no. 2, 1993.

[2] Koehn, P., F.J. Och and D. Marcu, "Statistical phrase-based translation", Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Vol. 1, 2003.

[3] Zhao B., S. Vogel, "Adaptive parallel sentences mining from Web bilingual news collection", International Conference on Data Mining. 2002.

[4] Utiyama M, Isahara H, Reliable measures for aligning Japanese-English news articles and sentences. In: 41st Annual meeting of the Association for Computational Linguistics, proceedings of the conference, Sapporo, Japan, pp 72–79, 2003

[5] Philip Resnik, Noah A. Smith. 2003. The Web as Parallel Corpus. Computational Linguistics, 29. 349 – 380

[6] Fung, P., P Cheung, "Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM", Conference on Empirical Methods on Natural Language Processing. 2004.

[7] Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar Sentences across Multiple languages in Wikipedia. Procesding of the 11th Conference of the European Chapter of the Association for Computational Linguistics. 62-69.

[8] Munteanu, Dragos Stefan and Daniel Marcu (2005) 'Improving Machine Translation Performance by Exploiting Non-Parallel Corpora.' Computationbal Linguistics, 31(4): 477-504.

[9] Munteanu, Dragos Stefan and Daniel Marcu (2006) 'Extracting parallel sub-sentential fragments from non-parallel corpora', Sydney, Australia, Association for Computational Linguistics: 81-88.

[10] Abdul-Rauf, Sadaf and Holger Schwenk (2009) 'On the use of comparable corpora to improve SMT performance', Athens, Greece, Association for Computational Linguistics: 16-23.

[11] Abdul-Rauf, Sadaf and Holger Schwenk (2011) 'Parallel sentence generation from comparable corpora for improved SMT.' Machine Translation, 25(4): 341-375.

[12] Diep, Do Thi Ngoc, Laurent Besacier and Eric Castelli (2010) 'Improved Vietnamese-French Parallel Corpus Mining Using English Language', Paris, France: 235-242.

[13] Jason R. Smith, Chris Quirk and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment. Annual Conference of North American Chapter of the ACL, Los Angles, California. 403-411.

[14] Alexandra Patry and Philippe Langlais. PARADOCS: A Language Independent Go-Between for Mating Parallel Documents, 2010

[15] Dan Stefanescu, Radu Ion, Sabine Hunsicker. 2012. Hybrid Parallel Sentence Mining from Comparable Corpora. Eurpean Association for Machine Translation (EAMT).

[16] Pablo Gamallo Otero, Isaac Gonzalez Lopez. 2010. Wikipedia as Multilingual Source of Comparable Corpora. Proceeding of the 3rd workshop on building and using comparable corpora, LREC. Malta. 2010.

[17] Chen Yuncong and Pascale Fung. Unsupervised Synthesis of Multilingual Wikipedia Articles. 2010. Proceeding of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing. 197-205.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar ,Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the ACL 2007 Demo and Poster Sessions. Association for Computational Linguistics. Prague, June 2007. 177–180.

[19] Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, March 2003. 19-51.

[20] A. Stolcke (2002), SRILM - An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, Denver. 901-904.

Ebrahim Ansari received his B.S. degree in Computer Science from Yazd University, Iran, in 2005, and his M.S. degree in Computer Engineering from the Engineering School of Shiraz University, Iran, in 2009. He is currently a Ph.D. degree student in the Computer Science and Engineering Department at Shiraz University, Iran. His research interests include: machine translation, association rule mining and distributed processing.

Mohammad Hadi Sadreddini is Associate Professor in the Computer Science and Engineering Department at Shiraz University, Iran. He received his B.S. degree in Computer Science in 1985, his M.S. degree in Information Technology in 1986 and a Ph.D. degree in Distributed Information Technology, in 1991, from Ulster University, in the UK. He has been working in Shiraz University, Iran, since 1993. His research interests include: association rule mining, bioinformatics, and machine translation.

Alireza Tabebordbar received his associated degree in Computer Science from Bahonar University, Iran in 2007, and B.S degree in Computer Science from Safahan Higher Education in 2009. He is currently M.Sc. Student in field of Computer Engineering at Computer Science and Engineering Department of Shiraz University, Iran. His research interests include: Natural Language Processing, Information Retrieval, and Data mining.

Richard Wallace received his BS from the University of Idaho, Moscow, Idaho (1980); his Masters in Computer Science from the University of Dayton, Dayton Ohio (1986), and is currently pursuing his PhD in Computer Architecture from Complutense University of Madrid, Madrid, Spain under Dr. José Luis Vázque-Poletti. He is an Advising Engineer for multiple aerospace companies in the Dayton, Ohio area. His work is focused on real-time, critical systems. He has worked on two IEEE standards VHDL (IEEE-STD-1076) and PSL (IEEE-STD-1850) and has been the author of over 30 technical reports and publications on the theory and application of large distributed event-based systems with an emphasis on hybrid computing architectures.