An Evaluation on Performance different metrics on extraction of Persian-English Parallel sentences

Marziyeh Homayouni[1], *[*]**Amin Keshavarzi**[2]
[1,2]Department of Computer engineering, Marvdasht branch, Islamic Azad University Marvdasht, Iran
*Corresponding author : keshavarzi@miau.ac.ir

Summery
Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Persain). MT systems are highly dependent on the amount of training data. Through past years, different methods have been proposed to extract parallel sentences from the web or available corpora. In this paper we have presented a method to create Persian-English comparable corpus from Wikipedia articles and extract parallel sentences from that. In order to create a Persian-English comparable corpus we have used WordNet to classify and extract similar articles in Wikipedia. Also we have evaluated the performance of different calssification algorithms in extracting Persian-English parallel sentences. Experimental results show the efficiency of the proposed approach in comparison with the other state of the art methods. This approach is language independent and it could be applied to other language pairs that have enough Wikipedia sources.

Keywords:
Parallel sentences, Comparable Corpus, Wikipedia, Information Retrieval, Statistical Machine Translation.

1 Introduction

In recent years, machine translation (MT) systems have obtained reasonable results when applied to languages such as English-French, English-Chinese, English-Germany; however, for many languages, especially for low resource languages, machine translation needs more parallel sentences to get a better results. There are two main challenges for low resource languages; difficulty of creating corpus for these languages, difficulty of implementing methods available for rich-resource languages for low resource languages (due to of difference in syntax and structure). Statistical machine translation (SMT) systems 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 applied to decode and find the best translation for each source language sentence [1,2,3].

Wikipedia is a rich resource, containing articles in a variety of domains. Wikipedia articles have useful characteristics such as links, interlanguage links and category tags, which are beneficial for Information Extraction (IE) task. The main propose of IE is processing unstructured data and converting them to the structured data. English Wikipedia, with more than 4 million articles, is the first language among the others in terms of article quantity and Persian with more than 300,000 articles is the 20th language in Wikipedia (at the time of writing this paper). However, Persian Wikipedia has not been investigated enough.

Different methods have been proposed to extract parallel sentences from Wikipedia [4]. However, they have extracted parallel sentences by aligning documents via interlanguage links between source and target languages. Also its not possible to get the similar results for low-resourced languages. Due to of differences in syntax and their structure. Therefore, extraction of parallel sentences for low resource languages such as Persian is more complicated. In this paper, first, an approach is proposed to cluster Wikipedia articles using WordNet. Then, we have extracted parallel sentecens from these clusters by applying different machine learning algorithms. Also we have performed an evaluation on performance of these algorithms in extraction of parallel sentences.

To create Persian-English parallel corpus, Wikipedia articles are extracted and filtered. Afterwards, WordNet is used to cluster similar articles in Wikipedia. Then the performance of different machine learning algorithms have been evaluated.

Our paper is structured as follows: In Section 2, the related works are reviewed. Afterwards, the procedure of creating comparable corpus and extracting parallel senteces are described in Section 3. Results are discussed in section 4 and Section 5 concludes the paper.

2 Related Work

Fung and Cheung present a method to extract parallel sentences from very non-parallel corpora by exploiting bootstrapping on top of IBM Model [5,6]. They claim that their “find-one-get-more” strategy principle allows them to add more parallel sentences from dissimilar documents, to the baseline set. Since they use similarity metric like other approaches, primary steps of their method are alike

Manuscript received July 5, 2016
Manuscript revised July 20, 2016

the former 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 by using extracted sentence pairs, and refine the mining process iteratively until convergence [5].

[7] , first used a dictionary to translate some of the words of the source sentences, and then used these translations to query a database for finding matching translation candidates and extracting final parallel sentences. In other work, [8] 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 (2011) present another technique similar to [4] 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 evaluate the degree of parallelism between candidate sentences [9].

[4] extract similar sentences from Wikipedia article pairs by considering that Wikipedia consists of documents with several languages. They investigated two approaches. First approach used a machine translation to translate Wikipedia documents from source language to target language. In the second approach, a bilingual lexicon was used to extract parallel sentences from Wikipedia aligned documents. Finally, word overlap between sentences was used as a similarity measure.

[11] investigate potentiality of Wikipedia as comparable corpus and train a classifier to detect parallel sentences. They extracted a large number of parallel sentences from Wikipedia aligned documents and used different features to model their sentences. They show that their extracted parallel sentences from Wikipedia could improve the performance of an SMT system [10].

[12] introduce an automatic method to build comparable corpora from Wikipedia using the categories as topic restrictions. Their strategy relies on the fact that Wikipedia is a multilingual encyclopedia containing semi structured information. They proposed three methods to create comparable corpora: First, a machine translation was used to translate the given subject to target language and extract articles in the source and target languages, which contained the given topic in category tag. In second and third methods, they used some criteria based on interlanguage links to increase the homogeneity of extracted comparable corpora.

[13] 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. Afterwards, top N similar sentences in the target language are selected as the translation candidates. Finally, candidate sentences are weighted based on some features and sentence pair with the highest score is selected as the potential parallel sentence.

3 Proposed method

Section 3.1 presents the WordNet based approach for creating comparable corpus from Wikipedia documents. Section 3.2 introduce our features that used for extraction of Persian-English parallel sentences. Section 3.3 evaluate the performance of different algoithms in extracting parallel sentences.

3.1 Creating Persian-English Comparable Corpus

This section presents the procedure of extracting articles from Wikipedia. In order to extract parallel sentences from Wikipedia, creating a comparable corpus for Persian-English languages is necessary [14]. Therefore, we have used WordNet to cluster similar Wikipedia articles Our proposed approach in creating comparable corpora from Wikipedia is applicable for all other Wikipedia language pairs that have enough Wikipedia resources.

In order to extract Persian-English Wikipedia documents a web crawler is designed to download the articles in both languages. Extracted articles were filtered as they contain non-textual elements like images. Afterwards, WordNet has used to claculate the similarity of articles.

Section 3.1.1 describes the process of creating similarity vectors. Section 3.1.2 presents the method of creating comparable corpora and clustering similar articles.

3.1.1 Clustering Articles

In order to cluster similar Wikipedia articles, In the first step, we define a subject to cluster the similar articles. Defined subject can be the name of place or a issue. After that, we have used WordNet to find the synomym of the defined subject and create a distinct list from the extracted subject. Finally we have used lucene as a information retreival tool to cluster the articles that have the highest similarity with the defined subject or one of it's synonum. Following describe the steps needs to cluster the articles related to the "music"

1-Extract Wikipedia articles from www.wikipedia.org/en
2- Use WordNet to find the synonym of music
3-create a list of selected subject and it's synonym

4- Use lucene to calculate the similarity of articles with created list (we have used Tf-idf similarity metrics to calculate the similarity of articles)
5-cluster articles with the highest measure of similarity.
6- follow the similar procedure for other subject
in order to calculate the similarity of articles we have used the C# version lucene.

3.1.2 Clustering and creating comparable corpus

In order to cluster similar Persian articles and create comparable corpora for Persian-English languages, we have used the Inter-Language-Links of English clustered articles. Therefore, the interlanguage link of English clustered articles haved used to create Persian counterparts.

3.2 Extracting Parallel Sentences

In pervious section, similar Persian-English articles were clustered and a comparable corpus was created for each Persian-English aligned cluster. In this section, the goal is to implement an effective method to extract Persian-English parallel sentences from Wikipedia aligned clusters. With considering CP and CE two aligned Clusters in Persian and English languages respectively, the main goal is to find a potential parallel sentence in CE, for any selected sentence in Cp. Therefore, calculation of the similarity scores between each Persian sentence in Cp and English candidate sentences in CE is essential. In the rest of this paper, for two aligned Persian and English clusters, the following notations are used:

- CP and CE: English Clusters respectively
- CTP : Translation of Cluster CP
- SP: A selected Persian sentence from Cp
- STp: Translation of sentence Sp

To extract parallel sentences from Wikipedia two features are defined; Translation Similarity and Word alignments. Word alignments features are based on IBM alignment models (Brown et al. 1993). Translation similarity is based on the similarity of key words between Persian-English sentences.

3.2.1 Features

The main problem in machine translation tasks for under resource languages is the lack of parallel data. Although for Persian language there are some readily available parallel corpora, [15, 16] the problem is the limited the size of these corpora and the low quality of the sentences they have. Therefore, we have used both translation similarity and word alignment similarity features to extract Persian and English sentences. In order to implement translation similarity feature an SMT system is needed for translating the texts from Persian to English. The translation modules in this work is built using Moses toolkit [17] with the default setting and is as follows:
- GIZA++ [18] 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. Both language models were built using the SRILM toolkit [19] and our monolingual corpus.
Using our machine translation system (i.e. Persian-English machine translation systems) with above configuration, all Persian clusters are translated to English language. Therefore, a new translated cluster is generated: CTp, the translation of Persian cluster. The process of calculating the translation similarity score is as follows:

3.2.2 Translation Similarity Score f1 (SP, SE)

In order to calculate the similarity of each Persian sentence with English ones in aligned clusters we have used lucene.
First, all English sentences (SE) in CE were indexed by the IR system. Then each Persian sentence in translated Cluster CTp is considered as an independent query (only non-functional words were kept), and the IR system extracts and calculates scores between the query and top sentences in index database.

$$f_1(S_P, S_E) = Sim(S_p, S_E) \quad (1)$$

At this point, for each Persian-English sentence we have extracted a similarity score, which extracted from our implemented Information Retrieval machine.

3.2.3 Word Alignments Features f2 (SP, SE)

In this section, in order to improve the process of Persian-English Parallel sentence extraction, we have used another feature in addition to the above described score. Therefore we have used GIZA++ to train a model for calculating the alignment similarity of sentence pairs. Word alignments of sentences in both directions are examined (Persian-English and English-Persian). In the first step, we added the candidate sentences to the end of our training corpus and trained a model. We have used GIZA++ for creating our model. Those sentence pairs which have more similarity would have a higher probability and sentences which doesnt have anything in common or have low similarty would get lower probability.

3.3 Model Creation

In order to extract similar Persian-English sentences we have to assign the optimal weight of each feature (translation similarity and word alignment). Therefore we trained four classifiers and examined the efficiency of each model in assigning optimal weight to our features. Thus, 300 Persian-English sentence pairs are extracted manually, which contains 150 parallel sentences and 150 non-parallel sentences as a training set. In the first step we created a logistic regression classifier. The classifier assigns 0.35 percent of weights to the word alignment similarity and 0.65 percent of weights to the translation similarity feature. In the second step we trained a model using Chi-Squre metric. The model assigns 0.40 percent of weights to the word alignment similarity feature and 0.60 percent of weights to the translation similarity feature. Afterwards, a linear regression classifier was trained, which assigns 0.42 percent of weights to the word alignment similarity and 0.58 percent of weights to the translation similarity feature. Finally a classifier was trained using gain ratio metric. The model assigns 0.49 percent of weights to the word alignment similarity feature and 0.51 percent of weights to the translation similarity feature.

Up to now, we have created four different models and assigned the optimal weights of each feature. In the next section we are going to evaluate the performance of created models in extracting parallel Persian-English sentence pairs.

4 Experimental Results

To evaluate the accuracy of proposed clustering approach, the precision of clusters was examined. Also, In order to evaluate the performance of created models, we extracted the 400 sentences in each model. As presented in Figure 1 logistic regression model has the highest accuracy among the other models. Therefore logistic regression model selected for further processing an evaluation.

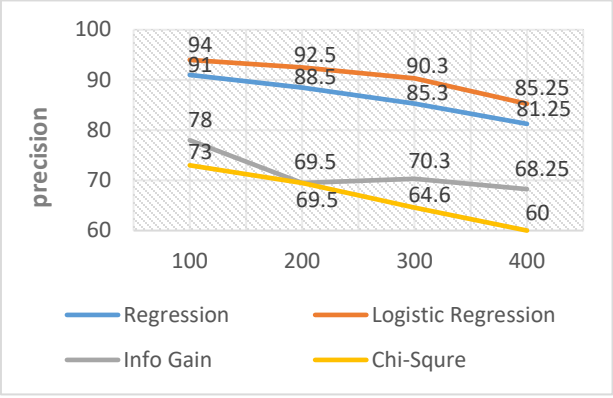

Fig.1 The accuracy achieved by the four selective model

4.1 Document clustering accuracy

The accuracy of clustering method was evaluated by computing the precision of extracted clusters. Therefore, 10 clusters were annotated and the precision of each cluster was examined manually. As Figure 2 shows, the proposed clustering approach, which was based on WordNet similarity. One advantage of our proposed method in comparison with other classic approaches is the non-dependency of this method to any specific language resources such as external dictionary.

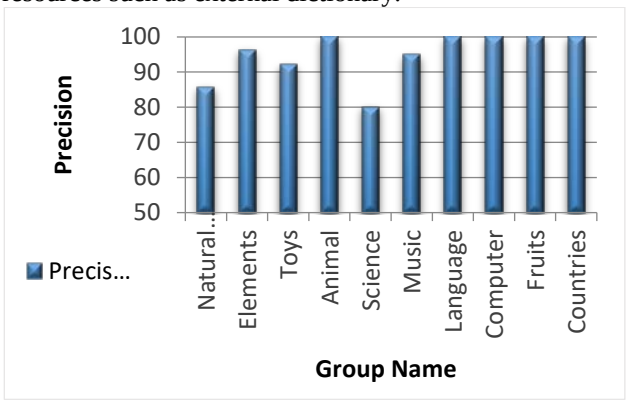

Fig.2 The accuracy of clustering method

4.2 Evaluating Regression Algorithm Deeply

As is presented in Fig.1 , Logistic Regression Algorithm has better performance in assigning the optimal weights to our features. Therefore we evaluated the performance of this algorithm more deeply. In the first step, we have selected four clusters and annotated the top 300 sentences in each of them. Afterwards, we have evaluated the precision of annotated sentence in each cluster. Figure3 demonstrated the results obtained from each cluster.

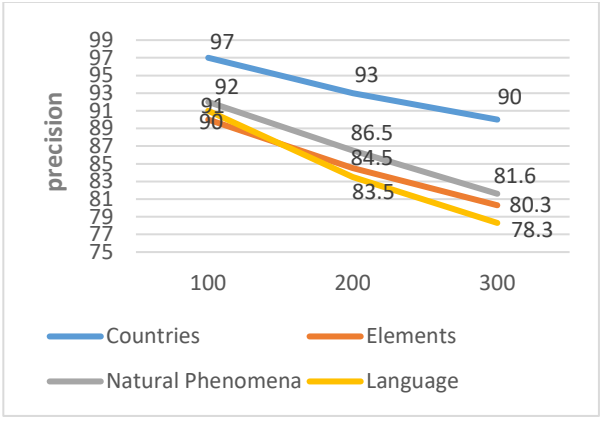

Fig.3 Accuracy different clusters in Extraction of parallel sentences

4.3 Machine Translation Evaluation

In order to evaluate the impact of adding extracted sentences to an existing machine translation system, result corpus was added to a baseline system. Eventually, 15778 parallel sentences were selected from different clusters and a test set with 400 sentences manually extracted from Wikipedia documents. An SMT system used with the default configuration explained in section 3.2.2 As a parallel corpus, we have used available parallel corpora (TEP and Mizan). Based on the results represented in table 1, extracted sentences could improve the performance of an SMT system about 6.6 BLEU score.

Table 1 Machine Translation Evaluation

SMT Evaluation	*Bleu Score*
Base line System	16.50
Base line System+ Wiki Extracted Sentences	23.10
Number of Extracted Sentences	15778
Number of Sentences In Test Set	400

5. Conclusion

This paper focused on extracting parallel Persian-English sentences from Wikipedia. A cluster level approach was introduced bu using WordNet to classify Wikipedia articles. Moreover, two set of features have used to increase the Performance of extracting Persian-English parallel sentences. First, Wikipedia articles were clustered based on WordNet similarity. Afterwards, sentences in each cluster were weighted based on features related to word alignments and translation. After wards we have evaluated the performance of different models in assigning the optimal weights to our feature. The experiment results showedthat our proposed method is enough accurate in extracting parallel sentences. Applying the extracted sentences on the baseline statistical machine translation system has a large effect on translation accuracy and improves the Persian-English SMT. Based on the approach procedures, our proposed method could be applied to other language pairs.

Reference

[1] Brown, P. F. , S. A. D. Pietra, V. J. D. Pietra and R. L. Mercer. (1993) The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19: 2.

[2] Stolcke. A. (2002) SRILM - An Extensible Language Modeling Toolkit. roc. Intl. Conf. on Spoken Language Processing 2: 901-904.

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

[4] Adafre, S. F and de Rijke, M. (2006) Finding similar Sentences across Multiple languages in Wikipedia. Proceeding of the 11th Conference of the European Chapter of the Association for Computational Linguistics:62-69.

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

[6] Och. F, Ney. H (2003) A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1): 19-51.

[7] Munteanu, D. S and Marcu, D. (2005) Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics 31:477-504.

[8] Munteanu, D. S and Marcu, D. (2006) Extracting parallel sub-sentential fragments from non-parallel corpora. Association for Computational Linguistics,81-88.

[9] Koehn. P, Hoang. H, Birch. A, Callison-Burch. C, Federico. M, Bertoldi. N , Cowan. B, Shen. W, Moran. C, Zens. R, Dyer. R, Bojar. O ,Constantin. A, Herbst. E. (2007) Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the ACL- Demo and Poster Sessions. Association for Computational Linguistics: 177-180.

[10] Torsten Zesch and Christof Müller and Iryna Gurevych (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech,

[11] Smith, R. Quirk, C, Toutanova, K. (2010) Extracting Parallel Sentences from Compar ble Corpora Using Document Level Alignment. Annual Conference of North American Chapter of the ACL 25(4) 403-411.

[12] Otero, P. and Lopez, I. (2010) Wikipedia as Multi-lingual Source of Comparable Corpora. Proceeding of the 3rd workshop on building and using comparable corpora:21-25.

[13] Stefanescu. D, Ion. R, Hunsicker. S. (2012) Hybrid Parallel Sentence Mining from Comparable Corpora. European Association for Machine Translation (EAMT).

[14] Homa Baradaran Hashemi, Azadeh Shakery, and Heshaam Faili (2010). Creating a Persian-English Comparable Corpus, Lecture Notes in Computer Science Volume 6360, 2010, pp 27-39

[15] Homa Baradaran Hashemi, AzadehShakery (2014) . Mining a Persian–English comparable corpus for cross-language information retrieval. Information Processing & Management,Volume 50, Issue 2, Pages 384–398

[16] AzadehShakery, Zahra Rahimi (2011). Topic Based Creation of a Persian-English Comparable Corpus. Lecture Notes in Computer Science Volume 7097, pp 458-469.

[17] Yuncong, P and Fung, P. (2010) Unsupervised Synthesis of Multi-lingual Wikipedia Articles Proceeding of the 23rd International Conference on Computational Linguistics (Coling) :197-205.

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

[19] Mohammad Taher Pilevar, Heshaam Faili, Abdol Hamid Pilevar. (2011) TEP: Tehran English-Persian Parallel Corpus. Lecture Notes in Computer Science Volume 6609, pp 68-79.