IJICTR International Journal of Information & Communication Technology Research

ITRC

Volume 5- Number 3- Summer 2013

Evaluating Fidelity of Persian-English Sentence-Aligned Parallel Corpus

Masoomeh Mashayekhi
Computer Engineering Department
Iran University of Science and Technology
Tehran, Iran
M_mashayekhi@com.iust.ac.ir
ma_mashayekhi@yahoo.com

Morteza Analoui
Computer Engineering Department
Iran University of Science and Technology
Tehran, Iran
anaLoui@iust.ac.ir

Received: January 27, 2013- Accepted: June 29, 2013

***Abstract*— Bilingual corpus is one of the most important resources for Natural Language Processing applications and researches. The quality of bilingual corpora can influence the result of researches that used it as a resource. When translation machine is used to verify corpus quality, the quality of translation machine can affect the evaluation of corpus. One way for evaluating software or resources in ISO is verifying its own features. The expectation of finding translation for each word in each sentence by using a bilingual dictionary is verified in this paper as a factor for evaluating fidelity of corpus. Computing this expectation needed a pre-processing step that is designed with considering the differences between English and Persian languages. This method is a combination of a rule-based method with the information of a dictionary.**

Keywords- Evaluation, Bilingual Corpora, Aligned Corpora, Parallel Corpora, Bilingual dictionary

I. INTRODUCTION

Sentence-level aligned parallel corpora have several applications of high importance, especially in NLP, like cross-language information retrieval [1], machine translation [2], lexicography [3], and language learning. In modern linguistics, a corpus can be defined as a body of naturally occurring language or languages. It should be noted that computer corpora are rarely haphazard collections of textual material. They are generally assembled with particular purposes in mind, and are often assembled to be representative of some language or text type [4]. During The last decade, numerous projects are defined and fulfilled aimed to build up as lager parallel bilingual corpora as possible. Usually, one side of such projects is English language. Chinese-English [5], French-English [6], Hungarian-English [7], Swedish-English [1] are most frequent examples.

Before ARCADE project started, there was no formal evaluation exercise for parallel text alignment. And worse still, there was no multilingual aligned reference corpus to serve as a "gold standard" or any established methodology for the evaluation of alignment systems [8].

ARCADE project used a gold standard to measure recall, precision and F-measure for evaluating parallel alignment. Sentence is the unit of granularity used for computation of recall and precision. In ARCADE II, recall, precision and F-measure were computed only at

the character level [9]. Other papers also used these measures to evaluate alignment [10].

One way of evaluating a bilingual corpus is using it in SMT and evaluating its output. Bin LU setup an SMT system and tested translation in both directions. [11] The BLEU scores are shown for evaluating the corpus. Both training without optimizing parameters using minimal error-rate and training with parameters optimization of minimal error-rate is implemented. Measuring the complexity and variety of corpus is another way to evaluate. Correctness of the corpus word frequency distribution is also checked. [12] These measurements are used for a monolingual Arabic corpus.

Diana Santos showed several simple principles for evaluating a monolingual Portuguese corpus resource. [13] Some of these evaluating forms are sentence separation, extraneous characters and so on. These measures are evaluated manually. Bo Li and Eric Gaussier found translated word pairs in the corpus. [14] They used this measure to evaluate the quality of some corpora and introduced an algorithm for improving the quality of corpus.

In this paper, a strategy is developed to evaluate the quality of English-Persian parallel corpus before using it in SMT. Since there is no gold standard corpus, we can't find recall, precision and F-measure for our corpus. So, the percentage of words in each sentence that have translation in the alignment pair is computed and reported to evaluate the parallel corpus quality of faithfulness.

II. The English-Persian Parallel Corpus

An English-Persian parallel corpus is used to perform a pre-use evaluation on it. This corpus is a 1-million sentence corpus that is aligned in sentence level. The corpus is constructed manually. The texts used for this corpus are from classic literature. Some novels and their translations such as Anna Karnina, David Copperfield and Don Quixote are aligned to build an English-Persian corpus. The corpus is in XML format.

III. Word Alignment of a Sentence Pair

Bilingual dictionaries are an essential resource in many multilingual natural language processing (NLP) tasks such as machine translation [15] and cross-language information retrieval (CLIR). Let`s have the following primitive definitions in an English-Persian sentence-aligned parallel corpus:

$$C= \{A_i \mid A_i \text{ is an alignment between an English and a Persian sentence}\} \quad (1)$$

$$A= \{(E, F)\mid F \text{ is the translation of } E\} \quad (2)$$

Where E is an English sentence and F is its Persian translation. C is the whole corpus. We can define E and F as follow:

$$E= \{e_i: \text{an English word}\mid 0<i<n_e\} \quad (3)$$

$$F= \{f_i : \text{a Persian word} \mid 0<i<n_f\} \quad (4)$$

e_i represents the i^{th} English word and f_i has the same definition correspondingly. n_e is the number of words in an English sentence and n_f is the number of words in a Persian sentence. For each *A* alignment of corpus, Mef can be defined on the basis of the expectation of finding the translation for each English word ei in the English sentence E in the Persian sentence F. Let θ be a function indicating whether a translation from the vocabulary T_e of e_i in the dictionary D is found in the sentence F. θ is defined as follows:

$$T_{ei}= \{ t \mid t \text{ is every phrase in } D(e_i) \} \quad (5)$$

$$\theta(\omega,F)= 1 \quad \text{iff } T(\omega) \cap F \neq \emptyset \quad (6)$$

$$\theta(\omega,F)= 0 \quad \text{elsewhere} \quad (7)$$

M_{ef} is then defined as:

$$M_{ef}(E,F)=\text{Expectation}(\theta(\omega,F)\mid\omega\in E)= \sum\theta(\omega,F)\Pr(\omega\in E)$$

$$\text{for every } \omega\in E \quad (8)$$

$$M_{ef}(E,F)= |E|/|E \cap D_e| \sum \theta(\omega,F)\Pr(\omega \in E) \quad \text{for every } \omega \in E \cap D_e \quad [14] \quad (9)$$

Where D_e is the English part of an independent bilingual dictionary, and where the last equality is based on the fact that the corpus and the bilingual dictionary are independent from one another, the probability of finding the translation in F of a word ω is the same for ω in E∩De. Given the natural language text, our evaluation will show that the simple presence/absence of a criterion can perform very well. This leads to $\Pr(\omega\in E)=1/|E|$, and finally to:

$$M_{ef}=1/|E\cap D_e|\sum\theta(\omega,F) \quad \text{for every } \omega \in E \cap D_e \quad (10)$$

M_{ef} is the percentage of English words translated in Persian part of corpus

IV. Pre-Processing

The differences between English and Persian language and the properties of the bilingual dictionary forced us to apply a pre-processing to English and Persian parts of corpus. These differences can be named as follow:

A. *Abbreviations*

All the abbreviations in the English sentences should be converted to the main form. The English abbreviations probably come with apostrophe.

B. *Plural form*

The plural form of the nouns should be converted to single form. So, plural 's' should be omitted and all irregular plural forms should be converted to single.

C. *Irregular verbs*

In bilingual dictionaries, there is no definition for other forms of irregular verbs except the present form. So, for using dictionary, irregular verbs should be converted to present form.

D. *Determiner*

There isn't any word in Persian as determiner. So there is no translation for 'the' in sentence pairs. 'the' is omitted in pre-processing step. Although we can align 'the' with 'آن' or 'این', in most of the sentences there is no alignment for this word. Since 'the' is not an important word in sentences, omitting it does not affect our evaluating. "a" and "an" are aligned with

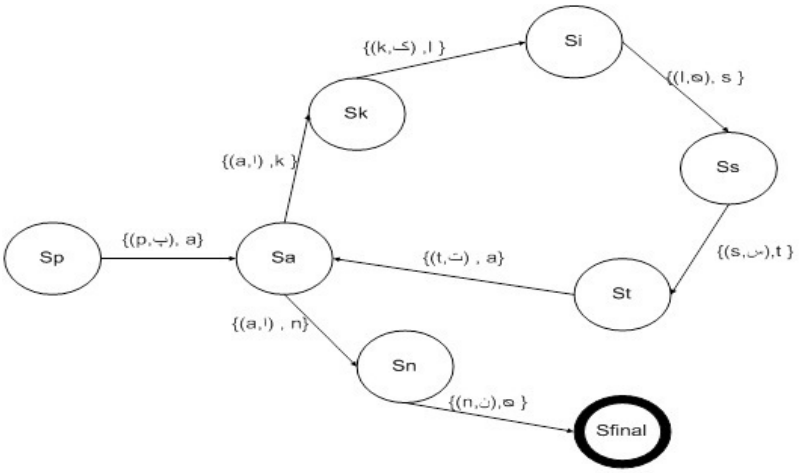

Figure 1. The process of matching two words in finite state automaton

"ی" at the end of the word or "یک","یکی" before the word in Persian. . If there isn't any of such words, we omit them and don`t calculate them in percentage. The word 'the' sometimes changes the meaning of its next word. In this implementation, as we do not use the meaning of words considering its pervious word and we do not use the meaning of the words in the context, we can't calculate the meaning of the word after 'the' in the context. So we just try to find, "یک","یکی","آن' or 'این' for aligning the word 'the'.

E. Present form

The 'ing' and 'ed' should be omitted from the verbs to have the present form. So some lemmatization is performed to the sentences.

F. Pronouns

In most of Persian sentences, Pronouns are dropped and the identifier of verb is the subject. This rule is named as "prodrop". So pronouns should be verified separately.

G. Persian verbs

In all English-Persian dictionaries, the Persian verbs are in infinitive forms. So, they should be converted to infinitive form.

All of these changes are recorded. Because these changes are performed without having POS-tags and for example, it may be possible to omit 'ing' from a noun. In this case, the change is receded.

V. FINDING TRANSLATION FOR EACH ENGLISH WORD

The translation of some English words in Persian is more than one word and these words can be far from each other in Persian sentences and may have distortion. So finding the translation of each word should be performed to whole Persian sentence.

Arian pour dictionary (A very famous English-Persian dictionary) is used as a resource. It has about 50300 English words with many useful phrases and their translations.

In all bilingual dictionaries, there is no entry for Proper nouns. A finite state automaton is used to find proper nouns in sentence pair.

A. Proper nouns

The finite state automaton works like a transliteration machine. An English word and a Persian word that is probably its equal are defined as inputs of the FSA (finite state automaton). The transition conditions of this automaton consist of a pair and a character. The pair includes an English character and the Persian character that is its equal. The input is like this:

$$\{(a,b),y\} \quad (11)$$

Where *a* is the English character, *b* is the Persian character and *y* is the next character of English word. The process of finding word 'پاکستان' for its English word 'Pakistan' is shown in fig.1.

All steps for finding aligned words are done in an application that gives a pair from the corpus, tokenizes English sentence as words, finds translation of the word in dictionary, chunks the translation in words, and finds all words of translation in order to match two words in English and Persian sentences. This process is done until 3-gram, it means that if there is no match for an English word, it is combined with next word in English sentence and this combination is verified in the dictionary. After checking all English words with dictionary, the proper noun step is performed. Pronouns, 'to be' verbs, plural 's', and some propositions are checked separately and by a rule-based method. Finally, the average is computed for whole corpus.

There is an example that shows the result of this application:

English sentence: *I remember the precise moment, crouching behind a crumbling mud wall, peeking into the alley near the frozen creek.*

Persian sentence:

دقیقا آن لحظه یادم مانده؛ پشت چینه مخروب‌های دولا شده بودم و کوچه کنار نهر یخزده را دید می‌زدم

Remember	یاد ماند	Peek	دید زد
Precise	*دقیقا*	*Alley*	*کوچه*
Moment	*لحظه*	*Near*	*کنار*
Crouch	*دولا شد*	*Freeze*	*یخزده*
Behind	*پشت*	*Creek*	*نهر*
Crumble	مخروب	I	*م (شناسه فعل)*
mud wall	چینه		

So the percentage of matching words in this pair is 94.73%.

Fig. 2 shows another example in the application that is written for this purpose.

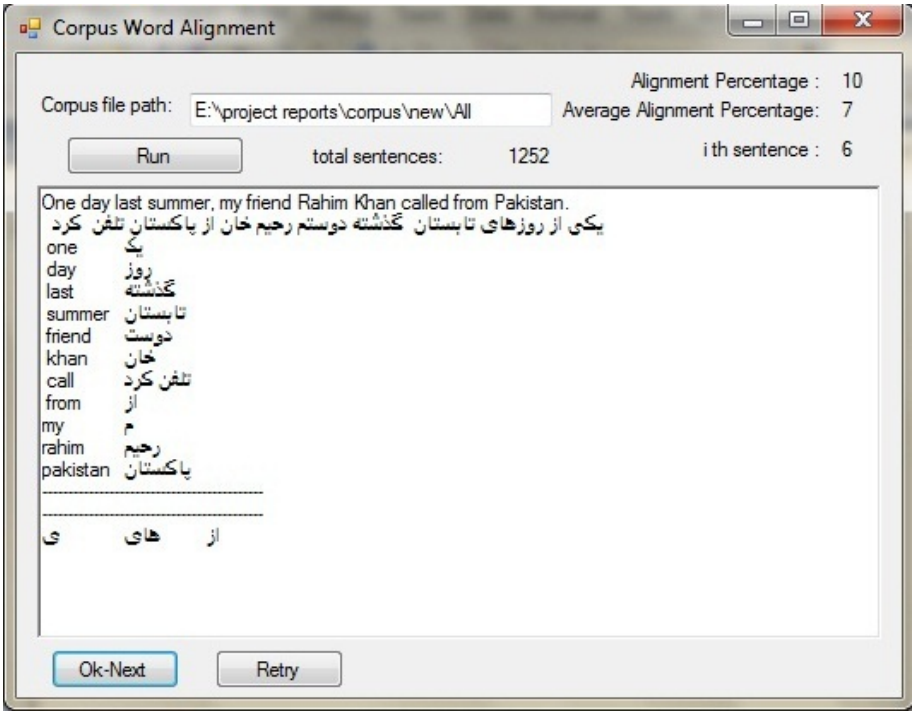

Figure 2 An Example of word alignment of a pair sentences in the written Application

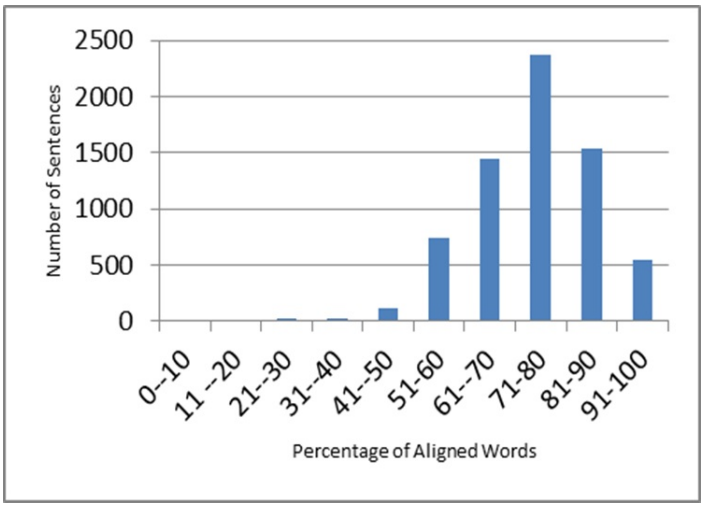

Figure 3. Distribution of sentences by Percentage of aligned words

VI. VERIFYING THE APPLICATION

This method is applied to a part of corpus and is manually checked. There are 6854 sentences in this part of corpus. Sentences are categorized in classes by their percentage of aligned words. Numbers and sentence percentages in each class are shown in Fig. 3.

This result shows that most of sentences are in the class of 70 and 80 percentage aligned words.

Finally, the whole corpus is checked by aligned words percentages. Our corpus is separated in parts based on the books. The result is shown with separation of books. Fig. 4 shows the results of 25 parts of corpus. The whole corpus has 68 parts.

VII. CONCLUSION

This method is a way to evaluate fidelity of a sentence-aligned bilingual corpus. We used this method to evaluate an English-Persian parallel corpus that is built from novels. This method is a combination of a rule-based and a data-based method; because it used a bilingual dictionary as data for alignment and also some rules to coordinate sentences with the entry of bilingual dictionary. These Rules are designed considering the differences between English and Persian languages. Some lemmatization changes are applied to the input sentences.

A finite state automaton is designed for finding corresponding proper nouns in bilingual sentence pairs. This way is like a transliteration but in a basic way, because there is a little amount of proper nouns in a bilingual sentence pair.

This word alignment method can be used as a secondary method for word alignment in a corpus. Of course we can use this method to align sentences of two languages. For example, for bilingual comparative texts, we can find the percentage of aligned words with this method and find most related sentences and offer them as an aligned pair of a parallel corpus.

The result shows that books that their main language is not English have less average of aligned words' percentage. So we can conclude that it is not a logical way to construct a bilingual corpus that used translation of other language texts for both part of it.

The percentage of aligned words for books that tell a story is more than books that their content is not a story. For example the book "Create your own future" is a book about time management and as Fig. 4 shows, it has less than 50% of aligned words. When we check the translation of this book manually and we found out that it has short sentences and the translator has tried to translate the concept of sentences, not the words.

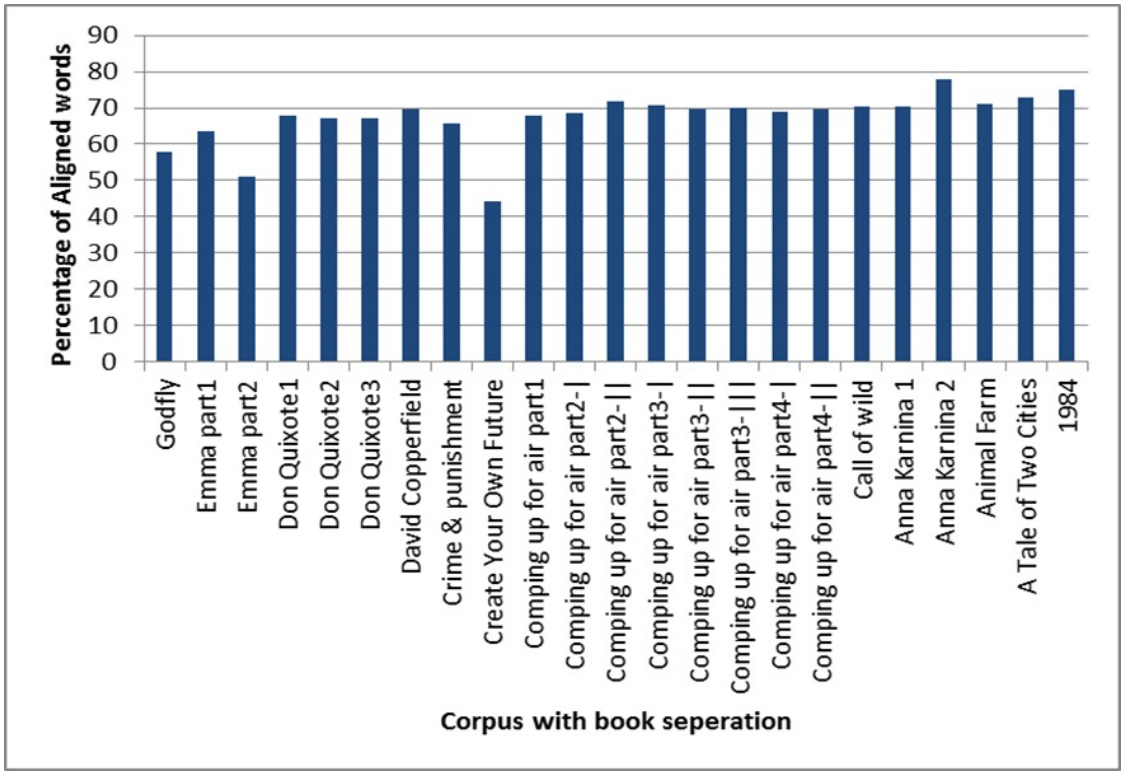

Figure 4. Result of implementing method to whole corpus
(Average of percentage of aligned words in sentences with separation of books)

REFERENCES

[1] T. Talvensaari, J. Laurikkala, K. J¨arvelin, M. Juhola, and H. Keskustalo, "Creating and exploiting a comparable corpus in cross-language information retrieval," ACM Transactions on Information Systems (TOIS), February 2007.

[2] M. Guid`ere, "Toward corpus-based machine translation for standard," Translation Journal, vol. 6, 2002.

[3] R. Krishnamurthy, "Corpus-driven lexicography," International Journal of Lexicography - Oxford Univ Press ,vol. 21, p. 231–242, July 2008.

[4] Leech, G. "Corpora and theories of linguistic performance", in Svartvik, J. Directions in Corpus Linguistics, pp 105-22. Berlin: Mouton de Gruyter, 1992.

[5] L. Sun, S. Xue, W. Qu, X. Wang, and Y. Sun, "Constructing of a large-scale chinese-english parallel corpus," COLING '02: Proceedings of the 3rd workshop on Asian language resources and international standardization, pp. 1-8, 2002.

[6] S.F. Chen, "Aligning sentences in bilingual corpora using lexical information," Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 1-8, 2002.

[7] K. T´oth, R. Farkas, and A. Kocsor, Sentence alignment of hungarian-english parallel corpora using a hybrid algorithm, vol. 18, Acta Cybern, 2008, p. 463–478.

[8] P. Langlais, M. Simard, J. Véronis, S. Armstrong,P. Bonhomme, F. Debilil, P. Isabelle, E. Souissi, and P. Théron, "ARCADE: A cooperative research project on parallel text alignment evaluation," in Proceedings of the First International Conference on Language Resources and Evaluation, Grenada,Spain, 1998.

[9] Y. C. Chiao, O. Kraif , D. Laurent , T. Nguyen , and Semmar, "Evaluation of multilingual text alignment systems: the ARCADE II project," 2006..

[10] Philippe Langlais, Michel Simard , and Jean Véronis, "Methods and practical issues in evaluating alignment techniques," in Proceedings of the 17th international conference on Computational linguistics, Montreal,Quebec, Canada, August 10-14, 1998.

[11] Bin LU, Tao JIANG,Kapo CHOW, and Benjamin K. TSOU, "Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT".

[12] Y. Benajiba, and P.Rosso, "Towards a Measure for Arabic Corpora Quality," 2007.

[13] D. Santos, and P. Rocha, "Evaluating CETEMPúblico, a free resource for Portuguese," in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,ACL '01, Morristown, NJ, USA, 2001.

[14] Bo Li, and eric Gaussier, "Improving corpus comparability for bilingual lexicon extraction from comparable corpora," in Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 2010.

[15] Franz Josef Och, and Hermann Ney, "A systematic comparison of various statistical alignment models," Computational linguistics, vol. 29, pp. 19-51, 2003.

Masoomeh Mashayekhi received her B.Sc. degree in Hardware Engineering from K.N.Toosi University of Technology, Tehran and her M.Sc. degree in Artificial Intelligence from Iran University of Science and Technology, Tehran. Her current research interests include Natural Language Processing, Machine Translation and Transliteration. In addition, she has some experience in Evaluating NLP Resources.

Morteza Analoui is an associate professor of Computer Engineering in Iran University of Science and Technology, Narmak, Tehran, Iran. His academic and industrial research works are in the area of "virtual networks", "AI and cognitive science", "Optimization" and "NLP".