International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Sentence Alignment for English Urdu Language Pair

Syed Abdul Basit Andrabi , Abdul Wahid

Abstract:Sentence aligned parallel text is an important resource in statistical machine translation; therefore, Sentence alignment is a crucial part of machine translation.The sentence alignment task comprises of recognizing the correspondence between words sentences and paragraphs of the source and target languages. Different researchers proposed several sentence alignment algorithms for aligning sentences of the source and target language. In this paper, we have explored sentence alignment algorithms based on character length , word length and lexical matching and carry out performance analysis of Gale and Church Algorithm on English and low resource language Urdu.

Index Terms*: *Low resource languages,sentence alignment, parallel corpus ,statistical machine translation, parallel corpus

I. INTRODUCTION

Statistical machine translation is a challenging task for language pairs for which low resources are available and where the word ordering of sentences is also different. The languages such as English and Urdu are different word order languages, the word order of English language is subject-verb-object (SVO) whereas the word ordering of Urdu language is Subject-object-verb (SOV).The first task of statistical machine translation is a collection of bilingual corpora, then their sentence alignment. Sentence alignment is an essential and crucial task in machine translation. Sentence aligned parallel text is an essential resource for NLP (Natural Language Processing) tasks which include multilingual information extraction, text processing, machine translation and sentimental analysis. The sentence alignment task comprises mapping sentences, paragraphs and words from a source language to the corresponding target language. So that they can be used in translation. As we know that the manual alignment of sentences is time consuming, expensive and we cannot generate a good amount of the corpus through manual sentence alignment, so it is necessary to have some automated means of sentence alignment methods. A sentence alignment algorithm for practical use should be (1) Fast enough to process large corpora,(2) Robust to deal with noise commonly present in the real data and (3) It should be less error-prone [1] . The main sentence alignment approaches are of two types one is based on length of characters and word in sentence and other is based on lexical matching approach the other one can be a hybrid of both length and lexical based. Length based approach depends on the length of characters or words in the sentence. Although a huge amount of bilingual text is available on the web for different language pairs since Urdu is resource-poor language, therefore, a vast amount of parallel corpora are not available. In this paper, we have collected some religious texts available in UMC005 corpus of both language pairs for sentence alignment. Several algorithms are proposed for sentence alignment task which are either based on length or lexical matching some of them are mentioned in this paper.

Revised Manuscript Received on 22 May 2019.
***** Correspondence Author
Syed Abdul Basit Andrabi*, Department of CS & IT, Maulana Azad National Urdu University Hyderabad, India.
Abdul Wahid, Department of CS & IT, Maulana Azad National Urdu University Hyderabad, India.

II. CORPORA CLASSIFICATION

The term corpus is derived from a Latin word corpus which means body. In modern times it represent collection of texts of various languages, dialect or subset of the language used for linguistic (grammatical) analysis.It can also be defined as a group of machine-readable texts or more closely a definite collection of machine-readable content which is sampled so that it can be a representative of whole language . The corpus can be classified into various types but in field of computer science we are only interested in the classification of monolingual and multilingual classification [2] , [3] . The classification of corpus is shown in figure 1.

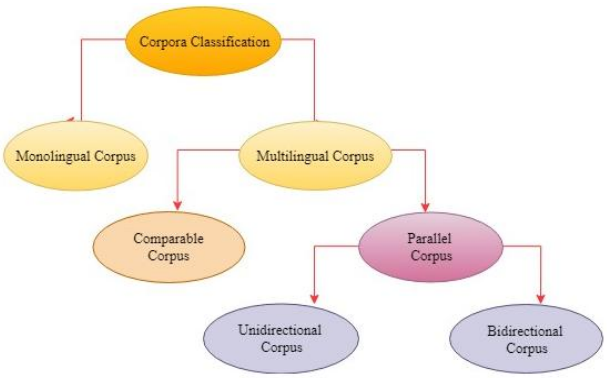

Figure 1. Classification of corpora

Monolingual Corpora: These are the texts which are in one single language. Multilingual Corpora: These are the texts which consist of multiple languages and can further be divided into following types:
A.

Parallel corpus: A parallel corpus comprises of set of aligned sentences that are translations of each other. The parallel corpus represents texts which are in different languages in which one text represents the source and other one represents the translation. These texts are also called as bi-texts. The parallel corpus can be further subdivided into two categories: one is unidirectional corpus and other is bidirectional corpus.
B.
C. Comparable Corpus: Comparable corpus is set of text that share the same idea , but the way of representation is different. For example news broadcast of different channels has some same content , but they present it in different way.
Bilingual corpus is an important and essential resource for statistical machine translation.Little work has been done on the development of parallel corpus for English Urdu language pair. Parallel corpus of some religious texts are available. We have used UMC005 [4] corpus in this paper. UMC0005 is a parallel corpus of English and Urdu language. It actually has four sources Quran, Bible,Penn, Emille corpus and Treebank. From this corpus Quran and Bible are available for free of charge for research , educational and nonprofit use. However, for Emile corpus, we have to take permission from ELRA (European Language Resource Association).

III. Background of Urdu Language

The word Urdu is derived from theTurkish word Urdu which means Army camp Urdu is an asian low resource language that is mainly spoken in the Asian Subcontinent like Pakistan and in some states of India and Bangladesh. Urdu is a national language of Pakistan and is an official language in some states of India like Jammu and Kashmir , Telengana, Delhi,and Andra Pradesh etc. There are a vast number of speakers of Urdu language it is spoken by more than 100 million speakers in more than 20 countries across the world [5]. Due to digital resource scarcity not enough work has been done for Urdu. Urdu is a free order language and was developed between 6th and 13th century. Urdu vocabulary is derived mainly from three languages which are Arabic, Persian and Sanskrit making its syntactic structure complex. The Urdu in written in Nastaliq style from Right to Left same as Arabic and there are no small and capital words . For example if we have a sentence “I bought a new house today “ in English in Urdu we can write it in different forms as shown below

میں نے آج نیا مکان خریدا or آج میں نے نیا مکان خریدا

IV. Sentence Alignment Methods

As already mentioned sentence alignment approaches can be classified based on sentence length , word correspondence, and hybrid where more than one approaches are combined. The various approaches mentioned in this paper are shown in figure 2.

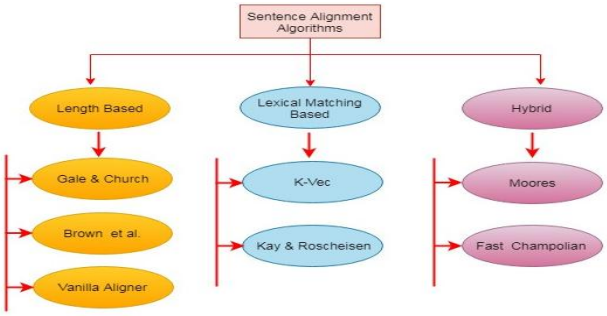

Figure 2 .Classification of sentence alignment methods

The length based approach depends upon the number of characters,the number of words, sentence position, sentence,length ratio in two languages co-occurrence frequency etc.. The lexical approach depends on already existing knowledge such as synonym’s , antonym’s and translation of words.
Gale and Church Algorithm: This algorithm is based on a statistical model. The main idea of this algorithm is that longer sentences in one language are usually translated into longer sentences in the other language and shorter sentences are usually translated into corresponding shorter sentences [6]. This algorithm counts the number of characters in the source and target language and align them according to sentence lengths.
Brown . et al also proposed sentence alignment algorithm that works at word level this is same as gale and church algorithm where aligned sentence is dynamically chosen using dynamic programming concepts.Brown algorithm makes use of statistical technique i:e it is based on the word length for aligning sentences in parallel corpora. In this algorithm the number of tokens were counted in both the language pair and no attention was paid to lexical information. The advantage of this algorithm is that it can be used for very large collections of texts as it is fast. This algorithm was first uses on English -French Haenard corpus and it was claimed that it achieved an accuracy of 99%. The algorithm was inspired by the fact if we give two sentences to the human and ask him to align them , it is natural for him to look up into words, so this algorithm also counts words [7] .
K-Vec Agorithm [8]. It is lexical matching based algorithm where the source and target text is splited into k pieces, then the word for which we are looking is checked in each of the k pieces. Here the distribution of words in each k points is seen In this way we can get the translation of each word. This algorithm was mainly developed for the languages which are not alike, as it is independent on sentence boundary.
Vanilla aligner [9]. It is an extension of gale and church algorithm presented by pernella et al in 1997. This algorithm uses the same approach as gale and church with a little difference that is this algorithm works with bi-texts in SGML (Standard General Mark up Language) which has a standard structure making it easy to identify sentence boundaries by using end line characters. The BR line break can also be used to mark end of the sentence as its function is to terminate a line of text.

Text translation alignment proposed by Kay and Roscheisen [10] which is based on a lexical matching approach ,where the main aim is to find similar word distribution. The words which are translations of one another have similar word distributions in the given languages. In this paper the results is in pairs of words , sentences and paragraphs. The problem with this approach is that its efficiency gets degraded when it is applied on large corpora.

Fast and accurate sentence alignment algorithm [11] proposed by Moore is based on composite approach that means it is using both length as well as lexical approach. The algorithm first uses modified Brown algorithm then using a lexical approach to find corresponding sentence pairs. This algorithm does not uses anchor points that were used by earlier approaches. The Moore,s approach has a problem if the data is sparse. Fast Champllion is a sentence alignment algorithm proposed by Peng Li et al. [1] , this algorithm is based on a hybrid approach using both sentence length and lexical matching. This algorithm first uses the length based approach to split the input bilingual text into small fragments that are already aligned and then align these fragments one by one, thereby reducing the running time complexity of algorithm , thus making it time efficient. In 2014 Hai-Long Trieu et al. [12] improved Moore's sentence alignment method. In this algorithm they used the technique of word clustering approach to overcome the weakness of Moore,s approach that happen when the data is sparse.

V. PERFORMANCE ANALYSIS

The Performance of sentence alignment algorithms depend on several factors like corpus size , noise in corpus that is sentences which are not translations of each other which may be due to the data extraction from different sources the another factor can be the linguistic difference due to which the algorithm may not perform well for that language pair ,the other is syntactical structure The above mentioned sentence alignment algorithms are applied for English German and English French languages but have not been applied to English Urdu language pair, as the structure and morphology of Urdu language is different as compared to French and German. From the syntactic structure Urdu is free order language and follows (SOV) subject object verb order where as English, French and German are (SVO) Subject verb object order. Therefore the performance of the algorithms may vary. Sentence alignment is crucial part of machine translation therefore it is important that the best algorithm is selected for sentence alignment otherwise the statistical machine translation system will be less reliable. Python script was used to calculate the number of words in both languages with utf-8 encoding. The description of corpus used is shown in table 1

Table 1. Description of corpus used

Language	Number of Sentences	Number of words
English	7400	192565
Urdu	7400	186167

we have used LF aligner that uses Gale and Church algorithm to align sentences of English and Urdu language. We executed many times by changing the number of sentences to get the error rate in sentence alignmnet .The results are shown in the graph show in figure 3.

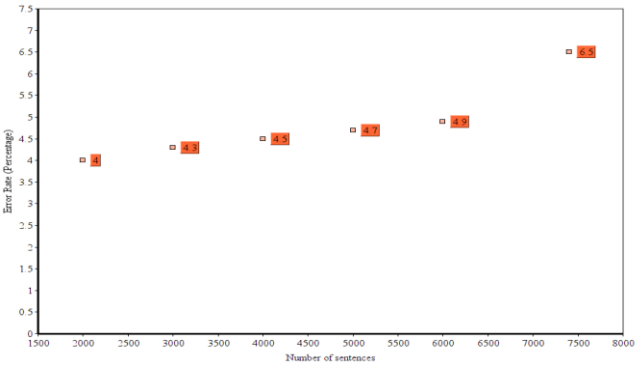

Figure 3.Performance of Gale and Church algorithm

VI. DISCUSSIONS

In this paper we have mentioned some algorithms based on statistical information like length , lexical information and hybrid approach.The Gale and church algorithm and Brown algorithm were the first algorithms which were based on statistical information. The gale and church was based on character length of sentence and Brown was based on word length of sentence .These algorithms perform well for European languages but do not have good accuracy for low resource languages.. We tried to use Gale and Church algorithm for English to Urdu language pair to find the error rate and it was found that the error rate varies with the increase in number of sentences .There are various other tools available like Nova text aligner , ABBYY aligner[13] , Garguntau [14] , Champollian aligner [15] and Nova text aligner [16] These tools did not have support for Urdu language. The method of sentence we mentioned were length based and lexical matching based , The length based sentence alignments methods are fast but are not so much robust as no lexical information is used , on the other hand lexical based alignment algorithms are robust as they use lexical information from source and target texts but they are slower as they require high computation cost. From it we conclude that we should go for hybrid approach to utilize the advantage of both approaches

VII. CONCLUSION

In this paper we have mentioned several sentence alignment methods based on different approaches like statistical approach and lexical approach .These algorithms perform well for the languages that share common features , but they do not show a good performance for the languages that are syntactically different like English and Urdu. We used Gale and church algorithm for English and Urdu languages to check the performance and it was found that the error rate increases with the increase in number sentences . Keeping in view the resource poor language these algorithms need to be enhanced to achieve good results.

It was also found that we should go for the hybrid sentence alignmnet approach so that we can utilize the benefits of both approaches.

REFERENCES

1. Peng Li, Maosong Sun, and Ping Xue. Fast-champollion: a fast and robust sentence alignment algorithm. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 710–718. Association for Computational Linguistics, 2010.
2. Tony McEnery and XIAO Zhonghua. Parallel and comparable corpora. Corpus-Based Perspectives in Linguistics, 6:131, 2007.
3. Sulema Torres-Ramos and Raymundo E Garay-Quezada. A survey on statistical-based parallel corpus alignment. Research in Computing Science, 90:57–76, 2015
4. Bushra Jawaid and Daniel Zeman. Word-order issues in English-to-Urdustatistical machine translation. 2011.
5. Sarmad Hussain. Resources for urdu language processing. In Proceedings of the 6th workshop on Asian Language Resources, 2008.
6. William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. Computational linguistics, 19(1):75–102,1993.
7. Peter F Brown, Jennifer C Lai, and Robert L Mercer. Aligning sentences in parallel corpora. In Proceedings of the 29th annual meeting on Association for Computational Linguistics, pages 169–176. Association for Computational Linguistics, 1991.
8. Pascale Fung and Kenneth Ward Church. K-vec: A new approach for aligning parallel texts. In Proceedings of the 15th conference on Computational linguistics-Volume 2, pages 1096–1102. Association for Computational Linguistics, 1994.
9. Danielsson, Pernilla, and Daniel Ridings. "Practical presentation of a“vanilla” aligner." In TELRI Workshop in alignment and exploitation of texts, February. 1997..
10. Martin Kay and Martin Röscheisen. Text-translation alignment. Computational linguistics, 19(1):121–142, 1993.
11. Robert C Moore. Fast and accurate sentence alignment of bilingual corpora. In Conference of the Association for Machine Translation in the Americas, pages 135–144. Springer, 2002.
12. Hai-Long Trieu, Phuong-Thai Nguyen, and Kim-Anh Nguyen. Improving moores sentence alignment method using bilingual word clustering. In Knowledge and Systems Engineering, pages 149–160. Springer, 2014
13. ABBY Sentence aligner accessed on 04 April 2019 url https://abbyy-aligner.soft112.com/download.html
14. Garguntau sentence aligner accessed on 04 April 2019 url https://sourceforge.net/projects/gargantua/
15. Champollian aligner accessed on 04 April 2019 url http://champollion.sourceforge.net/
16. Nova text aligner accessed 04 April 2019 url https://nova-text-aligner.soft112.com/