International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 4, April - 2014

Comparison of Various Bilingual Sentence Alignment Methods for Parallel Corpora Development

Sneha. D. L
M.Tech, SCSE
VIT University, Chennai Campus

G. Bharadwaja Kumar
Associate Professor, SCSE
VIT University, Chennai Campus

***Abstract*— Natural Language Processing is the field of AI, Linguistics and Computer Science mainly concerned with the Human (languages) Computer Interactions. Machine translation takes the input in one language and converts it into other language by preserving its meaning. The existing literary works in more than one language are useful resources for many practical applications, such as translation studies, language learning aids, writing and data-based machine translation. These bilingual works need to be aligned as precisely as possible in order to be of any use in parallel corpora development, which is a notoriously difficult task.**

In this paper, the performance of various sentence alignment approaches are compared through experiments conducted on manually built Telugu and Kannada parallel corpora. A few enhancements on the existing methods for effective and intelligent sentence alignment process for building parallel corpora are suggested based on the learning and experimentation results.

***Keywords*— NLP; Computational Linguistics; Machine Translation; Sentence Alignment; Parallel Corpora; Telugu; Kannada.**

I. INTRODUCTION

Recently, there has been extensive research going in development of bilingual corpora. A parallel corpus is a collection of texts in different languages where one of them is the original text and the other is their translations. The first step to enrich parallel corpora is to enhance the parallelism between units on both texts. This process is called "alignment". Alignment can be done at different levels, from characters, words, segments, sentences and paragraphs.

Aligning bilingual corpora has been proved to be useful in many ways including sense disambiguation, bilingual lexicography and machine translation. The input for sentence alignment is a pair of texts or documents and output is the sentences which are aligned. The main task here is to identify the correspondences between the sentences from two languages. Alignment usually starts with a pre-aligned parallel text such as speaker information, time stamp, chapters, paragraphs etc. Sentence alignment in parallel corpora development has been proved to be very useful in machine learning strategies.

In the present scenario, the challenging task in the field of Natural Language Processing is machine translation which takes the input in one language and converts it into other language by preserving its meaning. This paper throws light on the importance of Machine translation in present context and evaluates the performance of sentence alignment methods using the manually built parallel bilingual corpora for aligning the sentences in Telugu and Kannada Languages.

I. RELATED WORKS

An initial sentence alignment method was suggested by Brown [1] for aligning sentences in Parallel Corpora. Soon after which Gale and Church [2] proposed Simple probabilistic model in order to calculate the Probability score for each sentence pair based on sentence-length. Bob Moore [3] designed the Bilingual Sentence Aligner with an idea to combine sentence length based methods that are based on the distribution of length variable with word correspondence. Later, Fast-Champollion which is a fast and robust sentence alignment algorithm that employs a combination of both length-based and lexicon-based algorithms by optimizing the process of splitting the input bilingual texts into small fragments was developed by Peng Li and Maosong Sun [5]. It is an enhancement to Champollion algorithm proposed by Xiaoyi Ma [4]. Champollion considers a match possible only when lexical matches are present. It assigns higher weight to less frequent words, which are considered a stronger indication that two segments are a match. Champollion [4] achieved high precision and recall on manually aligned Chinese-English parallel text corpus. Fast-Champollion [5] considered reducing the running time by first splitting the bilingual input texts into small aligned fragments and then further aligning them one by one. Fast-Champollion [5] is fast, robust and effective enough for practical use, especially for aligning large amount of bilingual texts or long bilingual texts.

II. CHALLENGES IN MACHINE TRANSLATION

A sentence alignment algorithm should ideally be highly accurate, fast and require no special knowledge about the corpus or the two languages. The task of alignment for many languages is difficult and highly challenging because of the following reasons:

- For some languages, Sentences might be split, merged, deleted or inserted and it is difficult to use the statistical analysis of sentence lengths.
- Morphology of different languages is different and there are substantial additions and deletions that can occur on both sides.
- Stemming in vast corpus is another challenge and a lot of sentences align many to many which make the task more difficult.
- Another challenge is dealing with changes of sentence order: i.e., segments A and B in one language corresponding to segments B' A' in the other language.
- Non-1:1 alignments and incomplete translations make the Machine Translation even more challenging task.

III. BACKGROUND

In this section, the data we used for the experimentation is described. We manually extracted the data from the news magazine "Web Duniya". The magazine is released in two languages - Telugu and Kannada. There are substantial additions and deletion of texts in either of the issues of the magazine and sentence lengths of both the languages are not proportional which makes it difficult to use statistical analysis to do the alignment. For training, the data is manually aligned from Telugu to Kannada language using the Quill pad [9]. Quill pad is the Number One predictive transliteration tool for inputting Indian languages. Launched in 2006, Quill pad is the first Indic transliteration solution to use statistical machine learning method for intelligently converting user entered free-style phonetic input to its accurate representation in a chosen Indian language.

Text in a source language and the corresponding text in a target language are given to the various alignment systems. First, all the source sentences and the target sentences are chunked into smaller units based on the language specific chunkers. After chunking, the alignment of sentences takes place. We evaluate the performance of the methods by measuring the Precision and Recall parameters.

IV. SENTENCE ALIGNMENT METHODS

Basically the Sentence Alignment Approaches can be classified as Length-based methods, Lexical methods and Combined Methods.

In Length-based method, sentences that correspond to each other are similar in length. In Lexical methods, the corresponding sentences contain more corresponding words. Whereas, the combined methods use lexical cues in length-based settings.

We compared the performance of the following unsupervised methods.

1) Gale & Church Algorithm
2) Bilingual Sentence Aligner (Moore)
3) Hun Align Algorithm (ver. ga et al)
4) Giza++

Gale-Church's Length Based model defines a distance based on the costs of aligning source to target sentences. Probability score for each sentence pair based on sentence-length. Shorter regions of text tend to have shorter translations and longer regions of text tend to have longer translations. Although this method is extremely simple, it is also quite accurate and language-independent.

Bilingual Sentence Aligner was developed my Moore at Microsoft research labs. The idea is to combine sentence length based method with word correspondence. Sentence length methods are based on the distribution of length variable and the method uses Poisson distributions, rather than Gaussians, so that no hidden parameters need to be iteratively re-estimated and eliminates the need for anchor points. Training of the IBM Model-1 during runtime uses alignments obtained from the first pass. This method assumes that the larger corpus size the more effective. This method is a modification to IBM Translation Model 1, which eliminates rare words and low probability translations to reduce the size of the model by 90% or more. The word-correspondence-based model reduces the search space.

Hun align algorithm takes the input which is tokenized. The output for Hun align algorithm is bisentences, which is a sequence of sentence pairs from both languages. Hun align also produces many-to-one and one-to-many alignment links, which are needed to ensure that all the input sentences appear in the final alignment. But, Hun align method is not designed to handle corpora of over 20k sentence scopes by splitting larger corpora, this causes worse dictionary estimates and Hun align does not deal with changes of sentence order: i.e., In few languages segments A and B corresponds to segments B' A' in the other language.

GIZA++ is a program for learning statistical translation models from bitext. It Implements full IBM-4 alignment model with a dependency of word classes as described in (Brown et al) [1]. IBM-5 with features such as: dependency on word classes and smoothing are implemented. HMM alignment model is also implemented with features such as: Forward-Backward algorithm, Baum-Welch training, empty word, and transfer to fertility models, dependency on word classes, GIZA++ also implements a variant of the IBM-3 and IBM-4 models which allow the training of -p0 Smoothing for fertility, distortion/alignment parameters which yields in significant more efficient training of the fertility models and correct implementation of pegging as described in (Brown et al) [1].

V. EXPERIMENTATION AND RESULTS

Task:

– *To align Telugu sentences to its corresponding Kannada sentence.*

Dataset:

– *Size of Telugu Corpus: 17KB*

– *Size of Kannada Corpus: 18KB*

– *Total 180 sentences to be aligned.*

Challenges:

– *Non-1:1 alignments*

– *Insertions*

– *Improper Alignments*

– *Deletions*

– *Sentence Reordering*

– *Incorrect translations*

We used the F-score parameter which is defined in terms of precision and recall for measuring the performance of the sentence alignment methods. Precision is defined as the fraction of retrieved documents that are relevant to the search whereas as recall is the fraction of the documents that are relevant to the query that are successfully retrieved and F-score is the harmonic mean of precision and recall. Table 1 indicates the precision, recall and f-score results for varying mean and variance values for Gale and Church Alignment Method.

Mean	Variance	Precision	Recall	F-Score
1	1.0	0.1494	0.2798	0.2798
1	3.0	0.5230	0.6113	0.5637
0.7	**6.8**	**0.5775**	**0.6757**	**0.6228**
1	7.0	0.5666	0.6690	0.6144
1	9.0	0.5718	0.6757	0.6194

TABLE I. Gale-Church Alignment Results

From Table 1, it is observed that using Gale and Church algorithm the f-score values change with varying mean and variance. The optimal results are found with mean value of 0.7 and variance value of 6.8.

Table 2 indicates the Moore Alignment Method results. It is observed that the precision and f-score values are similar to the values measured using the Gale and Church Method. But, the time taken for alignment is much less in Moore method when compared to the Gale and Church Method.

Algorithm	Time Taken in seconds
Forward pass time	0.27
Backward pass time	0.03
total time	0.3

TABLE II. MOORE ALIGNMENT RESULTS

Table 2 shows the results in terms of time measured in seconds for three iterations using the forward pass and backward pass of forward-backward algorithm. In comparison, the time taken is 3 times lesser than the previous method (0.89 s). Since, there is no use of EM and hidden parameter evaluation and the reduced search space makes this model practical faster.

Following are the results evaluated using the Statistical Machine Translator GIZA++ and Hun Align Algorithm.

Alignment Method	Score
GIZA++(SMT)	0.8756
Hun Align	0.4076

TABLE III. Alignment Results Using Giza And Hun Align Methods

We can see that the overall result of the experimental evaluation has been that an improved alignment quality yields an improved subjective quality of the statistical machine translation system as well. While these algorithms perform sentence alignments effectively, there are few defects in the present existing systems. Following are a few drawbacks of the sentence alignment models:

- Hun Align algorithm doesn't work well for huge corpus (over 20k) and it needs a dictionary for aligning sentences. Thus, the precision and recall will drop as the size of the dictionary decreases.
- Gale and Church method uses dynamic programming approach which makes it difficult for huge corpus as the search space for alignment increases.
- Hun aligns does not deal with changes of sentence order: i.e., segments X and Y in one language corresponding to segments Y' X' in the other language.
- Following are few sentences wrongly aligned using the above methods.

Telugu sentences	Wrongly aligned Kannada sentence
మా కోడలిది చాలా జాలిగుండోయ్	ನನ್ನ ಸೊಸೆದು ಬಹಳ jaligundoy ಆಗಿದೆ
ఈ క్రూర విధి మనల్ని ఇంత త్వరగా విడదీస్తుందని నేననుకోలేదు	ಇದು ತಕ್ಷಣ ಕಾಡು vidadistundani nenanukoledu ನಮ್ಮ ಕರ್ತವ್ಯ
కత్రినా లడాయికి ఖాన్‌లిద్దరూ ఫుల్ స్టాప్... వాటేసుకుని స్నేహం...!!	ನಿಲ್ಲಿಸಿ ಫುಲ್ ಕತ್ರಿನಾ ladayiki khanliddaru ... Vatesukuni ಸ್ನೇಹ ...!
నీ బొజ్జ గుర్తొస్తుంది మావయ్యా అని మళ్లీ తడుముకోకుండా అన్నాడు చింటూ	ನಿಮ್ಮ ಹೊಟ್ಟೆ gurtostundi ಚಿಂಟು ಚಿಕ್ಕಪ್ಪ ಹೇಳಿದರು ಎಂದು ಮರು tadumukokunda

TABLE IV. INCORRECT ALIGNMENT EXAMPLES

Based on our experimentation and learning we propose the following methods for effective and intelligent sentence alignment process in building the parallel corpora:

- **Bootstrapping Technique:** In this technique, the algorithm learns how to align sentences from training dataset and aligns sentences on its own.
- **Case Marking Technique:** Generate highly accurate parallel corpora manually. Using this corpus, compile a set of rules or grammar based on the commonly observed sequence or case markers during alignment. Use this as benchmark for intelligent alignment process.
- **Ranking Technique:** For a given set of sentences that could be the possible alignment matches, a rank or score can be calculated. Based on these ranks, the alignment algorithm picks the target sentence which has the highest score to find the perfect match.
-

Telugu Sentence	Kannada Sentence	Quality Score
ఎలా చెప్పగలవు?	**ಹೇಳಲು ಹೇಗೆ?**	0.139
ఎలా చెప్పగలవు?	**ಏನು ಹೀಳಲಿ?**	0.209
ఎలా చెప్పగలవు?	**ಹೇಗೆ ಹೇಳ್ತೀಯ?**	**0.213**

TABLE V. Ranking Technique

Table 4 demonstarted the proposed ranking technique with a simple example. The scores are calculated using the Hun Alignment method, since the third row indicates the highest score (0.213), we choose the third row as the correct sentence alignment match for the given example.

VI. CONCLUSION

The above mentioned sentence alignment methods can be used in various fields such as:

- Teaching second languages and Terminology Studies
- Automatic Translation
- Multilingual Information Retrieval
- To build bitext database
- In Contrastive linguistics
- Translation studies - EFLClassroom
- Lexicology
- Speech Recognition and Parsing.

Further lines of research include:

- Increase the size of the existing bilingual Corpus (Telugu-Kannada) and evaluate the performance with different kinds of inputs.
- Developing a Hybrid Model that does Statistical Machine Translation for more than one language.
- Develop the algorithm based on our proposed enhancements.

Thus, the paper discussed in detail various unsupervised sentence alignment models. We have performed various experiments to assess the effect of different alignment models by implementing the four algorithms using the manually built bilingual parallel Corpus for Telugu and Kannada languages. This paper also suggested a few enhancements on the existing methods that will help in effective and intelligent sentence alignment for building parallel corpora.

REFERENCES

[1] (Brown , 1991)Brown P,J.Lai and R.Mercer " Aligning Sentences in Parallel Corpora " 47th Annual meeting for the Association of Computational Linguistics.

[2] Gale and Church (1993) "A Program for Aligning Sentences in Bilingual Corpora" Computational Linguistics, also presented at ACL-91

[3] Robert C. Moore (2002) "Fast and Accurate Sentence Alignment of Bilingual Corpora" Microsoft Research, Redmond, WA 98052, USA.

[4] Xiaoyi Ma(2006) "Champollion: A Robust Parallel Text Sentence Aligner" Linguistic Data Consortium, 3600 Market St. Suite 810 ,Philadelphia, PA 19104

[5] Peng Li and Maosong Sun(2010) "Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm Department of Computer Science and Technology" State Key Lab on Intelligent Technology and Systems National Lab for Information Science and Technology

[6] Dan Tufiş(2007) "Exploiting Aligned Parallel Corpora in Multilingual Studies and Applications" Research Institute for Artificial Intelligence, Romanian Academy, 13, "13 Septembrie", 050711, Bucharest, Romania.

[7] Andr´e Santos1(2011) "A survey on parallel corpora alignment" Universidade do Minho..

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

[9] Ram Prakash H [2006], "Quillpad Multilingual Predictive Transliteration System ", Tachyon Technologies P Ltd.

[10] Stanley F. Chen, "Aligning Sentences In Bilingual Corpora Using Lexical Information ", Aiken Computation Laboratory , Division of Applied Sciences,Harvard University,Cambridge, M A 02138.

[11] http://telugu.webdunia.com/