Research Journal of Applied Sciences, Engineering and Technology 7(6): 1187-1198, 2014
DOI:10.19026/rjaset.7.379
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: March 19, 2013 Accepted: May 10, 2013 Published: February 15, 2014

Research Article

Domain biased Bilingual Parallel Data Extraction and its Sentence Level Alignment for English-Hindi Pair

Deepa Gupta, Vani Raveendran and Rahul Kumar Yadav
Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India

Abstract: Creation of Parallel Corpora and efficient corporal alignment at sentential level for structurally distinct languages having relatively low degree of correlation remains a challenge. This work emphasizes the importance of domain biased parallel data collection and a structured methodology to obtain the same for English-Hindi language duet. Further, its sentential alignment has also been undertaken since the participating languages are structurally distinct. In essence two aspects of this study is collection of parallel corpora from different domains and aligning the extracted parallel corpus at sentence level. The proposition is intended to help researchers in the field of Natural Language Processing help contribute better in terms of accuracy, precision and robustness of their proposition. This being possible only with availability of abundant parallel corpora and more so only if the parallel corpora are available domain wise and aligned at least at sentence level. The language pair considered for the development of the algorithm is English-Hindi. The algorithm being generic in nature makes our proposition scalable to other like structured language pairs.

Keywords: Cost calculation, Natural Language Processing (NLP), non-official data, normal distribution, official data, parallel corpus collection, semi-official data, sentential alignment

INTRODUCTION

Information plays an important role in today's world. As the gamut of information technology is expanding day by day, the need for its quick dissemination is also increasing. One major obstacle in this regard is that, not only is the information emanating from various corners of globe, it is being produced in various languages as well. Consequently, bilingual or even multilingual documents are quite common which span to include advertisements, web pages, Govt. and/or legal notices, etc. and are available as parallel texts which are text placed alongside its translation or translations. Large collections of parallel texts are called parallel corpora which is important to wide variety of applications like cross language information retrieval, sense based dictionary creation, POS taggers, Statistical Machine Translation (SMT) etc. The common resource for collecting parallel data is World Wide Web (WWW). The necessity of rich and abundant parallel corpora is increasingly becoming essential in various advanced language processing applications starting from SMT to automatic dictionary creation. More so in applications like MT were scarce data set produce mediocre results. Evidently, need for a structured methodology to parallel data collection is in order. Further, efficient utilization of the collected parallel corpus for mentioned applications require establishment of correspondences between texts. Alignment is the technique to achieve this goal by helping in establishing correspondence at various levels such as paragraph, sentence or word. The sentences can be aligned by using different methods. Some prominent ones include character length based alignment, word-pair correspondences and position of sentences in parallel corpus and hybrid approaches. Literature contains significant work on creating Parallel Corpora for structurally related languages and its process is now considered well established. On contrary, parallel data creation for structurally distinct language pair, especially English-Hindi duo remains relatively immature and less explored. However, domain wise parallel data collection and its sorting has never been explicitly promoted or addressed in any of past work, as far as authors knowledge extends. Domain biased data collection and sorting facilitates efficient and transparent evaluation of alignment algorithm. Since, the algorithm can then be specifically tested on Non-official type of data set like story books, advertisements, etc., where in alignments are not mere one to one correspondences unlike Official documents like Govt. articles. Toward this end our proposition aims to contribute an effective methodology to domain wise parallel data creation for English-Hindi duo and sentence level alignment of the collected data. Further, the work brings out a strategy on sentence level alignment of parallel texts for this is a prerequisite to many areas of applied linguistic research. This

Corresponding Author: Deepa Gupta, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India

exploration has immense importance since it aims to contribute a primary tool to a researcher in field of NLP to evaluate their contribution in terms of robustness, accuracy and reliability of their proposition which should be possible only with vast parallel corpora from various domains and more so only if they are aligned at the level of sentences.

LITERATURE REVIEW

This section includes details of existing efforts to parallel data creation apart from short description on various available sets of English-Hindi Parallel data. Further, few noteworthy work on sentence level alignments have been discussed in brief.

Parallel data collection for English-Hindi: Parallel corpus can be classified in three ways viz. number of languages, translation direction and level of alignment. Many of the parallel data extraction methods concentrate on the second type of the classification. Some of the earliest works include that of Baker *et al*. (2004) describing EMILLE corpus in terms of issues regarding parallel data collection. It explains the difficulties in collecting the PDF files and in data available as image. The work regards identification of correct source for Parallel Data collection as one of the major challenges encountered in data collection. The work by Chaudhury *et al*. (2008), focused on Gyan-nidhi corpus which contains parallel texts in multiple languages other than English and Hindi. The work extracts paragraphs from English-Hindi common books and aligned them using their machine translation based heuristic approach.

Bojar *et al*. (2010) proposed a method for collecting parallel corpus besides describing the various problems encountered in data of available English-Hindi parallel corpus. The work also proposed automatic methods of data cleaning and normalization. Also, Singh and Bandyopadhyay (2010) described about how the PDF files are converted to UTF-8 format while data collection for English-Manipuri duo, mainly from e-newspapers.

It is evident from the above discussion that literature lacks structured and robust parallel data collection methodology, especially for English-Hindi duo. Moreover, domain wise data collection has been completely ignored in many existing prominent propositions.

Existing data sets: As regards to existing data sets, there are few sets of English-Hindi parallel data available on WWW. For instance, EMILLE (Baker *et al*., 2004), which contains three major corporal classification viz. Monolingual, Annotated and Parallel corpus. It comprises of 200,000 words of text in English against its transliterated form in Hindi in addition to 14 other languages like Kannada, Tamil, Telugu, Punjabi, Gujarati, etc. The corpus contains data in two encoding schemes of which one is UTF-8 and other is the transliterated form. There are about 7000 sentence pairs in this corpus some of which are incorrectly aligned apart from containing multiple instances of spelling errors. Tides corpus, is another popularly used corpus that was collected during June 2002 for DARPA-tides surprise-language contest contains around 1.5 million words of English-Hindi parallel text. News articles, aligned at sentence level, form great portion of this corpus and thus one may regard it as Official data set. Multiple short comes of this data set make it unfit for usage. It contains Latin characters, besides being used incorrectly, in over 2000 sentences. Moreover, it contains multiple instances of meaningless sequence of identifiable characters in midst of various sentences. Daniel pipes corpus (website: http://www.danielpipes.org/) is yet another data set which is a collection of articles that describe Middle East. Originally written in English, it has been translated to 25 other languages including Hindi. For Hindi alone, there exist 322 articles which make approximately 6761 sentence pairs.

Amongst others are small data sets like Shabdanjali which is English-Hindi dictionary populated with 26,000 entries and ACL 2005, a subset of EMILLE (source: http:// www. ces. unt. edu/~rada/wpt05/). The Agriculture-domain parallel corpus is English-Hindi-Marathi-UNL parallel corpus developed by the Resource centre for Indian language Technology Solutions (source: http:// www. cfilt. iitb. ac. in/ download/ corpus/ parallel/ agriculture_ domaip_ parallel_corpus.zip) and contains only about 527 parallel sentences.

Markedly, many of parallel data source mentioned above have associated problems. Some are not open source, few contain errors and misaligned sentences and remaining are rather small data sets. More over the aspect of domain sorted data has not been brought out in all of the fore mentioned. Consequently, creating one's own reliable and relatively error free parallel data is in order. In such case, one should look up to WWW where parallel data is available in abundance. However, the data set cannot be used directly and the method proposed in this study a suitable strategy to generate usable parallel data set.

Sentential alignment: There are four major classes of sentential alignment methodologies viz. Length-based, Position-based, Lexicon-based and Hybrid approaches. Much of these sentence alignment techniques consider European languages for efficacy evaluation. Experimentation on English-Hindi pair has not achieved the maturity that European languages have attained, as far as the author's knowledge goes. Some of the initial noteworthy works include that of Gale and Church (1991) which developed alignment strategy for French-English parallel corpus. The work exploits

sentence length in terms of characters to evaluate likeliness of an alignment of some number of sentences in source language to some number of sentences in target language. The algorithm uses a dynamic programming technique to conclude on the type alignment viz. substitution, contraction, expansion, insertion, deletion or merging. Further, Sennrich and Volk (2011) proposed an iterative MT-based sentence alignment technique for German-French which considers initial alignment with length-based method of Gale and Church (1991).

Later the work by Kay and Roscheisen (1993) proposed aligning the sentences by using Lexicon based method. The method assumes that first and last sentences align. Then, until most sentences are aligned, envelopes of possible alignment were to be formed. Next, pairs of words that tend to co-occur in these potential partial alignments had to be chosen. Lastly, pairs of source and target sentences which contain many possible lexical correspondences were determined. The method can be computationally intensive depending on the chosen envelope size but pays huge dividends in robustness terms for non possible alignments never match. However, good lexicon design remains a challenge.

Further, Church (1993) aligned the sentences using location based approach with the assumption that "beads of the sentences in the two texts will have similar positions". In some cases the word positions are used to determine the position of the sentences. The fore mentioned approach works well in case of languages that are syntactically appreciably close. However, very few works report the usage of this approach for it is evident that the method would completely fail in case of syntactically different languages. For instance, it works well on foreign languages like French, English, Portuguese, etc., which have sentence syntax in SOV form but prove to be a complete failure for English-Hindi duo.

Later hybrid approaches to sentential alignment were experimented. Wu (1994) developed a Chinese-English corpus using lexical clues combined with the method by Gale and Church (1991) to obtain satisfactory alignment. Haruno and Yamazaki (1996) extended the method by Kay and Roscheisen (1993) and used two kinds of word correspondence viz. correspondence identification by external bilingual dictionary and correspondences obtained statistically to obtain the alignment. Yu *et al*. (2012) considered aligning sentences by combining power of Length and Lexicon mapping techniques i.e., Gale and Church (1991) and Kay and Roscheisen (1993). The method segments parallel text into words. Translated word pairs were formed and a matching penalty was determined for the words on these two sets. Further, by considering length of sentences a length penalty was assigned to each sentence. Finally, matching penalty and length penalty were compared to devise a new measure and the sentences were aligned according to this metric. Also, Singh and Bandyopadhyay (2010) developed a Manipuri-English sentence alignment algorithm, wherein Gale and Church (1991) method was used for finding the rough alignments. These alignments were compared against an external bilingual dictionary to find the wrongly aligned sentences and to make the needed corrections. Recently, Aziz and Specia (2011) report an sentential alignment by TCA specially written for aligning Brazilian Portuguese to other text like English and Spanish. However, most desired alignment type is substitution (1-1). Prior to this, minor data collection task has been described focusing on data collection of single genre. All the methods exploiting hybrid approach report improvement in alignment accuracy over conventional approaches in handling structurally distinct languages.

It is evident from above intensive review that Length based methods are by far the most popular and underpinning method of almost all alignment algorithms. The method gained popularity owing to its implementation simplicity, simple statistical data dependence and fastness in result delivery. Also, it is language independent in that the method requires almost no prior knowledge of language pair for it is based only on length of sentences. However, often the task of sentential alignment gets skewed if one sentence maps incorrectly, thereby affecting further alignments. Consequently, a prominent observation is the reduced robustness is reduced if length-based methods are applied to structurally distinct language pairs. Lexicon based methods used word correspondences for sentence mapping. In terms of robustness, lexicon based method outwits the length based approach, provided it works on accurately designed lexicon base. Clearly, its bottleneck is the design/availability of robust dictionary that should contain all sense of a considered lexicon. The location based methods are not used often because in most cases the positions of sentences in the two texts may not be same. Also, they can be used only if the language pairs are similar in structure. Evidently, location and lexicon based alignment technique shall not be efficient in handling alignment task for structurally distinct language pairs like English-Hindi. It shall be shown in subsequent discussions that enhanced length based approach to sentential alignment that uses weighted cost based method coupled with data based heuristics is worthy of handling sentence level alignment of English-Hindi data.

PARALLEL DATA COLLECTION

The common resource for collecting the parallel data is World Wide Web (WWW). An intensive research of over nine months has revealed that most documents on WWW can be classified primarily into following three categories:

- Official data
- Semi-official data
- Non-official data

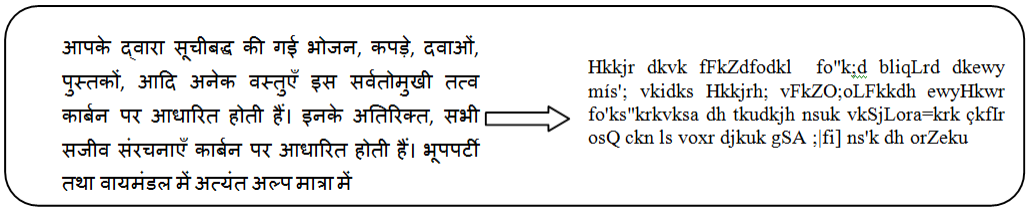

Fig. 1: Format I

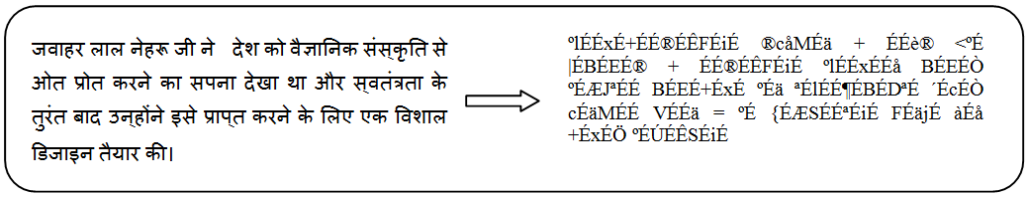

Fig. 2: Format II

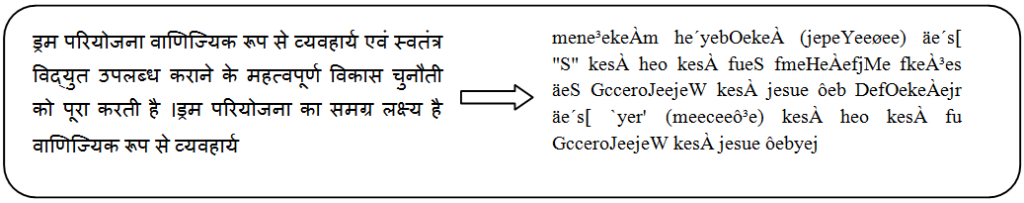

Fig. 3: Format III

The Official documents include the government documents, application forms, etc. The NCERT books (http://ncert.nic.in/ncerts/textbook/textbook.htm) home appliance manuals, etc., fall under Semi-official category of data set. The Non-official data spans to include story books, advertisements, recipes, etc.

Obscurities in parallel data collection: Numerous issues pertaining to the collection of parallel data exists. However, significant ones which are encountered for almost all language pair and also observed in English-Hindi duo are enlisted below.

Scarcity of uniformly encoded data: English-Hindi parallel resources exist in multiple formats such as. pdf, .html, .docx, .rtf, .txt, .doc, scanned documents, etc. and possess different encoding schemes. docx, .rtf, .txt, .doc, scanned documents, etc. and possess different encoding schemes. Requirement is the need of a common encoding scheme such as UTF-8 encoding.

Identifying suitable sources for data collection: Identifying the proper source for collecting the parallel data is another issue regarding parallel data collection. In some web sources even if English-Hindi data is available the information is not same in the parallel corpus. Clearly, identifying the proper source for collecting the data culminates into challenge and adds to the degree of obfuscation in parallel data collection.

PROPOSED PARALLEL DATA COLLECTION METHODOLOGY

The proposed parallel data collection strategy considers the parallel data existing in HTML, PDF, RTF, docx, doc and .txt formats. Scanned document processing requires additional modules like OCR and thus are omitted. Most data are found to exist as PDF files, rendering them useless for NLP tasks that perform word level processing.

Over 75-80% available PDF files have been have been found to occur in three major formats for English-Hindi duo, which has been discussed in succeeding subsections.

PDF data collection: The PDF files are not easy to save in the Unicode format compared with the other types of data. The English PDF files have a common encoding scheme such as ANSI. While Hindi PDFs follow different encoding schemes. Converting a Hindi PDF to text produces a document which is in unreadable format. The following steps have been proposed to convert Hindi PDF to process-able format:

Step 1: Save the PDF as text: On saving PDF as text one finds Hindi characters appearing in different formats that are not readable. As mentioned earlier, the formats are predominantly only in three types of forms.

Table 1: Statistics of parallel data

Data type	Language	No. of paragraphs	No. of sentences	No. of words
Official data	English	1390	12,229	1,11,884
	Hindi	1390	12,220	1,41,578
Semi-official data	English	1300	32,407	4,10,813
	Hindi	1300	32,335	4,92,558
Non-official data	English	25	4660	99,632
	Hindi	25	4450	1,31,072

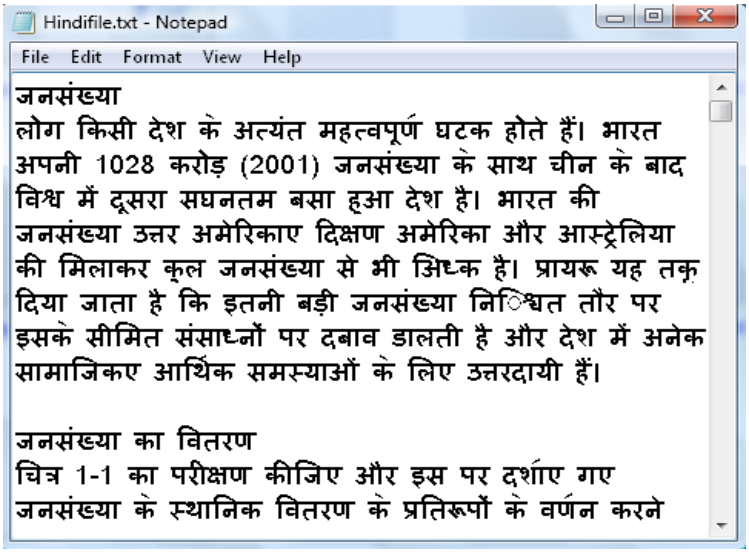

Fig. 4: Decoded Hindi text

Figure 1 to 3 shows the three formats of the encoded files. Customarily, NCERT books are found to translate to format shown in Fig. 1, while Government documents and documents of Reserve Bank of India have found to translate to formats shown in Fig. 2 and 3, respectively.

Step 2: Data pre-processing: After the conversion of the PDF files to text format several pre-processing needs to be done. This step involves the removal of bad characters, pictures, etc., in addition to the removal of English text from the Hindi file. For both English and Hindi PDFs, the pre-processing is mandatory.

Step 3: Decrypting using mapping table: This step requires the use of a manually constructed mapping table to decode the unreadable file into readable .txt file. The mapping table is constructed using the original pdf file and the text data. We have presented this table in the post reference section in Table 1. The table shows mapping for three formats of text data. Certain characters require extra mappings other than those shown in the table. The result of step 3 is illustrated in Fig. 4.

Step 4: Post-processing: Post processing is essential post decoding for some characters are incorrectly mapped. For instance "अिध्क", "ि्स", etc., should come post conversion. The characters need to be replaced. Moreover, two or more English words may be combined together. These defects also need to be eliminated.

Collecting allied data (non-PDF): HTML data, doc files, docx data and the text documents can be directly converted to UTF-8 format. In fact, HTML data is available in UTF-8 format itself. However, the doc files and text files are available in the ANSI encoded form. Therefore, the ANSI encoded text has to be converted to UTF-8 encoded form. In some cases, data may require prior cleaning, as the data saved in text format may contain bad characters. Also are the cases encountered where the Hindi file may contain English text? The cleaning step deals with the removal of all the fore mentioned characters from the text.

Paragraph alignment: After the preprocessing conversion, post processing steps and post cleaning, all the files arrive at a common encoding scheme, here UTF-8. Post parallel data collection the texts are aligned at paragraph level. The delimiter for the paragraph is taken as the blank line between the

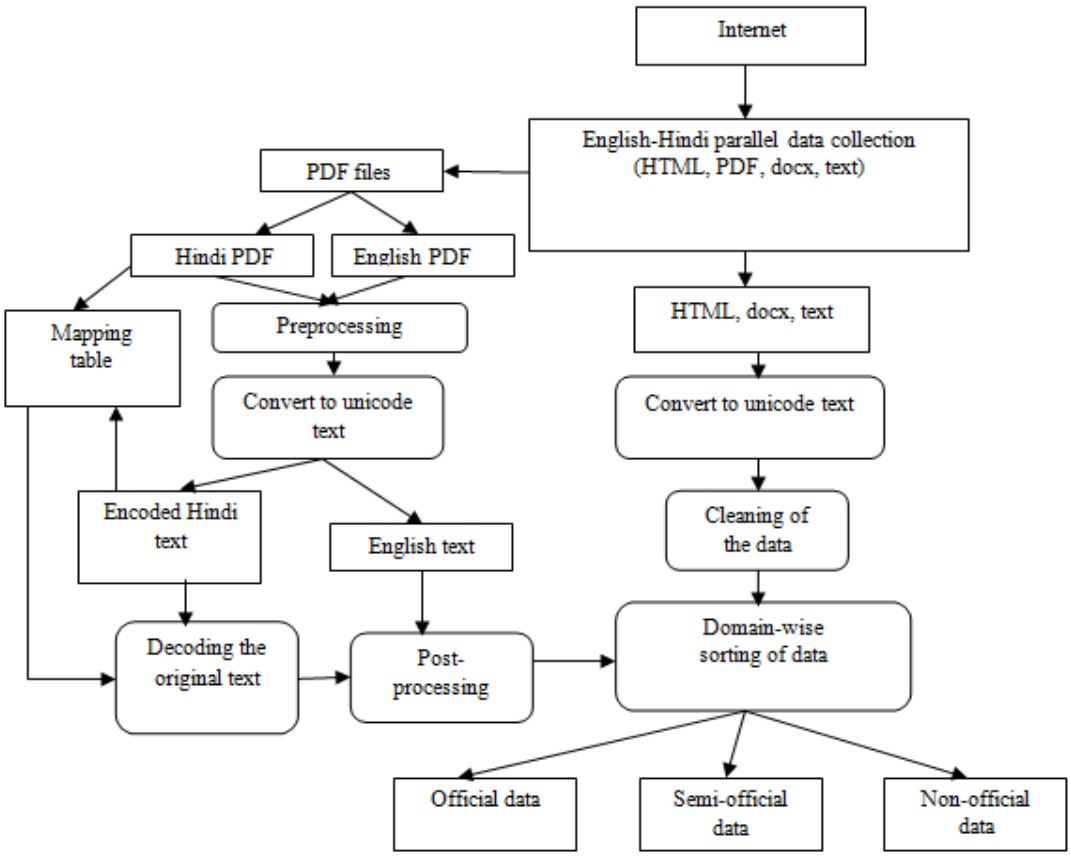

Fig. 5: Detailed design for data collection

paragraphs. For a sentential alignment program to operate, the number of paragraphs in bilingual file should agree. Thus this step deals with alignment of the paragraphs in the parallel corpus and the removal of unaligned paragraphs which can be considered inconsequential.

Domain-wise sorting of the data: The next step which is of utmost importance is the domain wise sorting of the data. Table 1 shows the statistics of the parallel data when sorted into their respective domain. A pre-written code can serve the purpose which should be designed to sort data based on the information obtained from keywords used in document title, URL, etc. To ensure correctness in data sorting, manual check may be undertaken. It can be inferred from Table 1 that intra-sentence count difference in the region of tens in parallel corpus of Official data and thus are nearly close in quantity when compared to Semi-official and Non-official data whose intra-sentence count disparity is roughly 100 and 200, respectively, suggesting high degree of one to one correspondences in former than later. In view of above analysis with further on atomized parameter i.e., word level analysis reveals that intra-word count contrast in Official data set has 35% drop which is noteworthy for this data set for it is 3 times more protracted corpus than Non-official data. Similar noteworthy, comparison exists for Semi-official vis-à-vis Non-official wherein former is about 6 times bigger in order than later. It is also evident from Table 1 that in data sets of all domains the Hindi word count exceeds corresponding English word count. This is due to fact that a word in English usually requires more than single word in Hindi to support its meaningful translation. For instance, the phrase "Leather Belt" translates to "चमडे की बेल्ट" and "is going" translates to "जा रहा है". In both cases it is observed that an English phrase requires more words in Hindi for meaningful translation. Evidently, sentential alignment will be reasonably more accurate in case of Official data than Semi-official or Non-Official data. Thus, it is extremely vital to classify data domain wise and test the proposed algorithm on these domain specific data to conclude on algorithm's robustness.

An intensive survey of multiple sentence alignment algorithms show that the description of data base for algorithm robustness test is omitted. All most all work reported have not focused on this aspect for evaluating the efficacy of their proposed algorithm, as far as author's knowledge extends.

The complete process of parallel data collection has been summarized in Fig. 5, intended to give quick insights to our proposed parallel data collection strategy.

Sentential alignment: The length-based alignment technique considers character length for alignment. It assumes lengths of a sentence and its translation to be strongly correlated which is mostly the case with a parallel corpus. Figure 6 provides the correlation of English and Hindi (L_1 and L_2, respectively) parallel sentence lengths based on number of characters for a random data set which justifies the fore mentioned and corroborate the assumption. The key aspect in sentential alignment is to establish matching for it is not necessary that each L_1 sentence be translated to a single L_2 sentence. In realistic translations one may find different variations of translation patterns.

Our study considers six patterns which are frequently encountered by refer them as α-alignments in subsequent discussions and Table 2 illustrates three of the typical α-alignment:

- **Substitution (1:1):** One L_1 sentence translates into one L_2 sentence only.
- **Contraction (2:1):** Two consecutive L_1 sentences translate into a single L_2 sentence.
- **Expansion (1:2):** One L_1 sentence of source language translates to two L_2 sentences.
- **Merge (2:2):** Two sentences jointly translate to two sentences (It is not two 1: 1 translations.).
- **Deletion (1:0):** A L_1 sentence is not translated at all.
- **Insertion (0:1):** A new sentence is inserted in the L_2 text that has equivalent in the L_1 text.

The sentences in Table 2 have "." (Full stop) and "।" (called Poornaviram) as sentence delimiter for English and Hindi respectively, while a new line character separates the two adjoining sentences.

Obscurities in English-Hindi sentence alignment: The prominent issues encountered in process of English-Hindi sentential alignment are presented in the succeeding subsections.

Difference in structure of the sentence: Unlike English, French, Portuguese, etc., structure of Hindi language is different. In some cases, a phrase in English may correspond to one Hindi word leading to noteworthy difference in sentence lengths of the two languages.

Punctuation and other issues: Sometimes ';' and ',' in English may translate to a full stop in Hindi or vice versa. Such instances induce sufficient level of contingency making sentence alignment task obscure. One may also encounter instances of mismatch of punctuations like ':', numbering, etc.

The following subsections describe the alignment algorithm by Gale and Church (1991) and our proposed enhancement over it to handle structurally distinct language considered viz. English-Hindi duo.

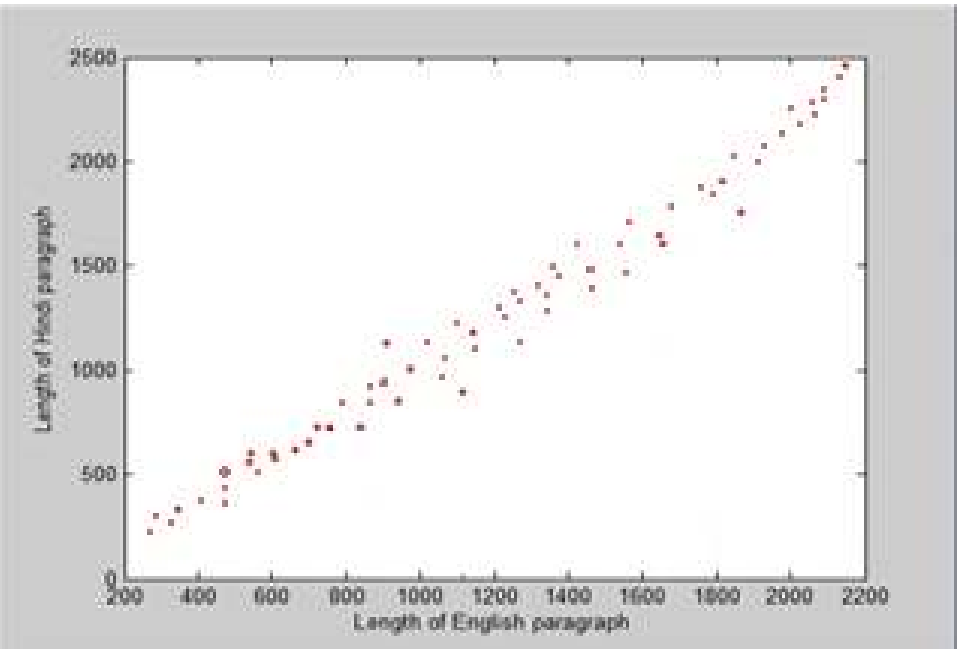

Fig. 6: Paragraph length correlation plot for a random English-Hindi parallel dataset

Table 2: Illustration of few α-alignments

Model	English text	Hindi text
2 : 1	Rajya Sabha is considered as the federal chamber. So it enjoys certain special powers under the constitution.	एक परिसंधीय सदन होने के नाते राज्य सभा को संविधान के अधीन कुछ विशेष शक्तियाँ प्राप्त हैं।
1 : 2	The 'Council of States' which is also known as Rajya Sabha, a nomenclature that was announced by the chair in the house on the 23rd August, 1954 has its own distinctive features	"राज्यों के परिषद" जिने राज्य सभा भी कहा जाता है, एक ऐसा नाम है जिसकी घोषणा सभा पीठ द्वारा सभा में 23 अगस्त, 1954 में की गई। इसकी ख़ास विशेषताएँ हैं।
2 : 2	The jackal stayed in the tub until he was sure that the dogs had gone away. Then slowly he crawled out of the tub.	उतनी देर सियार टब में ही छिपा रहा। जब उसे विश्वास हो गया कि कुत्ते चले गए तब वह धीरे-धीरे टब से बाहर निकल आया।

Length based method for sentence alignment: The conventional alignment algorithm proposed by Gale and Church (1991) assumes that both document of L_1 and L_2 have the same number of paragraphs and the i^{th} paragraph of L_1 corresponds to the i^{th} paragraph of L_2. Thus the sentence-level alignment has to be carried out for each paragraph. According to the algorithm, at every instance of sentence pair, evaluate the probability of the α-alignments is to be evaluated and a score to each alignment is to be assigned intended to reflect the degree to which the two segments related to one another. It is determined probabilistically with the help of statistics obtained from sample data. The cost of each α -alignment is obtained in terms of a distance measure between the lengths of the sentences currently under consideration. The cost evaluation with respect to character-length assumes that each character of language L_1 gives rise to a random number of characters in L_2.

Moreover, random variables are assumed to be independent and identically distributed on account of large data and this randomness is thus modelled by Normal distribution with mean c and variance s^2 where c is the expected number of characters in L_2 per character in L_1 and s^2 is its variance. These parameters are estimated from sample corpus data and given in (1) and (2), respectively:

$$c = \frac{Number\ \ of\ \ characters\ \ in\ L_2}{Number\ \ of\ \ characters\ \ in\ L_1} \quad (1)$$

$$s^2 = \frac{\sum (L_1\ paragraphlength - L_2\ paragraphlength)^2}{Total\ L_1\ paragraphlength} \quad (2)$$

Gale and Church (1991) considered language independent values of these parameters for European languages and assumed $c = 1$ and $s^2 = 6.8$ for their experimentation.

However, the calculated value of these parameters obtained from our sample data set is: $c = 12569/11065 = 1.1359$ and $s^2 = 6.96851$ and the same has been used in our experimentation. The distance measure d compares the difference in the sum of the lengths of the sentences of L_1 and L_2 respectively. If we denote the two lengths as l_1 and l_2, then probability score (δ) is calculated as shown in (3):

$$\delta = [(l_2 - l_1 c)] / [\sqrt{(l_1 s^2)}] \quad (3)$$

Subsequently, cost of α-alignment is calculated and is given by (4):

$$Cost\ (l_1;\ l_2) = -\log\ (P\ (\alpha\text{-align}) / \delta\ (l_1;\ l_2;\ c;\ s^2)) \quad (4)$$

The negative log is used so that cost is regarded as a measure of distance. The above probability is calculated using Baye's theorem in terms of constant times as shown in (5):

$$P\ (\alpha\text{-align}) * P\ (\delta/\alpha\text{-align}) \quad (5)$$

The constant can be ignored since it will be same for all proposed matches. The required conditional probability of P (δ/α-align) may then be estimated by the relation in (6):

$$\begin{aligned} P\ (\delta/align) &= P\ (z>|\delta|) + P\ (z<-|\delta|) \\ &= (1- \Phi\ (|\delta|)) + (1 - \Phi\ (|\delta|)) \\ &= 2\ (1- \Phi\ (|\delta|)) \end{aligned} \quad (6)$$

where, $z \sim N\ (0, 1)$ and $\Phi\ (\delta)$ is as shown in (7):

$$\Phi(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz \quad (7)$$

Finally, one out of the six possible alignments which would be the needed alignment is to be extracted. A dynamic programming, which tries to minimize the cost by considering the alignments made till that point, serves the purpose. The cost is to be computed using the recurrence relationship proposed by Gale and Church (1991) for English-European languages.

TEST DATA STATISTICS AND EVALUATION MEASURES

The documents were from WWW and had been converted to UTF-8 encoding scheme according to proposed strategy. The data was then sorted into three categories viz. Official, semi-official and Non-official data. A pre-written code served the purpose which was

Table 3: Development and test data statistics

Data type	Language	# of para.	# of sent.	Avg. sent./para.
Training data statistics				
Development set	English	245	1994	8.139
	Hindi	245	2057	8.396
Test data statistics				
Official data	English	70	475	6.790
	Hindi	70	477	6.810
Semi-official data	English	36	500	13.890
	Hindi	36	509	14.139
Non-official data	English	27	574	21.260
	Hindi	27	579	21.440

Table 4: Probability of α-alignment on development set

Type	Frequency	P (match)	Cost (l_1, l_2)
1 : 1	1905	0.9592	1.809
1 : 2	64	0.0322	149.210
2 : 1	9	0.0045	234.675
2 : 2	2	0.0010	300
0 : 1	2	0.0010	300
1 : 0	3	0.0015	282.390

designed to sort data based on the information obtained from keywords used in document title, URL, etc.

The experimentation conducted focuses on highlighting importance of domain biased data for algorithm's accuracy test apart from proposing an enhancement on length based method for sentential alignment of English-Hindi duo. The programming language used was Python 2.6.6 and Natural Language Tool Kit (NLTK) was the tool kit used. Approximately 2000 sentences were chosen in random from each of the three domains to create a development set which comprised of 50% official data, 25% semi-official and non-official data each and aligned them at the paragraph level. The statistics pertaining to development set is presented in Table 3.

The probability of match i.e., P (α-align) for the α-alignment on development set is shown in Table 4, as calculated using (5). Also, match probability was evaluated on each of three data sets separately. Its quantification has been omitted due to its low degree of relevance.

To prove efficacy of our proposition, we compare the outcomes of our algorithm against the results obtained using distance measure in Gale and Church (1991), on same domain biased data set comprising of approximately 500 sentences from each domain. The measures utilized for effectiveness evaluation and comparison are manually worked out % (P) recession and % (R) ecall computed in the manner shown below:

$$\%P = \frac{No.\ of\ correctly\ aligned\ sentences}{No.\ of\ alignments\ obtained\ from\ system} \times 100$$

$$\%R = \frac{No.\ of\ correctly\ aligned\ sentences}{No.\ of\ alignments\ according\ to\ Gold\ Standard} \times 100$$

EXPERIMENTATION USING LENGTH BASED APPROACH AND RESULT ANALYSIS

The alignment statistics obtained using the length based method by Gale and Church (1991) on official, semi-official and non-official data is presented in Table 5 to 7, respectively. In the tabulation, "Available" stands for the count of actual number of available type of α-alignment determined manually, while "Obtained" stands for the count obtained by the algorithm. "Correct" is the total number of correctly identified alignment type by the algorithm, which has been determined after a manual check.

Table 5: Evaluation on official data set using length based approach

Type	Available	Obtained	Correct	% P	% R
1 : 1	462	403	358	83.23	77.85
1 : 2	6	23	2	12.50	33.33
2 : 1	3	21	1	7.14	33.33
2 : 2	0	3	0	-	-
0 : 1	0	1	0	-	-
1 : 0	1	1	0	-	-
Total	473	452	361	79.87	76.32

Table 6: Evaluation on semi-official data set using length based approach

Type	Available	Obtained	Correct	% P	% R
1 : 1	479	431	256	59.40	53.44
1 : 2	19	30	5	16.67	26.32
2 : 1	1	17	1	4.00	2.50
2 : 2	0	2	0	-	-
0 : 1	0	1	0	-	-
1 : 0	0	1	0	-	-
Total	499	482	262	54.45	52.60

Table 7: Evaluation of non-official data set using length based approach

Type	Available	Obtained	Correct	% P	% R
1 : 1	437	223	127	56.95	28.93
1 : 2	45	107	10	9.35	22.22
2 : 1	31	105	18	17.14	58.06
2 : 2	9	17	0	-	-
0 : 1	3	0	0	-	-
1 : 0	12	0	0	-	-
Total	537	455	156	34.29	29.05

The algorithm is first evaluated on Official data and the results have been summarized in Table 5. It is found that most of the sentences belong to 1:1 type. Amongst 473 available alignment types almost all belong to 1:1 α-alignment type while few are of 1:2 and 2:1 none of other types. This corroborates with the discussion made using Table 1. The length based approach identifies 403 alignments as 1:1 type of which only 358 are correct. None of remaining identified alignments come close to their correct count. Consequently, other alignments also get skewed resulting in severely degraded precision and recall values. Evidently, performance using length based approach on structurally distinct language duo results in complete failure even on Official data set. Table 6 quantifies evaluation done on Semi-official data set. Compared to Official data, the Semi-official set also contains non 1:1 type alignment like 1:2. However, 2:2, 0:1 and 1:0 were not found in the test data set. Out of 479 available 1:1 α-type, 431 were obtained amongst which correct alignments extended to mere 256 in count

causing a significant turn down of 36% in precision and 24% in recall relative to the figures obtained for Official data set. It is so observed since this type of data contains 1:2 and 2:1 alignments more in number compared to the official data. Also inferred is only some percent of the 1:2 alignments are found to be correct. It can be inferred from Table 7 that out of 437 available 1:1 alignments only 127 were correctly obtained. Thus, precision of the alignments are further degraded on account of similar reasons and even more on account of presence of higher 1:2, 2:1, 2:2, etc., types of α-alignments. Clearly, Non-official data sets pose dual challenge to sentence alignment. Firstly, they contain multiple non 1:1 type alignments and secondly, large intra-word count disparity leading to failure of conventional length based approach. A reduction of approximately 46 and 49% in precision and recall, respectively, is observed relative to the respective quantities obtained for Official data set.

From these results it can be concluded that large intra-word count disparity in Semi-official and even more in Non-official data set caused significantly low, with an average of 34.29% precision and 29.05% recall, in the parameters that govern the accuracy of alignment algorithm owing much to the features of language duo considered and the length based algorithm exploited. Evidently, biasing the distance measure in accordance to the domain type of data is in order. This is the proposition made as discussed in the succeeding section.

Proposed enhancement and experimentation: The proposed enhancement has been discussed in the succeeding subsections along with the experimentation and result analysis.

Proposed enhancement: The proposed modification on conventional length based algorithm is primarily to handle structurally distinct language pairs, here English-Hindi duo, for conventional methodology was developed to tackle like structured languages like English-European, etc. It demonstrates handicapped-ness in handling the former. The analyses of statistics of sentence alignment accuracy using the conventional approach reveals important fail factor of its inability to handle alignment of structurally distinct language pairs. This happened since it identified numerous false instances of α-alignment types like 1:2, 2:1, 2:2, 1:0 etc., leading to skewed effect in detection of appropriate α-alignment type leading to a cascaded effect of incorrect alignment detection. Evidently, it contributes to multiply fall in accuracy of alignment. The prime factor to above was the variation of word count disparity present in data sets of various domains as discussed under Section paragraph alignment with Table 1. The failure also justifies the need for sentential alignment algorithm's efficacy evaluation on domain classified data prior to concluding on its robustness. Toward this end we propose an anti-skewing approach to tackle the problem by a weighted cost calculation strategy as described below.

Weighted distance measure: The weighted distance approach is a consequence of analysis made from performance of conventional length based approach on domain biased data sets. The authors would thus like to reiterate that study and performance analysis of an alignment algorithm on domain biased data set is of utmost importance to complete its robustness test which has been found ignored in all prominent propositions made thus far. The modified cost calculation proposed is improvement devised to increase the penalty gap between the alignments. The increase in the penalty gap will result in high penalty for the alignments which are less in number, thereby increasing the correctness of the alignments which are more in number. It can be inferred that our cost calculation strategy is weighted in that it has an empirically determined constant involved in the computation of the cost:

$$D(i,j)=\begin{cases} D(i,j-1)+w*cost(0{:}1\,align0,tj{:}0,0) \\ D(i-1,j)+w*cost(1{:}0\,alignsi,0;0,0) \\ D(i-1,j-1)+cost(1{:}1\,alignsi,tj;0,0) \\ D(i-1,j-2)+w*cost(1{:}2\,alignsi,tj;0,tj-1) \\ D(i-2,j-1)+w*cost(2{:}1\,alignsi,tj;si-1,0) \\ D(i-2,j-2)+w*cost(2{:}2\,alignsi-1,si;tj-1,tj \end{cases} \quad (8)$$

Intuitively, the idea of weighted distance measure is to assign weight to costs of α-alignments other than 1:1 resulting in abrupt increase in 1:1 type due to high penalty gap. This reduces the error multiplication on account of cascading phenomenon due to detection of lower compound alignment types, like 1:2, 2:1, 2:2, etc., followed by improvement in precision and recall. The effectiveness of the bias constant depends on the data set involved. Experimentally verified weights for English-Hindi duo are 100 each for Official and Semi-official data set and 10 for Non-official data set. The cost D (i, j) is to be computed using the recurrence relationship biased with the proposed weights given in (8).

Heuristic measure: Mild heuristics have demonstrated prevention of pronounced dynamics of typical alignments errors.

Minor alignment strategy supplements, as proposed in the following, have exhibited sufficient potential in increasing the accuracy of existing alignment algorithm for English-Hindi duo. These are enlisted below and are generic in nature making them usefully scalable to other like language pairs.

Like phonetics: Words phonetically similar in both source and target languages can be used as a key to map

Table 8: Evaluation of official data set using proposed weighted distance measure

Type	Available	Obtained	Correct	% P	% R
1 : 1	462	462	406	87.79	87.79
1 : 2	6	5	2	40.00	33.33
2 : 1	3	4	1	25.00	33.33
2 : 2	0	0	0	0	0
0 : 1	0	1	0	0	0
1 : 0	1	0	0	0	0
Total	473	472	410	86.50	86.68

Table 9: Evaluation of semi-official data set using proposed weighted distance measure

Type	Available	Obtained	Correct	% P	% R
1 : 1	479	487	392	80.49	81.84
1 : 2	19	11	6	54.55	31.58
2 : 1	1	1	1	100	100
2 : 2	0	0	0	-	-
0 : 1	0	0	0	-	-
1 : 0	0	0	0	-	-
Total	499	499	399	79.96	79.96

Table 10: Evaluation of non-official data set using proposed weighted distance measure

Type	Available	Obtained	Correct	% P	% R
1 : 1	437	456	334	73.25	76.43
1 : 2	45	37	23	62.16	51.11
2 : 1	31	32	20	62.50	64.52
2 : 2	9	7	1	14.29	05.26
0 : 1	3	0	0	-	-
1 : 0	12	3	0	00.00	00.00
Total	537	535	378	70.65	70.39

sentences. It is mostly seen in nouns which usually do not change in any of the languages. For instance, English words could be written in Hindi without any translation such as "school" in English sentence may be used as "स्कूल" in the Hindi sentence.

Special words: Certain words like abbreviations, reference numbers, date, rarely used words; etc., serve as anchors in sentential alignment for it reliably indicates sentential position. Somers (1998) reports a 1.4% increase in alignment accuracy for English-European parallel text, using special words alone.

Experimentation on proposed enhancement: The outcomes after applying the proposed improvement strategies and heuristics prove the efficacy of the enhancement devised based on the conclusions derived from Domain biased data set, when compared to corresponding results obtained using distance measure of the conventional Length based approach. The same data set has been used for efficiency quantification in Table 8 to 10.

A noteworthy 7 absolute units (approx.) of increment in precision and 10 units in recall is observed by comparing Table 5 and 8. The "Available" and "Obtained" alignments nearly tally perfectly. Also, most of "Obtained" are correct, at least in 1:1 alignment type. Furthermore, the biased distance measure has proved to be successful in handling sentence alignment in Semi-Official data set. An increased number of 1:1 alignment detection in conjunction with highly suppressed false detection of compound α-alignments led to a significant increase of 26 and 27% in precision and recall, respectively, in comparison to those determined by conventional approach in Table 6. A remarkable observation is inferred from Table 10 which explicitly highlights the effectiveness of the bias incorporated in the distance measure which assigns weight to costs of α-alignments other than 1:1 resulting in abrupt increase in 1:1 type due to high penalty gap for remaining alignments. This reduces the error multiplication on account of cascading trend upon detection of alignment types, like 1:2, 2:1, 2:2, etc. This led to a significant improvement of precision and recall by approximately 38 and 40%, respectively, relative to corresponding figures obtained in Table 7.

CONCLUSION

The study proposes firstly, the importance of Domain biased data set for the evaluation of robustness of an alignment algorithm, which has not been given due importance in all prominent works, as far as authors' knowledge goes. The authors classify domain of major importance as Official, Semi-Official and Non-official. Secondly, an effective strategy to parallel data creation from WWW resource has been proposed intended to aid a researcher in field of NLP to evaluate his contribution in terms of robustness, accuracy and reliability of his proposition which should be possible only with vast parallel corpora from various domains. The proposed method of extraction can be applied to data of .pfd, .doc, .docx, .txt and .html format. Later, the parallel data may be sorted in three major domains viz. Official, Semi-Official and Non-official data set, in order to carry out the needed experimentation. One can benefit from vast parallel data only if it is aligned at least at the level of sentences. Thus, the work thirdly proposes a light weight sentence aligner, for structurally distinct language duets, based on simple probabilistic model motivated by the observation that longer regions of text tend to have longer translations and shorter regions of text tend to have shorter translations. An experimentation based on length based algorithm by Gale and Church (1991) was carried out on English-Hindi duo and was shown to be handicapped in handling structurally distinct languages which consists increased intra-word count disparity instances and even more when domain based data is considered. Toward this end a weighted distance measure was devised which considers biased cost calculation using empirically determined weights. The incorporated bias in the distance measure assigned weight to cost of α-alignments other than 1:1 resulting high penalty gap for non 1:1 type alignments. This led to a significant improvement of precision and recall when compared to those obtained using conventional length based approach. While small amendment demonstrated

noteworthy potential in sentence alignment of structurally distinct language pair, it shall be interesting to devise a hybrid approach that should consider an iterative lexicon matching, exploiting intensive NLP techniques, in conjunction with the proposed weighted cost based alignment approach to further improve alignment task. Our future study shall investigate this intersection intended to further improve the parameters that govern the accuracy of the corpus alignment algorithm for structurally distinct languages.

REFERENCES

Aziz, W. and L. Specia, 2011. Fully automatic compilation of Portuguese-English and Portuguese-Spanish parallel corpora. Proceeding of 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011). Cuiaba, Brazil.

Baker, P., A. Hardie, T. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizaukas, O. Hamza, D. Maynard, V. Tablan, C. Ursu, B.D. Jayaram and M. Leisher, 2004. Corpus linguistics and south asian languages: Corpus creation and tool development. Lit Linguist Comput., 19(4): 509-524.

Bojar, O., P. Straňák and D. Zeman, 2010. Data issues in English-to-Hindi machine translation. Proceeding of the International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta, May 17-23.

Chaudhury, S., D.M. Sharma and A.P. Kulkarni, 2008. Enhancing effectiveness of sentence alignment in parallel corpora: Using MT & heuristics. Proceeding of the International Conference on Natural Language Processing (ICON-2008). Macmillan Publishers, India.

Church, K.W., 1993. Char align: A program for aligning parallel texts at the character level. Proceeding of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio.

Gale, W.A. and K.W. Church, 1991. A program for aligning sentences in bilingual corpora. Proceeding of 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, California, pp: 1-8.

Haruno, M. and T. Yamazaki, 1996. High performance bilingual text alignment using statistical and dictionary information. Proceeding of the 34th Annual Meeting on Association for Computational Linguistics, pp: 131-138.

Kay, M. and M. Roscheisen, 1993. Text translation alignment. J. Comput. Linguistics-Special Issue Using Large Corpora, 19(1): 121-142.

Sennrich, R. and M. Volk, 2011. Iterative, MT-based sentence alignment of parallel texts. Proceeding of the Nordic Conference of Computational Linguistics. Riga, May 11-13, pp: 1-10.

Singh, T.D. and S. Bandyopadhyay, 2010. Semi-automatic parallel corpora extraction from comparable news corpora. Polibits, 41: 11-18.

Somers, H., 1998. Further experiments in bilingual text alignment. Int. J. Corpus Linguist., 3: 115-150.

Wu, D., 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. Proceeding of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico, pp: 80-87.

Yu, Q., A. Max and F. Yvon, 2012. Aligning bilingual literaryworks: A pilot study. Proceeding of the Workshop on Computational Linguistics for Literature. Association for Computational Linguistics, Canada, pp: 36-44.