Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments

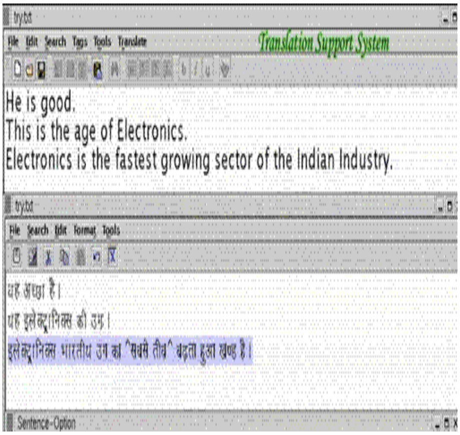
[1]Pardeep Kumar[2] Vishal Goyal
[1]M.Tech(ICT) Student, [2] Senior Lecturer
Department of Computer Science
Punjabi University, Patiala
E-mail: pardeep_800 yahoo.com

ABSTACT- In this survey paper, we have taken problem of "development of Hindi-Punjabi parallel corpus using existing Hindi to Punjabi machine translation system and using sentence alignment**"**. The alignment based on the length based technique, location based technique and lexical techniques. We will use Hindi-Punjabi machine translation system (i.e h2p.learnpunjabi.org). These tasks are need to Hindi-Punjabi parallel corpus. Sentence alignment is useful to developing Hindi-Punjabi parallel corpus and Hindi-Punjabi dictionary. The accuracy is basically depending upon the complexity of the corpus, more the complexity less the accuracy. Complexity means how to distribution of sentence in the target file. If any of these categories 1:1, 1:2, 2:1, 1:3, 3:1 sentences occur simultaneously in a paragraph. Our objective in this research paper is to developed Hindi-Punjabi parallel corpus using latest and existing techniques and method with a high accuracy and time efficiency.

Keyword- Parallel Corpus, Hindi-Punjabi, Sentence Alignment, length based, Location based

1. INTRODUCTION

A **parallel corpus** is a corpus that contains a collection of original texts in language L_1 and their translations into a set of languages L_2 ... L_n. In most cases, parallel corpora contain data from only two languages. Where the texts, paragraphs, sentences, and words are typically linked to each other.


try.txt
File Edit Search Tags Tools Translate
Translation Support System
try.txt
File Search Edit Format Tools
Sentence-Option

He is good.
This is the age of Electronics.
Electronics is the fastest growing sector of the Indian Industry.

यह अच्छा है।
यह इलेक्ट्रानिक्स की उम्र।
इलेक्ट्रानिक्स भारतीय उग का ^सबसे तीव्र^ बढ़ता हुआ खण्ड है।

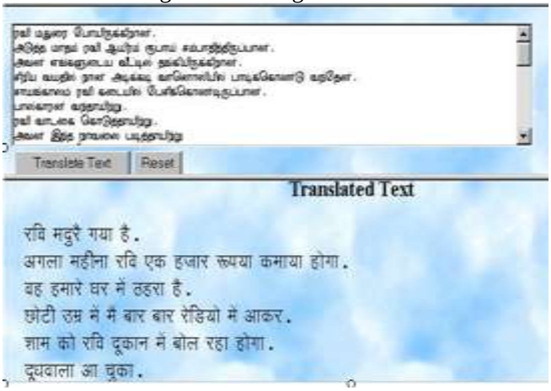
Fig 1.1

Alignment of corpus is basically of three types:

- Sentence-wise
- Paragraph-wise
- Word-wise

1.1 Sentence-wise:

Sentence alignment of Parallel corpus is the identification of the corresponding sentences in both halves of the parallel text. Alignments of parallel corpora at sentence level are prerequisite for many areas of linguistic research. During translation, sentences can be split, merged, deleted, inserted or changed in order. Basically the shorter sentences are aligned with shorter sentences and longer sentences are aligned with longer sentences.



[illegible]

रवि मदुरै गया है.
अगला महीना रवि एक हजार रूपया कमाया होगा.
वह हमारे घर में ठहरा है.
छोटी उम्र में मै बार बार रेडियो में आकर.
शाम को रवि दूकान में बोल रहा होगा.
दूधवाला आ चुका.

Fig 1.2 During translation sentence can be suplit

1.2 Paragraph-wise:

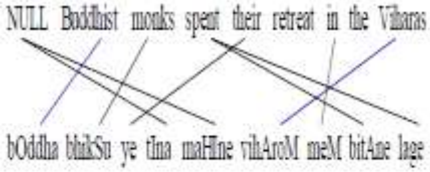
Paragraph alignment of Parallel corpus is the identification of the corresponding paragraph in both of parallel text in terms of number of sentences in it.

[illegible]

Fig 1.3

1.3 Word-wise:

Word alignment of Parallel corpus is the identification of the corresponding words in both halves of the parallel text. Automatic word alignment means that without the human interaction the parallel corpus should be aligned with the machine accurately.


NULL Boddhist monks spent their retreat in the Viharas
bOddha bhikSu ye tIna maHIne vihAroM meM bitAne lage

An **example** of an alignment between an English-Hindi sentence pair, blue links indicates alignment of cognates

2. TECHNIQUES USED IN DEVELOPMENT OF HINDI-PUNJABI PARALLEL CORPUS

- Length Based Technique.
- Location Based Technique.
- Lexical Based Technique.

2.1 Length Based Technique-

short sentences will be translated as short sentences and long sentences as long sentences.

2.2 Location Based Technique-

Based on statistical information, beads of sentences in the two texts have similar positions, these approaches do not attempt to align beads of sentences but rather just to align position offsets in the two parallel texts.

2.3 Lexical Based Technique-

Account the lexical information about texts, bilingual corpus is used to match the content words in one text with their correspondences in the other text and use these matches as anchor points in the sentence alignment process.

3. PROBLEM DEFINATION AND OBJECTIVE

The problem definition of project is "Development of Hindi-Punjabi parallel corpus using existing Hindi to Punjabi machine translation system and using sentence allignment". A collection of pairs of texts in two different languages where each member of pair is a translation of other is called a "Parallel Corpus".
The objective of our research is to Develop Hindi-Punjabi Parallel corpus using existing Hindi to Punjabi machine translation system. The development of Hindi-Punjabi parallel corpus has been misinterpreted as an alignment problem. The alignment problem is the next step of our research problem. If the parallel corpus is available of particular language pair, automatic alignment for that parallel corpus can be done, but there is no parallel corpus available for Hindi-Punjabi language pair. One of the PhD research scholar of Department of Computer Science, Punjabi University Patiala has already developed Hindi-Punjabi machine translation system with good accuracy. Thus we will make use of this machine translation system for developing the parallel corpus for Hindi-Punjabi language pair. Our work will also include development of a small statistical text analyzer which includes calculating sentence statistics, word statistics and character statistics. We will also develop an administration module for Hindi to Punjabi machine translation system, which will include the provision for adding new words in its dictionary, updating the correction (if any) etc.

4. RELATED WORK

Researchers have worked for non-Indian languages but very little work has been done for Indian languages & that is the focus of our project. Sentence alignment is a crucial part because it is the process of determining which sentence in a given source & target language sentences pair are translations of each other. A parallel text consists of a source language text & its translation into some target language.

Gale and church (1993) has developed French and English parallel corpus. They uses sentence length (measured in characters) to evaluate how likely an alignment of some number of sentences in L1 is with some number of sentences in L2.The algorithm uses a Dynamic Programming technique that allows the system to efficiently consider all possible alignments and find the minimum cost alignment.The method performs well (at least on related languages). It gets a 4% error rate. It works best on 1:1 alignments [only 2% error rate]. It has a high error rate on more difficult alignments. [4]

D.Wu, (1994) has developed Chinese and English parallel corpora in the Department of Computer Science and University of Science & Technology in Clear Water Bay, Hong Kong.Here two methods are applied which are important once.
Firstly, the gale's methods is used to Chinese and English which shows that *length-based methods* give satisfactory result even between unrelated languages which is a surprising result. Next, it shows the effect of adding lexical cues to a length –based methods. According to these results, using lexical information increases accuracy of alignment from 86% to 92%. [3]

Brown et al., 1991 Same approach as Gale and Church, except that sentence lengths are compared in terms of words rather than characters. Other difference in goal: Brown et al. didn't want to align entire articles but just a subset of the corpus suitable for further research. [2]

Sheng et al., 1994 It uses not only the length of sentences but also the length of texts, the length of upper and lower part of the candidate sentences, and some information like that to reinforce the effect of location of sentences in the text. In this sense it can be said that it is a further step of pure length-based method. [7]

Bridget and Ted (2003) has developed English-French and Romanian-English parallel corpus. The main approach is used for both English-French and Romanian-English. It is Perl implementation of IBM Model-2. In this process, approximately 50,000 sentences aligned pairs are used as training data for each language pair.The plain2snt program converts raw sentence aligned parallel text into snt format, where each word type in the source and target text is represented as a unique integer. This program also outputs two vocabulary files for source and target languages that list the word types and their integer values. A distortion factor is used to limit the number of possible alignments that are considered. The approach is tested using precision, recall, the f-measure and the alignment error rate (AER). The precision is .5292; recall .4706 and AER .5018 for Romanian-English language pair and for English-French precision .5305, recall .2136 and AER .4400. The precision of two language pairs is relatively similar, because they used approx. the same amount of training data for each language pair. [1]

***Kay & Roscheisen, 1993*_**Idea: Use word alignment to help determine sentence alignment.Then use sentence alignment to refine word alignment.

Method:

1. Begin with start and end of text as anchors
2. Form an *envelope* of all possible alignments (no crossing of anchors) where:
3. possible alignments must be at a certain distance away from the anchors
4. The distance increases as we get further away from the anchors
5. Choose pairs of words that co-occur in these potential alignments
6. Repeat steps 2-5 Pick the best sentences involved in step 3 (having the most lexical correspondences) and use them as new anchors.[5]

Haruno & Yamazaki, 1996 their method is a variant of Kay & Roscheisen (1993) with the following differences: For structurally very different languages, function words impede alignment. They eliminate function words using a POS Tagger.If trying to align short texts; there are not enough repeated words for reliable alignment using Kay & Roscheisen (1993). So they use an online dictionary to find matching word pairs. [6]

Zhonghua xiao, Tony Mc Enery (2002) has developed Asian language corpora and presented two corpora, developed at Lancaster University, together with exploration tools for use with these corpora. The standards we propose here work well with Asian language corpora, as demonstrated by our practice in corpus development; they also conform to the current trends in the international NLP community. The two corpora we developed also constitute an improvement to the Asian language resources. Asia is a continent of many languages and is potentially rich in language resources.The situation could also be improved by corpus builders working on Asian languages standardizing corpora so as to facilitate data interchange. We have learned from our work with CIIL on the EMILLE project that collaboration is better than competition. Our experience in collaborating with the Xara team also tells us that the cooperation between corpus creators and soft-ware developers can produce better corpora and better corpus tools. It is our belief that the cooperation and collaboration between centers and institutes worldwide will undoubtedly give rise to the further development of Asian language corpora. [8]

5. CONCLUSION

Thus In literature survey we have seen that most researchers on sentence alignment, especially if Bilingual texts are French, German, English or Chinese; use Hansards of these countries for a reliable common bilingual database. But no such Hansard exists in Hindi-Punjabi bilingual texts. This parallel aligned corpus development is the context of Indian languages. The parallel corpus as a translation memory (TM) will be a valuable source in improving the translation system and translator efficiency.

As discussed, there are many different approaches to sentence alignment. The first approach discussed was the character-length based approach. For the Hindi and the Punjabi languages, where there exists a similarity or co-relation between the sentence lengths in terms of number of characters per sentence, the character length approach is useful and also one idea to derive such formula is to use dynamic programming to find a minimum cost alignment, assuming a simple hidden generative model that emits sentences of varying lengths.

6. REFERENCES

[1] Bridget Thomson McInnes, Ted Pedersen, “The Duluth Word Alignment System”, participated in the 2003 HLT-NAACL Workshop on Parallel Text.

[2] Brown, P.; Lai, J.; and Mercer, R. (1991)."Aligning sentences in parallel corpora."

[3] D. Wu. “Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria” *In: Proc. of the 32nd Annual Conference of the ACL*: 80-87. Las Cruces, NM in 1994. http://acl.ldc.upenn.edu/P/P94/P94-1012.pdf

[4] Gale William A., Church Kenneth W., 1993, *A Program for Aligning Sentences in Bilingual Corpora,* AT&T Bell Laboratories

[5] Kay, M. and Röscheisen, M: Text-Translation Alignment, Computational Linguistics 19:1 (1994) 121-142

[6] John C. Henderson, “sentence Alignment Baselines” HLT-NAACL 2003Workshop: Building and Using Parallel Texts Data Driven MT and Beyond, Edmonton.

[7] Weigang Li, Ting Liu, Zhen Wang and Sheng Li: Aligning Bilingual Corpora Using Sentences Location Information, Proceedings of 3rd ACL SIGHAN Workshop, 141-147, (1994)

[8] Zhonghua xiao, Tony McEnergy, Paul Baker, Andrew Hardie “Developing Asian language corpora in (200809)” standard and practice in Department of Linguistics Lancaster University Lancaster.