Automatic Sentence Alignment Algorithm for Bilingual Hindi-Punjabi Parallel Text

Vishal Goyal[1] Gurdarshan Singh Sandhu[2]

ABSTRACT

This paper presents a project for automatically aligning sentences of bilingual parallel Hindi – Punjabi Parallel Texts. Bilingual Hindi-Punjabi Corpus has been collected from resources like CDAC Noida, Book Publishers and others. This automatic sentence alignment of bilingual Corpus is very beneficial in developing machine translation systems. The work involves the alignment of bilingual texts at sentence levels. The alignment algorithm used in this project used the concept of sentence length. The algorithm for aligning sentences is based on a simple statistical model of sentence lengths in terms of number of words in a sentence for aligning first at the paragraph level and then to sentence level. The program uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. This simple approach has shown remarkable results. An evaluation was performed based on a parallel corpus from different fields and almost all the sentences were correctly aligned. For very large corpus, the accuracy achieved is 74.6 %. This proposed algorithm can be implemented for other closely related language pairs.

Keywords: Bilingual corpus, statistical model, hindi punjabi Parallel text, sentence alignment, paragraph alignment

1. INTRODUCTION

In computational linguistics, a corpus is a collection of spoken or written utterances of natural language usually accessible in electronic form. Often, corpora represent a particular genre of text or speech. However, a corpus is always just a sample and can never completely represent a whole language. The expressive power of natural language cannot be captured by a finite data set. *Parallel corpora* are referred to as natural language utterances and their translations with alignments between corresponding segments in different languages. In a parallel corpus, the same body of text appears in two or more languages. Parallel corpora usually contain a common source document (the original) and one or more translations of this source (target documents). Bilingual parallel corpora are sometimes called *bitexts* or bilingual parallel text and corresponding parts within these corpora are called *bitext segments*. Parallel corpora have been exploited in many studies. Many applications use parallel corpora for translation studies and for tasks in multilingual natural language processing (NLP). Bilingual concordances have been used for some years in order to support human translation. In recent years, parallel corpora have become more widely available and serve as a source for data-driven NLP tasks. Automatic extraction of multilingual term databases, statistical machine translation, corpus-based bilingual lexicography are just some research fields that have been developed in

[1]Lecturer, Department of Computer Science, Punjabi University Patiala. E-mail : goyal_vishal@yahoo.com.sg

[2]M.Tech.(CSE) Student, Department of Computer Science, Punjabi University Patiala.
E-mail : gurdarshansandhu@yahoo.com

connection with a growing number of large parallel corpora.

The most widely used parallel corpora are derived from the English and French records of the Canadian Parliament, the so-called *Hansards* corpora. Like the Hansards, most parallel corpora contain only two languages, a source and a target language. However, multilingual parallel corpora with translations into more than one language are available and became very popular in recent studies. Examples of such corpora are the Multext East "1984" corpus for central and eastern European languages, the multilingual parallel corpus of European Parliament proceedings EUROPARL in eleven languages and the multilingual OPUS corpus. [1]

The task of *sentence alignment* is to estimate which sentence or sentences in one language correspond with which sentence or sentences in the other language. An aligned parallel corpus provides an aid to human translators since it is possible to look up all sentences in which a word or a phrase occurs to see the ways in which that word or phrase has been translated into the other language.

Sentence alignment also facilitates word alignment, since various statistical measures exist which determine instances where a word in one language consistently appears in sentences aligned with sentences containing the equivalent word in the other language. The sentence alignment task is to identify correspondences between sentences in one language and sentences in the other language. This task is a first step toward the more ambitious task finding correspondences among words.

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.

Automatic sentence alignment of Parallel Corpus means that without the human interaction the parallel corpus should be aligned with the machine accurately. We have used standard techniques for the establishment of links between source and target language segments. Researchers in both machine translation (Brown et al. 1990) and bilingual lexicography (Klavans and Tzoukermann 1990) have recently become interested in studying bilingual corpora, bodies of text such as the Canadian Hansards (parliamentary debates), which are available in multiple languages (such as French and English) [1]. Enthusiasm for this relatively new field of work was sparked early on by the apparent demonstration that very simple techniques could yield almost perfect results. For instance, two proposed methods (Gale's and Brown's) that completely ignored the lexical content of the texts, and relied almost entirely on the intuition that shorter sentences tend to translate into short sentences , while longer sentences tend to translate into longer ones. With simple programs in which this observation was encoded into a statistical method, both teams were able to achieve accuracy levels exceeding 98%. With the development of the corpus linguistics, more and more language resources have been established and used in language engineering research and applications. As we all know, there are different kinds of corpora for different kind of applications.

Sentence level alignments algorithms have so far proved very useful in a number of applications, which could be categorized as: bilingual lexicography, Automatic or machine –assisted translation verification, automatic acquisition of information about translation (Brown et al.,1993).

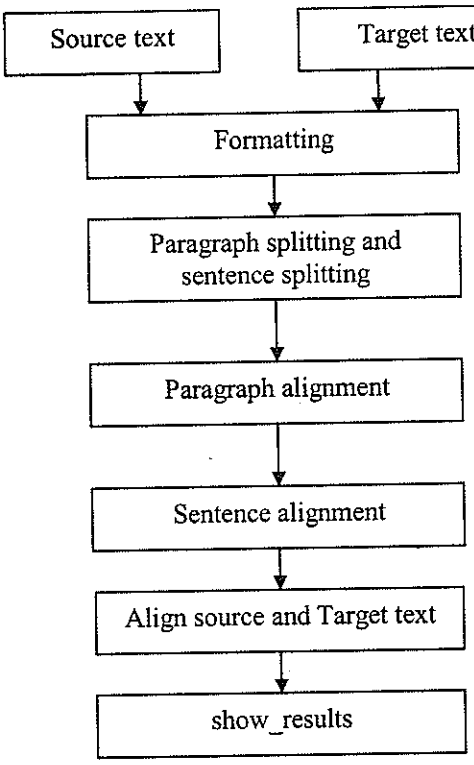
Basically, alignment aims to succeed the task of extracting structural information and statistical parameters from bilingual corpora. [18]

Researches have worked for non-Indian languages But very little work as been done for Indian languages and that is the focus of our project. Gale and Church (1993) uses sentence length criteria. The method performed 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. (About 86%). Brown et al. (1991) uses same approach as Gale and Church, except that sentence lengths are compared in terms of words rather than characters. Brown didn't want to align entire articles but just a subset of the corpus suitable for further research. Thus for higher level section alignment, they used lexical anchors and simply rejected sections that did not align adequately. The report was good on at least on 1:1 alignments, but note that sometimes small passages were misaligned because the algorithm ignores the identity of words (just looking at sentence length).Wu (1994) had applied Gale and Church's method to a corpus of parallel English and Cantonese (a version of Chinese) Text from the Hong Kong Hansard. He reports that some of the statistical assumptions underlying Gale Church's model are not as clearly met when dealing with these unrelated languages, but nevertheless, outside of certain header passages. The results are not much worse than Gale and Church. To improve accuracy, Wu explores using lexical cues, which heads this work in the direction of lexical methods. Wu 's 500 sentences test suite includes one each of a 3:1, 1:3 and 3:3 alignment- alignments considered too exotic to generable by most of the methods. Church(1993) argues that length-based methods work well on clean text but may break down in real-world situations (noisy OCR or unknown markup conventions). OCR programs can lose paragraph breaks and punctuation characters, and floating material (headers, footnotes, tables, etc.) can confuse the linear order of text to be aligned. In such texts, finding even paragraph and sentence boundaries can be difficult. Church's method is to induce an alignment by using cognates. The procedure works by finding the cognates at the level of character sequences. Fung & McKeown (1994) algorithm works on the text without having found sentence boundaries, in only roughly parallel text where some sections may have no corresponding section in the translation or vice versa and with unrelated language pairs. This technique is applied on English and Cantonese(Chinese). Kay & Roscheisen(1993) use a partial alignment of lexica items to induce the sentence alignment. The use of lexical cues also means that method does not require a prior higher level paragraph alignment. Chen(1993) does sentence alignment by constructing a simple word-to-word translation model as he goes along. The method has been used for large scale alignments: several million sentences each of English and French from both Canadian Hansard and European Economic Community proceedings. He estimates an erorr rate of 0.4 % over the entire text whereas others have either reported higher error rates or similar error rates over only a subset of the text. Most of the errors are apparently due to the not terribly good sentence boundary detection method used, and further improvements in the translation model are unlikely to improve the alignments, while tending to make alignment process much slower. The presented work limits matches to 1:0, 0:1, 1:1,2:1 and 1:2 ,and so it will fail to find the more exotic alignments that do sometimes occur. Haruno & Yamazaki (1996) has worked for structurally very different language and used an online dictionary to find matching word pairs.[15]

A very little work has been done on Indian language for aligning sentences for parallel corpus. And there has no work been done for Punjabi language. This paper focuses of Indian languages – Hindi and Punjabi.

2. Major Steps in Alignment Algorithm

In order to keep high precision in sentence alignment, several steps are used with the human and computer cooperation:



Source Text is the Hindi text in this project .This text is collected from different sources like different publishers, internet etc. This means that various formats e.g html, pdf are converted to plain text files.

Target Text is the Punjabi text here which is the exact translation of the source text.

Formatting is basically the conversion of raw data collected from various sources into a plain text files. The format of these text files is in Unicode Standard. The Font used in these files is Arial MS Unicode. The texts are encoded by using UTF-8 (Unicode).

Paragraph Splitter is module that identifies the paragraph boundaries in the source and target text files. In our case it is a new line character.

Sentence splitter is used to break the paragraph of texts into sentences and the texts are tokenized for both languages. The sentence boundary is identified by the occurrence of '|' called 'viram' sign or '?' character. Sometimes the end of sentence is a new line character when sentence is not ended by the 'viram' sign.

Paragraph and sentence alignment is based on the statistical parameters of the text like length of sentence in terms of number of words in it and total length of the paragraph in terms of number of sentences in it .The score is assigned to the sentence pairs on the basis of which the decision of alignment is taken that will be discussed in the next section.

Align Source and target text means after deciding which pairs of sentences can be matched are taken and store them in the database.

Show Results means to display the stored aligned data in the database. Here we can use third part software tool to display the results stored in the database. As results stored in Mysql database can be displayed in the proper format by using the Wamp server.

3. General Principle and Algorithm

The algorithm works on the principle that a shorter sentence tends to translate in shorter sentences and a longer sentence tends to translate in longer sentences. The dynamic programming concepts are used here that are used by the Gale and Church's algorithm in which very complex statistical formulas have been used whereas it uses the basic distance functions. This distance function actually used to find the best alignment of the sentences from the various possible cases. After mapping the paragraphs, the sentences are aligned according to the minimum distance function called scoring technique. We calculate a score between all combinations of sentence pairs. The corpus texts are segmented according to the natural units, such as: paragraph, sentence and word. The Hindi and Punjabi words are simply marked by spacing as in ordinary written text. An ID is given to every paragraph to indicate the relative position in whole text. The sentence alignment type between Hindi sentence unit and Punjabi sentence unit maybe 1:1, 2:1, 3:1, 1:2, 1:3,2:2, 3:2, 2:3. Links between parallel texts are showed by attributes of Alignment

Algorithm

1. Load the bilingual text.
2. Input these two files into the match (string, string) fuction that do the whole processing of the two files.
3. Staistical Analysis of two files are done i.e the number of sentences, words and paragraphs are counted.
4. Split the source paragraphs and target paragraphs into sentences in order and create arrays of sentences in each paragraph.
5. Then one by one each paragraph of source and target files are taken and sentences of these paragraphs are matched .The matching is based on the scoring given by the minimum distance function It consider only 1:1, 1:2, 2:1, 1:3 and 3:1 type of sentences alignments.
6. According to the minimum distance function the sentences are aligned and start inserting into the database. The aligned parallel corpora are formed i.e. the collection of pairs of sentences where one sentence is a translation of other.

Scoring Technique

While calculating the scores of sentence pairs within paragraph pairs, the length of sentence is measured in terms of number of words in it. This measurement on the basis of number of words give the precise results especially between Hindi-Punjabi texts. The following parameters are used while calculating the scores between the paragraph pairs:

- ✦ Whole text lengths: (L_s (source text), L_t (target text))
- ✦ Length of sentences: (L_{si} the i-th sentence of source text, L_{ti} i-th sentence of target text)

There are number of possible cases for the alignment of the sentences of the paragraphs pairs. So finding the distance for all the possible cases .The case having minimum score is taken into consideration that it is best aligned case of all the possible cases of sentence pairs within paragraph pairs. Accordingly aligning will be done. After aligning one paragraph the next mapped paragraph is taken and again find the case having the minimum score .The same procedure above will be executed for all the mapped paragraphs.

Minimum distance function

In this function, Paragraph having greater number of sentences is assumed as Source paragraph and other one is assumed as Target paragraph. Suppose the source text i.e. Hindi Text contains paragraph consisting of 3 sentences and target text i.e. Punjabi text contains paragraph consisting of 4 sentences. S_i and T_j are the length of ith and jth sentences in terms of words for source and target text paragraph respectively.

There are three possible cases how these sentences are corresponding to each other.

Case1: When First Sentence in Source text is mapped to two sentences of target Text

S1=T1, T2; S2=T3; S3=T4

Distance1=ABS(S1-(T1+T2))+ABS(S2-T3)+ABS (S3-T4)

Case2:

S1=T1; S2=T2, T3; S3=T4

Distance2=ABS(S1-T1)+ABS(S2-(T2+T3))+ABS (S3-T4)

Case3:

S1=T1; S2=T2; S3=T3,T4

Distance3=ABS(S1-T1)+ABS(S2-T2)+ABS (S3-(T3+T4))

Min (Distance1, Distance2, Distance3) will be taken as final result.

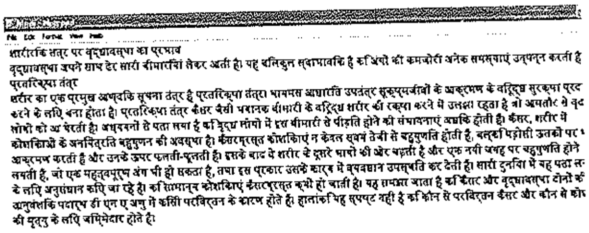
4. Collection of Parallel Corpus

Sample Parallel corpus has been arranged from CDAC (Center for development of advanced Computing), Noida for testing purposes otherwise there is no parallel Hindi Punjabi Text are available from any of the institutes, research organizations and internet. CDAC Noida is preparing this parallel text manually. In addition, Parallel Hindi Punjabi Text was arranged from book publishers from the translated version of books. Finally we were able to collect parallel text containing 20,000 sentences.

Sample Text

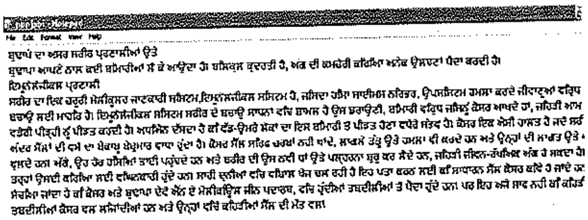
A sample text of our Hindi-English parallel Corpus in Unicode is showed below:

Hindi Text File :



शारीरिक सतर पर वृद्धावसथा का परभाव
वृद्धावसथा अपने साथ ढेर सारी बीमारियां लेकर आती है। यह बलिकुल सवाभावकि है कि अंगो की कमजोरी अनेक समसयाएं उतपन्न करती है
परतिक्रिया तंतर
शरीर का एक परमुख आणवकि सूचना तंतर है परतिक्रिया तंतर। थायमस आधारित उपतंतर सूकषमजीवों के आकरमण के वरिद्ध सुरकषा परद
करने के लिए बना होता है। परतिक्रिया तंतर कैंसर जैसी भयानक बीमारी के वरिद्ध शरीर की रकषा करने में उलझा रहता है जो आमतौर से वृद
लोगों को आ घेरती है। अधययनों से पता लगा है कि वृद्ध लोगों में इस बीमारी से पीड़ति होने की संभावनाएं अधकि होती हैं। कैंसर, शरीर में
कोशिकाओं के अनियंत्रित बहुगुणन की अवसथा है। कैंसरगरसत कोशिकाएं न केवल सवयं तेजी से बहुगुणति होती हैं, बलकि पड़ोसी ऊतकों पर
आकरमण करती है और उनके ऊपर फलती-फूलती हैं। इसके बाद ये शरीर के दूसरे भागों की ओर बढ़ती है और एक नयी जगह पर बहुगुणति होने
लगती हैं, जो एक महत्वपूरण अंग भी हो सकता है, तथा इस परकार उसके कारय में वयवधान उपसथति कर देती है। सारी दुनिया में यह पता लग
के लिए अनुसंधान कएि जा रहे हैं। कि सामान्य कोशिकाएं कैंसरगरसत क्यों हो जाती हैं। यह समझा जाता है कि कैंसर और वृद्धावसथा दोनों के
आनुवंशकि पदारथ डी एन ए अणु में कसिी परविरतन के कारण होते हैं। हालांकि यह सपषट नहीं है कि कौन से परविरतन कैंसर और कौन से को
की मृत्यु के लिए जिम्मेदार होते हैं।

Punjabi Text File:



ਬੁਢਾਪੇ ਦਾ ਅਸਰ ਸਰੀਰ ਪ੍ਰਣਾਲੀਆਂ ਉੱਤੇ
ਬੁਢਾਪਾ ਆਪਣੇ ਨਾਲ ਕਈ ਬੀਮਾਰੀਆਂ ਲੈ ਕੇ ਆਉਂਦਾ ਹੈ। ਬਿਲਕੁਲ ਕੁਦਰਤੀ ਹੈ, ਅੰਗ ਦੀ ਕਮਜ਼ੋਰੀ ਕਈਆਂ ਅਨੇਕ ਉਲਝਣਾਂ ਪੈਦਾ ਕਰਦੀ ਹੈ।
ਇਮਯੂਨੋਲੋਜੀਕਲ ਪ੍ਰਣਾਲੀ
ਸਰੀਰ ਦਾ ਇਕ ਜ਼ਰੂਰੀ ਮੋਲੀਕੂਲਰ ਜਾਣਕਾਰੀ ਸਿਸਟਮ,ਇਮਯੂਨੋਲੋਜੀਕਲ ਸਿਸਟਮ ਹੈ, ਜਿਸਦਾ ਥਾਇਮਸ ਸਾਈਮਸ ਨਿਰਭਰ, ਉਪਸਿਸਟਮ ਹਮਲਾ ਕਰਦੇ ਜੀਵਾਣੂਆਂ ਵਿਰੁੱਧ
ਬਚਾਉ ਲਈ ਮਾਹਰ ਹੈ। ਇਮਯੂਨੋਲੋਜੀਕਲ ਸਿਸਟਮ ਸਰੀਰ ਦੇ ਬਚਾਉ ਸਾਧਨਾਂ ਵਿੱਚ ਸ਼ਾਮਲ ਹੈ ਉਸ ਭਿਆਨਕ, ਬੀਮਾਰੀ ਵਿਰੁੱਧ ਜਿਸਨੂੰ ਕੈਂਸਰ ਆਖਦੇ ਹਾਂ, ਜਿਹੜੀ ਆਮ
ਵਡੇਰੀ ਪੀੜ੍ਹੀ ਨੂੰ ਪੀੜਤ ਕਰਦੀ ਹੈ। ਅਧਿਐਨ ਦੱਸਦਾ ਹੈ ਕਿ ਵੱਡ-ਉਮਰੇ ਲੋਕਾਂ ਦਾ ਇਸ ਬੀਮਾਰੀ ਤੋਂ ਪੀੜਤ ਹੋਣਾ ਵਧੇਰੇ ਸੰਭਵ ਹੈ। ਕੈਂਸਰ ਇਕ ਐਸੀ ਹਾਲਤ ਹੈ ਜਦੋਂ ਸਰੀ
ਅੰਦਰ ਸੈੱਲਾਂ ਦੀ ਵਧ ਦਾ ਬੇਕਾਬੂ ਬੇਸ਼ੁਮਾਰ ਵਾਧਾ ਹੁੰਦਾ ਹੈ। ਕੈਂਸਰ ਸੈੱਲ ਸਿਰਫ ਵਧਦੇ ਨਹੀਂ ਜਾਂਦੇ, ਆਲੇ ਦੁਆਲੇ ਤੰਤੂ ਉੱਤੇ ਹਮਲਾ ਵੀ ਕਰਦੇ ਹਨ ਅਤੇ ਉਨ੍ਹਾਂ ਦੀ ਆਕਤ ਉੱਤੇ
ਫੁਲਦੇ ਫਲਦੇ ਹਨ। ਮਗਰੋਂ, ਉਹ ਹੋਰ ਹਿੱਸਿਆਂ ਤਾਈਂ ਪਹੁੰਚਦੇ ਹਨ ਅਤੇ ਸਰੀਰ ਦੀ ਉਸ ਨਵੀਂ ਥਾਂ ਉੱਤੇ ਪਸਰਨਾ ਸ਼ੁਰੂ ਕਰ ਲੈਂਦੇ ਹਨ, ਜਿਹੜੀ ਜੀਵਨ-ਸੰਬੰਧਿਤ ਅੰਗ ਹੋ ਸਕਦਾ ਹੈ।
ਤਰ੍ਹਾਂ ਉਸਦੀ ਕਰਿਆ ਲਈ ਰੁਕਾਵਟਕਾਰੀ ਹੁੰਦੇ ਹਨ। ਸਾਰੀ ਦੁਨੀਆਂ ਵਿੱਚ ਖੋਜ ਚਲ ਰਹੀ ਹੈ ਇਹ ਪਤਾ ਕਰਨ ਲਈ ਕਿ ਸਾਧਾਰਨ ਸੈੱਲ ਕੈਂਸਰ ਕਿਵੇਂ ਹੋ ਜਾਂਦੇ ਹਨ
ਸੋਚਿਆ ਜਾਂਦਾ ਹੈ ਕਿ ਕੈਂਸਰ ਅਤੇ ਬੁਢਾਪਾ ਦੋਵੇਂ ਸੈੱਲ ਦੇ ਮੋਲੀਕਿਊਲ ਜੀਨ ਪਦਾਰਥ, ਵਿੱਚ ਹੁੰਦੀਆਂ ਤਬਦੀਲੀਆਂ ਤੋਂ ਪੈਦਾ ਹੁੰਦੇ ਹਨ। ਪਰ ਇਹ ਅਜੇ ਸਾਫ ਨਹੀਂ ਕਿ ਕਿਹੜੀ
ਤਬਦੀਲੀਆਂ ਕੈਂਸਰ ਦਾ ਕਾਰਨ ਹਨ ਅਤੇ ਉਨ੍ਹਾਂ ਵਿੱਚੋਂ ਕਿਹੜੀਆਂ ਸੈੱਲ ਦੀ ਮੌਤ ਦਾ।

Corpus Annotation Procedure:

The text is processed automatically by using tool making the annotation, alignment and manual correction easy and straightforward. Preprocessing is done for cleaning up the original files partly manually, for example, rtf., doc and pdf documents are converted to plain text files. The user can optionally give the location of the source and target files, specify the encoding for the input and output files. A text in a source language and the corresponding text in a target language are given to the alignment system. Our aim is to identify an appropriate translation for a particular sentence in the source language text among the sentences in the target language text. To do this, the source sentence is first compared with a set of probable sentences that could be the translation of the source sentence. A score of comparison is assigned for every such matching. The Automatic alignment is done which gives the percentage of sentences aligned.

5. Evaluation Method

The accuracy of the system is calculated by the following formula:

Accuracy percentage = (Number of correctly aligned sentences/Total number of sentences)*100

6. Results and Discussion

The number of sentences in parallel texts were 20,000. The average accuracy of the algorithm comes out to be 74.6 %. Basically, the accuracy is dependent upon the complexity of the corpus, more the complexity less the accuracy .Complexity means how the distribution of sentences in the target file. If any of these categories 1:2, 2:1, 1:3 and 3:1 occurs simultaneously in a one paragraph then it will be difficult for the program to align the sentences .The high frequency of these categories makes the corpus more complex. If the corpus has these types of cases individually distributed in the different paragraphs of the corpus then the results of this program is very fine. However, almost all the bilingual corpora used in research are clear (nearly without sentence omission or insertion) and literal translation bilingual texts. The performance tends to deteriorate significantly when these approaches are applied to noisy complex corpora (with sentence omission or insertion, less literal translation). The average percentage of accuracy is approximately 74.6 % for the sentence Alignment of Hindi-Punjabi text. The limitation of this algorithm is that it is unable to come up with the crossing alignment and also doesn't deal with the insertion and deletion cases. The Problem of adaptation of texts at the level the sentence is relatively complex especially in cases texts contain not only of the text, but also the elements of formatting (layout), paintings etc. If paragraphs are well-arranged in both bilingual texts, a paragraph alignment is advantageous and increases the accuracy of the alignment remarkably. So it is better to use this program for texts having well-arranged paragraphs.

7. Conclusions

Most of the researchers have worked on sentence alignment for French, German, English or Chinese, using Hansards of these countries for a reliable common bilingual database. But no such Hansard exists in Hindi-Punjabi bilingual texts. Thus we are using parallel corpus from the different sources mentioned above. The proposed algorithm uses the Gale's length based method for the sentence alignment and paragraph alignment, also the dynamic programming concept. The method is based on a simple statistic model. The model was motivated by the observation that longer regions of text tend to have longer translations, and that shorter regions of text tend to have shorter translations. The average percentage of accuracy is approximately 74.6% for the sentence Alignment of Hindi-Punjabi text. This work is very beneficial in developing Machine translation Systems. This proposed algorithm is also fairly useful for closely related language with little modifications. In the future, this work can be extended to make use of lexical constraints.

8. References

[1] Gale, W.A. and Church, K.W, *"A program for aligning sentences in bilingual corpora"*, Proc. of the 29th Annual Meeting of the ACL PP 177-184, 1991.

[2] Wu, D, *"Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria"*, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico (PP) 80–87, 1994.

[3] Robert C. Moore. *"Fast and accurate sentence alignment of bilingual corpora"*, S.Richardson (ed.), Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the

Americas, Tiburon,California), pp. 135–244, Springer-Verlag, Heidelberg, Germany, 2002.

[4] Melamed, I.D, *"A Geometric Approach to Mapping Bitext Correspondence*", IRCS Technical Report 96-22, University of Pennsylvania (1996).

[5] Weigang Li, Ting Liu, Zhen Wang and Sheng Li, *"Aligning Bilingual Corpora Using Sentences Location Information"*, Proceedings of 3rd ACL SIGHAN Workshop, PP 141-147, 1994.

[6] Chen, S.F, *"Aligning Sentences in Bilingual Corpora Using Lexical Information"*, Proc. Of 30th Annual Meeting of ACL (PP) 9-16, 1993.

[7] Kay, M. and Roscheisen, M, *"Text-Translation Alignment"*, Computational Linguistics 19:1 PP 121-142, 1994.

[8] Simard, M., Plamondon, P, *"Bilingual Sentence Alignment"*, Balancing Robustness and Accuracy. Machine Translation 13(1) PP 59–80, 1998.

[9] Karunesh Arora And V.N Shukla, *"GyanNidhi: A Parallel Corpus for Indian Languages including Nepali"* In Centre for Development of Advanced Computing,Noida.

[10] Abramowitz, M., and Stegun, I. Handbook of Mathematical Functions. US, 1964.

[11] Brown, P.; Lai, J.; and Mercer, R. *"Aligning sentences in parallel corpora"*, Proceedings, 47th Annual Meeting of the Association for Computational Linguistics, 1991.

[12] N. Collier K. Takahashi. *"Sentence Alignment in Parallel Corpora"* In centre for Computational Linguistic UMIST, 1995.

[13] Akshar Bharati, Sriram V, *"An Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information"* In International Institute of Information Technology, Hyderabad.

[14] Chris Callison-Burch, *"Statistical Machine Translation with Word- and Sentence- Aligned Parallel Corpora"* School on Informatics University of Edinburgh, 2 Buccleuch Place Edinburgh, EH8 9LW.

[15] *"Foundations of statistical Natural Language Processing"* by Christopher D. Manning

[16] http://en.wikipedia.org/wiki Natural_Language_Processing

[17] http://www.aaai.org/AITOPIC/html/nat.lang.html

[18] http://en.wikipedia.org/wiki/machine_translation

[19] http://Unicode.org

[20] http://www.e-novative.info/software/wamp.php

Author's Biography

Mr. Vishal Goyal, born in 1977, did his schooling from D. C. Model Higher Secondary School, Ferozepur Cantt, Punjab. Topped in Matriculation exams in the Ferozepur district. Completed his Secondary and Graduation from R.S.D. College, Ferozepur City. Got National merit Scholarship in Secondary class. Remained topper for all three years in graduation. Completed his M.C.A. in May, 2000 from Department of Computer Science, Punjabi University, Patiala. Awarded with merit scholarship through the study career. Started his career in software industry as senior software developer in New Delhi for two years. Then joined Doaba College Jalandhar City as regular lecturer for three years. He has been awarded with Young Scientist award in 2005 by Punjab Academy of Science for his excellent research in DNA Computing. During his service, he completed his M.Tech.(CS) through distance education from JRN Rajasthan Vidyapeeth Deemed University, Udiapur. Then he joined Department of Computer Science, Punjabi

University, Patiala as regular lecturer in February 2005. He is acting as Associate Placement officer for MCA and M.Tech. Classes. He is acting as convenor for alumni association of the department. He has presented about 17 papers in National Conferences and about 10 papers in International Conferences. He has recently visited Baba Ghulam Shah Badshah University, Rajouri, J&K for delivering lecture on Push Down Automaton and Turing Machines. He used to deliver invited talk on Information Technology at Workshop organized by IRDA for Block Development Officers. He used to deliver invited talks at various workshops, institutes. Now, he is pursuing Ph.D. in Computer Science on Part time basis at Department of Computer Science, Punjabi University Patiala under the able guidance of Dr. G.S. Lehal, Professor and Head, Department of Computer Science, Punjabi University, Patiala. His areas of interests are Natural Language Processing, Database Management Systems and Theory of Automata.

Mr. Gurdarshan Singh Sandhu, Born in 1983, passed his M.Tech (CSE) from Department of Computer Science, Punjabi University Patiala, India in 2007. He received his Bachelor Degree in Electrical Engineering from Punjab Engineering College, Chandigarh, India. Now, he is currently a Software Consultant in Fidelity information Services Pvt. Ltd [NYSE: FIS] Chandigarh, working in Dot net Technology. His area of interest is Natural Language Processing.

Detection and Correction of Page Orientation in Monochrome Textual Document Image

Vasudev .T[1] Hemantha Kumar .G[2] Nagabhushan .P[3]

ABSTRACT

While imaging a document with a scanner, it is quite possible that the document is fed into the scanner in portrait or landscape or portrait flipped or landscape flipped directions, in other words the direction of feed could be 0° or 90° or 180° or 270°. Additionally, a document suffers a skew also. Hence, correction of directional orientation followed by skew correction becomes the first phase of processing in Document Image Analysis. In this paper, we provide a solution to these problems by employing a non rotational approach in two stages. A macro level skew correction(if exists) is made to the document image in portrait or landscape directions. The direction of orientation detection and correction is done in the later stage. The performance on English text documents is presented based on extensive experimentation.

Key words: Textual document image, page orientation, skew alignment, portrait and landscape directions, non-rotational approach

1. INTRODUCTION

Document Image Analysis(DIA) involves different stages starting from image acquisition to image understanding[1-3]. Each stage has many issues to be addressed and attempted to evolve generic solutions to the problems in DIA. Many researchers are attempting over these problems and the results are converging to build generic models. Some important issues in the image acquisition stage, are, (i) skew in image, the tilt given to the document while placing the document in scanner in its normal position, (ii) orientation in image, rotation in document image due to misfeeding of document to scanner in other directions to its normal position, and (iii) bending deformation in image, the bending effect due to scanning of bound document.

Considerable amount of research is reported in literature to the above. The number of works is quite high on skew detection, Jonathan and Taylor[4] in 1998 has made a survey on skew detection work and summarized over 40 approaches on skew detection. According to this survey, majority of methods are limited to detection of skew less than 30° and very few methods upto 45°. This is because in normal situation the skew will not be more than 15°. Each of the methods discussed[4] has its own limitations and advantages. Shivakumar[5] in 2005 has suggested different approaches to detect the quantum of skew in skewed documents. All these approaches end at finding the angle of skew and suggest to rotate the document image in the anti-direction of detected skew angle. Few works are reported in literature on detection and correction of deformation in image[6-8]. Zheng Zhang[6] in 2004, Breuel[7] in 2005 and Vasudev et. al. [8] in 2005 have made attempts on detection and correction of deformations in document images.

[123]Department of Studies in Computer Science, University of Mysore Manasagangotri, Mysore, Karnataka, India – 570 006. E-mail : banglivasu@yahoo.com