ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

DEVELOPMENT AND ANALYSIS OF HINDI-URDU PARALLEL CORPUS

Mandeep Kaur

GNE, Ludhiana

Ludhiana, India

Navdeep Kaur

SLIET, Longowal

Sangrur, India

Abstract—This paper deals with Development and Analysis of Hindi-Urdu parallel corpus. One of the tasks is alignment of sentences and words in two corpus. So the eventual goal of alignment work is to find the sentences and word correspondence. For this task bilingual Hindi-Urdu corpus has been collected from the university. The algorithm takes text file as input and output of this algorithm is in XML format. The alignment algorithm uses the concept of sentence length of a text is highly correlated with the length of its translation, so we use the length of text in words. This simple approach has shown good results. An evaluation was performed based on parallel corpus from different fields and all the words were correctly aligned. In the end, we make an analysis of words having multiple translations. The variation of accuracy is depending on the corpus considered, however the method can also be useful for many language pairs (languages which are phonetically similar to each other).

Keywords - alignment; bilingual corpora; multilingual; parallel corpus

I. INTRODUCTION

In computational linguistics, a corpus is a collection of spoken or written utterances of natural languages usually access in electronic form. There are several ways of classifying corpora into different types and categories according to the properties. One ways is to distinguish between corpora that include only one language (monolingual corpora) and corpora that include several languages (multilingual corpora). Multilingual corpora can be divided into parallel and non parallel corpora. Parallel corpora are referred to as natural

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

language utterances and their translation with alignment between corresponding segments in different languages.

Parallel corpora usually contain a common source document and one or more translation of this source (target documents). Bilingual parallel corpora are sometimes called bitexts and corresponding parts within these corpora are called bitext segments. Parallel corpora are been exploited in man studies. Many application 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.

A. Alignment of parallel corpus

A parallel corpus is a text in one language together with its translation in another language. Parallel corpus is usually defined as a collection of original texts translated to another language where the texts, paragraphs, sentences and words are typically linked to each other.

Alignment of corpus is basically of three types:

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

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.

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

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. We use standard techniques for the establishment of links between source and target language segment that are explained

B. Role of automatic alignment in parallel corpus

Alignment in bilingual corpora has been an active research topic in the machine translation research groups. The subject of aligning the parallel corpora is expanding rapidly, particularly because of bottom up

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

machine translation paradigms such as example based machine translation and statistical machine translation. Text alignment is not used for the tasks such as bilingual lexicography or machine translation, but also in other language processing applications such as multilingual information retrieval (IR) and word sense disambiguation. Whilst resources like bilingual dictionaries and parallel grammars helps to improve machine translation quality, text alignment, by alignment two texts at various levels. For example documents, sections, paragraph, sentences and words help in the creation of such lexical resources.

Word alignment in bilingual corpus forms the foundation of most current approaches to statistical machine translation. Although the best performing systems are “phrase-based”, but possible phrase translation are normally first extracted from word aligned bilingual text segments.

II. OBJECTIVE

The lack of previous work on texts between Hindi and Urdu is the most prominent motivation for carrying out research in this field. The objective for this paper is alignment of parallel bilingual corpora of Hindi and Urdu. The alignment would follow from sentences alignment to an alignment of corresponding words and store them in xml file. The sentences will align such that the pair will be consisted of Hindi sentence and corresponding translated Urdu sentence. The lack of previous work on texts between Hindi and Urdu is the most prominent motivation for making a research in this field. There is a high similarity between Hindi and Urdu languages. This similarity makes one to guess intuitively that length based methods or methods using cognates will give good results for alignment of Hindi- Urdu texts. These languages are grammatically similar due to which we can align at word level .But we have to do concrete studies to prove or disprove such intuitions.

Basically our corpus is based on word alignment which allows for 1:1, 1:2 and 2:1 word alignments within a sentence. In some cases, there are certain words which, instead of being phonetically same, are divided into and represented as two words in the other language. These cases would require a 1:2 or 2:1 alignment of words.

Our first aim of the study is to see the efficiency of proposed methods for the languages in Hindi texts and make modification such that it will give better results for Hindi and Urdu parallel corpus.

Second aim is to do the analysis of parallel corpus in which we generate the automatic bilingual dictionary of aligned words and find the multiple translation words.

III. DESIGN AND IMPLEMENTATION

Basic concepts:

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

Source text is the Hindi text. This text is collected from the university. This is paragraph aligned text. The format of these files is in Unicode standard. The font used in these files is Arial MS Unicode. The texts are encoded by using UTF-8.

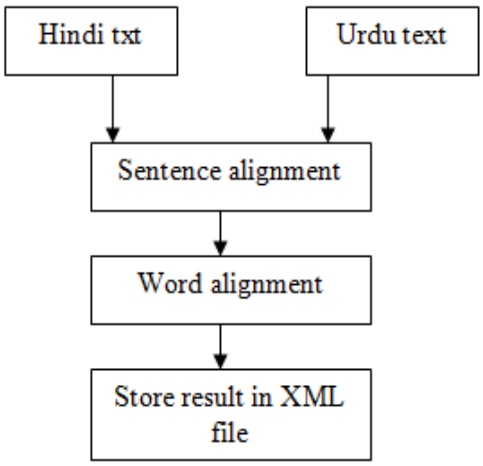
Target text is the Urdu text which is the exact translation of the source text. This text is collection from university. This is a paragraph aligned text.

Sentences alignment means deciding which pair of sentences can be the translation of each other in source and target language. In sentence alignment, we are taking care of 1:1, 1:2 and 2:1 alignment of sentences in both the languages.

Word alignment means deciding which pair of words can be the translation of each other in source and target language. In word alignment, we are taking care of 1:1, 1:2 and 2:1 word alignment .Result means after deciding which pairs of sentences and words are matched are taken and stored in XML format.

An overview of alignment algorithm:

In order to keep high precision in sentences and word alignment, several steps are used with the human and computer cooperation.



A. Sentence Alignment

Taking into account one aligned text pair at a time, the sentence alignment algorithm aligns sentences. We can very well employ length based method proposed by Gale and Church (1993) based on number of words for sentence alignment.

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

Assumptions: Assumptions for aligning sentences in Hindi and Urdu in this case are that firstly, the text being sent for sentence alignment should be aligned well; mapping of multiple sentences at once should not exceed 1:2 or 2:1. The program does not take into consideration the case where zero sentences in one case align to one or more than one sentences in another.

The algorithm starts by taking one sentence each from both the text. Then lengths of both sentences are compared according to the length relation by counting the number of words in two sentences. There can be three cases: Hindi sentence is longer than the Urdu sentence, Urdu sentence is longer than the Hindi sentence or both are similar in length.

If Hindi sentence is longer than Urdu sentence, it implies that it may refer to two sentences in the Urdu text. The thing here to check is the difference between the number of words in Hindi sentence and the number of words in the Urdu sentence is less than or equal to 20%. Then this 1:2 alignment is carried out.

If Urdu sentence is longer than Hindi sentence, it implies that it may refer to two sentences in the Hindi text. The thing her to check is the difference between the number of words in Urdu sentence and the number of words in the Hindi sentence is less than or equal to 20%. Then this 2:1 alignment is carried out

If neither of 2:1 or 1:2 alignments is possible, then simple 1:1 alignment is done.

The sentence pointers in the text are increased accordingly and the next pair of sentence is considered. This process is repeated till all the sentences of Hindi and Urdu in the paragraph are aligned. When the suitable candidates for sentence alignment are found, three things have to be done, the sentences are sent for word alignment, aligned sentences are written to the XML files and then sentence pointers with in the text are incremented accordingly.

Algorithm

1. Get the Hindi and Urdu text to be aligned
2. Set i to the first Hindi sentences and j to the first Urdu sentence in the respective text.
3. Count the number of words in both Hindi and Urdu sentence.

3.1. If the difference between the numbers of words in Hindi sentence at i^{th} position and the total number of words in the Urdu sentence at j^{th} and $j+1^{th}$ position is less than 20%, then

- Align the Hindi sentence at i^{th} position with two Urdu Sentences at j^{th} and $j+1^{th}$ positions
- Write the sentences to the output file
- Set i to i+1
- Set j to j+2
- Continue loop

Else

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

3.2. If the difference between the number of words in Urdu sentence at j^{th} position and the total number of words in the Hindi sentences at i^{th} and $i+1^{th}$ position is less than 20%, then

- Align Hindi sentences at i^{th} and $i+1^{th}$ positions with the Urdu sentence at j^{th} position
- Write the sentences to output file
- Set i to i+2
- Set j to j+1
- Continue loop

3.3. Else

- Align the Hindi sentence at i^{th} position with Urdu sentence at j^{th} position
- Set i to i+1
- Set j to j +1
- Continue loop

B. Word Alignment

One pair of sentences at a time is sent for word alignment. One thing to note is that this algorithm works well if order of words in sentences is nearly same. In case of Hindi and Urdu, which have a lot of similarity in their spoken form and have same phonetic sequences for common words, one obvious approach for word alignment could be based on cognates, as proposed by Church (1993).

Assumptions: The assumptions for word alignment are that firstly, multiple alignments should again not exceed more than 1:2 or 2:1. In cases it does, the algorithm would not perform well. Secondly, order of words in the sentence should be similar.

In word alignment, we use stop word to split the sentence into parts and generate a sentence id for each part of the sentence. Stop words are the words that are commonly occurring in the sentences. At a time one sentence id is sent for word alignment.

The process of word alignment starts in a similar fashion as sentence alignment, taking one word each from both the sentences. Again three conditions arise, one Hindi word corresponds to two Urdu words, two Hindi words correspond to one Urdu word or one Hindi word corresponds to one Urdu word. In order to consider all three, scores are calculated by counting the number of character in words of both languages and according to

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

which words are aligned. Those words are then written to the output file and word pointers within the sentence are accordingly incremented.

However, in case of Hindi characters, one character necessarily corresponds to only one sound. But in certain cases, where one character in Hindi is a "combination of two sounds", the phonetically corresponding character to it in the Urdu language is often written as a sequence of two characters instead of one character.

For example, the character "भ" in Hindi have same sound "bha" which is the combination of sound "bha" and "ha" and the corresponding character in Urdu is which is encoded as"بھ " which is composed of two characters in Urdu "ب" and "ھ". So the word length in two language are not exactly same but it can differ to some extended. The pivotal issue here is the calculation of scores by counting the no of characters in each corresponding words in Hindi and Urdu.

Algorithm

1. Get the Hindi and Urdu sentences to be aligned.
2. Spilt both the sentence into parts using stop words and generate a sentence id for each part. Then send one id at a time of both the sentences for word alignment.
3. Get i be the first Hindi word in the Hindi sentence and j be the first Urdu word in Urdu sentences.
4. Count the number of character in both Hindi and Urdu word.

. 4.1 If the difference between the number of character in Hindi word at i^{th} position and the total number of characters in the Urdu word at j^{th} and $j+1^{th}$ position is less than or equal to 20%, then

- Align the Hindi word at i^{th} position with two Urdu words at j^{th} and $j+1^{th}$ positions
- Write the words to the output file
- Set i to i+1
- Set j to j+2
- Continue loop

Else

4.2.1If the difference between the number of character in Urdu word at j^{th} position and the total number of characters in the Hindi word at i^{th} and $i+1^{th}$ position is less than or equal to 20%, then

- Align the Hindi word at i^{th} and $i+1^{th}$ position with two Urdu words at j^{th} positions.
- Write the words to the output file
- Set i to i+2

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

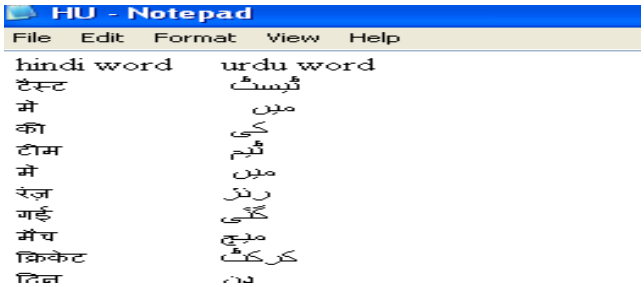
- Set j to j+1
- Continue loop

4.3. Else

- Align the Hindi sentence at i^{th} position with Urdu sentence at j^{th} position
- Set i to i+1
- Set j to j +1
- Continue loop

C. Stop words

During the process of word alignment, first we have to create a bilingual dictionary by taking the common words from both the languages which are commonly occurring in sentences. These words are called stop words. Following is the bilingual dictionary of Hindi Urdu words.



D. Project user interface

This project is based on aligning of sentences and words of parallel Hindi Urdu bilingual text. The font used for both the text files is MS Unicode. The files are read and the context of file id is splitted into sentences. The result of splitting is assigned to string .then the resulting is returned to our main part of the program which uses this for word alignment process. The interface for this program is to input the two files: one for Hindi and one for Urdu where each one is the machine translation of other. The aligner will do the sentence and word alignment of two input files and store result in XML files. Following are the two input file.

Fig. Input hindi file

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

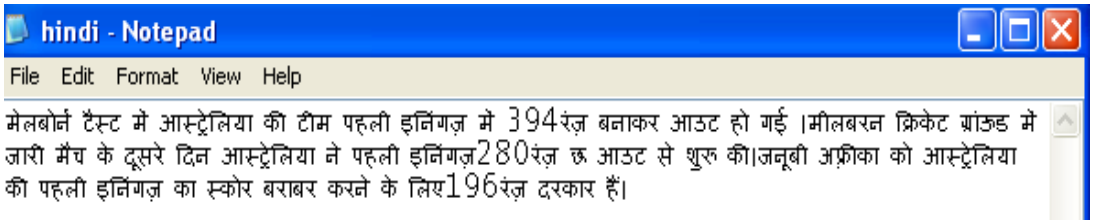
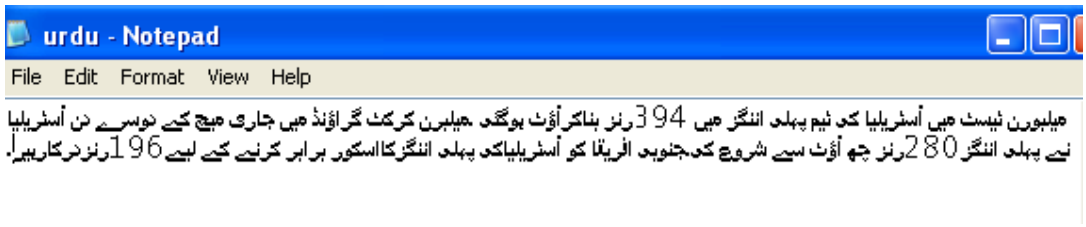
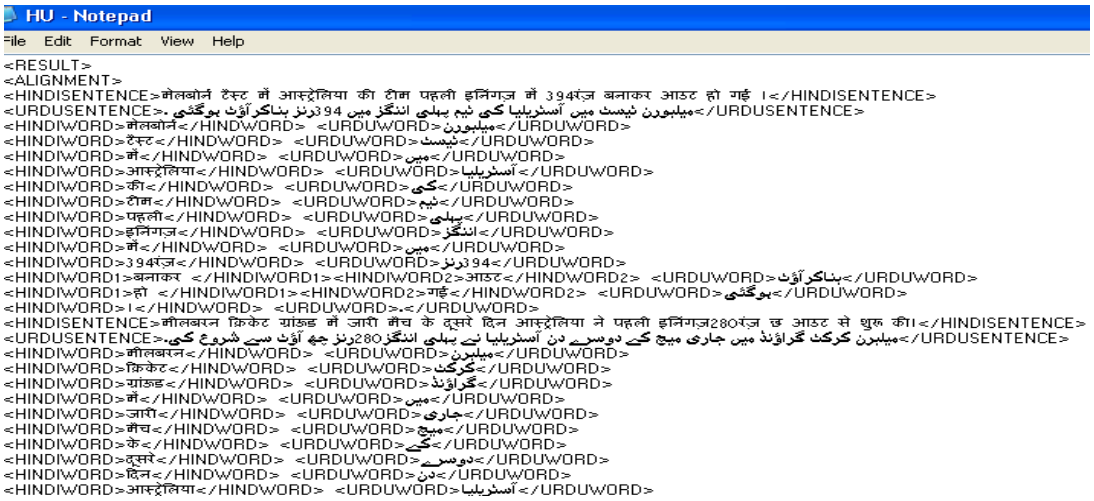


Fig. Input Urdu file



The aligner will do the sentence and word alignment of two input files and store result in XML files. Following is the result of aligner.

Fig. Final XML file



ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

IV. RESULT AND DISCUSSION

A. Result

The average accuracy of the algorithm comes out to be sentence alignment being 95% and word alignment-75.55%. Basically the accuracy is dependent upon the complexity of the corpus, more the complexity less the accuracy. Complexity means how the distribution of words is in the target file .if any of these categories 1:2 and 2:1 occur simultaneously in a one sentence then it will be difficult for the program to align the words. The high frequency of these categories makes the corpus more complex. If the corpus has these types of cases individually distributed in the different sentences of the corpus then the results of this program are very fine.

The performance tends to deteriorate significantly when these approaches are applied to noisy complex corpora. The problem of adaptation of texts at the level the word is relatively complex especially in cases where texts contain not only of the text, but also the elements of formatting (layout), painting etc. if sentences are well-arranged in both bilingual texts, a sentences alignment is advantageous and increase the accuracy of the alignment remarkably .so it is better to use this program for texts having well-arranged sentences.

B. Conclusion

Most of the research has work on sentence and word alignment for French. German, English or Chinese, using hansards of these countries for a reliable common bilingual database. But no such hansards exits in Hindi Urdu texts. Thus we are using parallel corpus from corpus mentioned above. The proposed algorithm uses the gale's length based method for the word alignment and sentences alignment. The method is based on a simple statistic model. The model was motivated by the observation that the longer regions of text tends to have longer translation and that the shorter region of text tend have shorter translations. This work is very beneficial in developing bilingual dictionary and machine translation system. Thus our objective of the project is achieved. This proposed algorithm also is also fairly useful for closely related language with little modifications.

C. Furure Scope

As an emerging research, there is so much of room for further development. In future, we would hope to extend the method by making use of linguistic information. The basic word alignment methods works on the word level of the plain text. Some discriminative methods are proposed to integrate various syntactic and lexical clues into the alignment models to improve alignment model to improve alignment quality.

REFERENCES

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V2I6M3-112012
VOLUME 2 ISSUE 6 November 2012

[1] D. Wu. 1994. 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.

[2] D. Melamed, "A Geometric Approach to Mapping Bitext Correspondence," Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA, 1996a.

[3] Dengjun Ren, Hua Wu,Haifeng Wang, ”Improving Statistical Word alignment with various clues” In proceeding of MT SUMMIT XI, Copenhagen, Denmark pages 391-397,2007

[4] Gale, W. A. and K. W. Church , (1993) "A program for aligning sentences in bilingual corpora," Computational Linguistics, vol. 19, pp. 75-102, 1993.

[5] Moore,R.C., "Fast and Accurate Sentence Alignment of Bilingual Corpora," AMTA 2002, pp. 135-144, 2002.

[6] Melamed,I. D.,"Bitext Maps and Alignment via Pattem Recognition," Computational Linguistics, 25(1), pp.107-130, March, 1999.

[7] Melamed,I. D., "Bitext Maps and Alignment via Pattern Recognition,"Computational Linguistics, 25(1), pp.107-130, March, 1999.

[8] XU Yang1, WANG Hou-feng1, LÜ Xue-qiang2 “Research of English-Chinese Alignment at Word Granularity on Parallel “Peking University, Beijing 100871, CHINA.

[9] http.//Unicode.org

[10] http://en.wikipedia.org/wiki/Natural_language_Processing