Enhancing effectiveness of sentence alignment in parallel corpora: Using MT heuristics

by

Sriram Chaudhury, Dipti Misra Sharma, Amba P Kulkarni

in

ICON-2008: International Conference on Natural Language Processing (ICON-2008)

Report No: IIIT/TR/2008/187

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2008
Enhancing effectiveness of sentence alignment in parallel corpora: Using MT & heuristics

Sriram Chaudary
Language Technologies Research Center
International Institute of Information Technology
Hyderabad, India
sriramchaudhury@gmail.com

Kiran Pala
Language Technologies Research Center
International Institute of Information Technology
Hyderabad, India
kiranap@gmail.com

Lakshminarayana kodavali
Aurora's Technological and Research Institute
Hyderabad, India
kodavali.lakshmi@gmail.com

Keshav Singhal
Shri vaishnav institute of technology and science.
Indore, India
mr.keshavsinghal@gmail.com

Abstract

India is a multilingual, linguistically dense and diverse country with rich resources of information. Parallel corpora have major role in multilingual natural language processing, computational linguistics, speech and information retrieval. This paper describes an alignment system for aligning English-Hindi texts in GyanNidhi corpus at sentence level. The criteria used for alignment is combination of linguistic, statistical information and simple heuristics. We use multi-feature approach with Anusaaraka (Machine Translation System), Hindi shallow-parser, Hindi WordNet lookup as primary technique with resources of target language to increase the level of alignment accuracy. Other features such as Named Entities, linguistic information, notation converters are used to match the words in between one-to-many bilingual sentences. Our experiments are based on the GyanNidhi corpus. We obtained 92.06% accuracy for English- to-Hindi sentence alignment with 95.68% precision and 88.09% recall for one-to-many sentence alignment. The study also suggests procedures for aligning parallel translated corpora by using a machine translation system.

1 Introduction

Parallel texts (corpora) are useful resources for acquiring a variety of linguistic knowledge resources (Matsumoto.Shimoto.ll.; and Utsuro.T, 1993). Corpus based text alignment is not only used for building lexicons or machine translation systems but also for other language processing applications such as multilingual information retrieval and word sense disambiguation. At the same time resources like bilingual dictionaries and parallel grammars help improve Machine Translation (MT) quality, and aligning two texts at various levels (i.e. documents, sections, paragraphs, sentences and words) helps in the creation of such lexical resources (Manning & Schtze, 2003). Aligning a parallel corpus means making each sentence of the source corpus correspond to an equivalent sentence of the target corpus. The corpus is a reliable source of database for the required language processing methods and a good source of reference for the linguistic analysis. The parallel corpus has been prepared in view of common translation practices. All these observations and conditions lead us to prepare an aligned corpus at sentence level and this should be treated as a lexical database consisting of parallel sentences.

In this paper we have described the procedures for sentence level alignment based on the GyanNidhi Parallel corpus (Arora et al). Motivation behind the system is to prepare a sentence aligned parallel corpus for research community. In the following sections we will describe the procedure using Anusaaraka, linguistic rules, along with Hindi WordNet and Shallow parser as major tools.
2 NLP Tools

2.1 Anusaaraka

Anusaaraka is an approach to Machine Translation based on Information Dynamics, inspired by Panini’s grammar. In this work we have used English-Hindi Anusaaraka MT system. (Anusaaraka source).

2.2 Hindi Shallow Parser

The Hindi shallow parser (Hindi Shallow Parser source) gives the analysis of a Hindi sentence at various levels. The analysis begins at the morphological level and cumulates at results of POS tagger and chunker. The final output combines the results of all these levels and shows them in a single representation (called Shakti Standard Format). Apart from the final output, one can also view the output after each level.

2.3 Hindi WordNet

The Hindi WordNet (Hindi WordNet source) is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles.

3 Sentence Alignment

The algorithm that is proposed does the sentence alignment using lexical information of the target language. The algorithm first passes the source language paragraph through Anusaaraka (Machine Translation system from English to Hindi). The source paragraph is broken into sentences and translated to target languages. The target language paragraph is also broken into sentences. Finally we will get both source and target languages in the form of sentences in the target language. The proposed algorithm uses shallow parser to extract the root words from translated source language and target language text for comparison. Scores were assigned according to number of matched possible words between translated source language and target language for each comparison. The algorithm then carries out the alignment of sentences using these scores. The precision of the alignment is 95.68%. The output of the system is shown in Figure-2.

3.1 Previous Methods

To align the parallel corpus various techniques were developed a decade ago; Brown et al. (1991) algorithm is dependent upon the length of the sentence in terms of words and Gale and Church (1993) algorithm is dependent on the length of the sentence in terms of characters, these are purely length-based approaches which have no concern with any meaning of the words, these are considered as linguistically poor but computationally better. If we consider algorithms based on lexical approaches, Kay and Roscheisen (1993) tried many possible lexical correspondences for sentence alignment in bilingual part. They have tried it on Scientific American articles and achieved 96% of accuracy. Similarly, based on the number of frequency of words in source language and the number of possible occurrences of source language words in target language, Warwick et al. (1989) have considered bilingual-dictionary to construct word pairs list. Mayers et al. (1998) propose an approach that is the basis of a MRD (Machine Readable Dictionary). They have tried on English-Japanese sentence alignment one-to-one and two-to-one, and they used bilingual dictionaries which contain root forms for the above experiment, and they preprocess the data to find the root form for each word in a sentence. More or less similarly other models are based on POS taggers (Haruno and Yamazaki 1996), and one of the methods has developed an algorithm based on cognates (Wu 2000). These are sophisticated techniques which use lexical information; eventually it is an expensive process and requires more lexical resources for both source and target languages, but in the scenario of Indian language processing monolingual resources are available in more quantity than the bilingual resources. Here in this method we are aligning sentences by comparison after translating source language text into target language.

3.2 Corpus

GyanNidhi is the first attempt at digitizing a corpus which is parallel in multiple Indian languages. The English and Hindi common books are extracted and the paragraphs are aligned. For testing and
evaluation, we used this paragraph aligned English-Hindi parallel corpus.

4 Framework

The detailed framework of the sentence aligner tool is described in the Figure 1. Workflow of the sentence alignment tool.

4.1 Assumptions

In Hindi Paragraph (HP), each line is separated by “|” or “?”. Paragraphs are separated with a blank line in both English and Hindi text.

If a sentence in English Paragraph (EP) is split into two sentences in HP, then these two Hindi Sentences (HS) must be sequential. The same rule applies when merging English sentences to form Hindi sentences.

4.2 Heuristics

1. Assigning less weightage (0.01) for grammatical words (e.g: kA, ke, Ora, hI, Apa...) matched between Translated English Sentences (TES) and HS.

2. Assigning more weightage (2.0) for Named Entities words (e.g: Hyderabad, Rama, Sita, IIIT,), Digits (e.g. 1947, 20,) and single frequency words in both Translated English Paragraph (TEP) and HP.

3. Assigning weightage 5.0, if word is Named Entities as well as single frequency in both TES and HS.

4. ES with scores less than (1.0) will not be matched

5. Two HS will be merged only if following conditions are satisfied.

   a. If length of TES(i) is more than length of HS(j)
   b. Difference in length of ES(i) and HS(j) is > 30%
   c. Atleast two non-grammatical common words should exist in between ES(i) and HS(j).
   d. Difference between length of ES(i) and length of HS(j)+HS(j+1) < 20%
   e. Atleast two non-grammatical common words between ES(i) and HS(j+1).

6. To merge two ES the conditions mentioned in step-5 should be satisfied such that they are applied to ES instead of HS.

7. If ES(i) is matched with HS(j) and ES(i+2) is matched with HS(j+2), then ES(i+1) should match with HS(j+1).

8. If ES (i) is matched with HS(j) and number of English Sentences (NES) and number of Hindi Sentences (NHS) are such that |i - j| > |N ES - N HS| then match ES(i) with next best score of HS.

4.3 Algorithm

Input: Set of English Paragraph (EP) and Hindi Paragraph (HP) in different files.

Output: Aligned English sentences (ES) and Hindi sentences (HS) with matching score for each paragraph.

begin
| Convert English text to Hindi using Anusaaraka.
| Convert Anusaaraka output and Hindi text into WX-notation.
| Use shallow parser (for root word comparison), Hindi Wordnet (HWN) (for Hindi synonyms) for finding equivalent words if present between TES and HS.
| Initialise: max score=0.0; Elist=[]; Hlist=[];
| *foreach* Para P of (TEP,HP) *do*
| TESlist=Split (TEP)
| HSlist=Split(HP)
| *foreach* TES i of TESlist *do*
| Elist.append(i)
| *foreach* HS j of HSlist *do*
| Compare words between i and j and calculate score
| if (Score > Max score) then
| hmax=j
| diff=abs(len(i)-len(j))
| end
| else if (Score == Max score)
| and (abs(len(i)-len(j)) > diff)
| then
| hmax=j
| end
| Hlist.append(hmax)
| end
| *foreach* E,H of list(Elist,Hlist) *do*
| Plist.append(ES,HS) Apply heuristics rules and realign the pair list.
| end
| end

Algorithm 1
5 Evaluation

5.1 Accuracy

The evaluation was done on GyanNidhi corpus. We had taken about 66 paragraphs randomly, with the variation of 3 to 19 numbers of sentences in each paragraph. The number of sentences on which experiment is done is 504, out of which 464 are aligned by the proposed algorithm in which 444 are aligned correctly. The accuracy of the algorithm was measured as 92.06%.

\[
\text{Accuracy} = \frac{\text{Number of aligned Sentences}}{\text{Total no of Sentences}}
\]
The accuracy can be visualized by the graph given in Figure-3, taking percentage scale on y-axis and number of test cases on x-axis.

5.2 Precision

\[
\text{Precision} = \frac{\text{Number of correctly aligned sentences}}{\text{Number of aligned sentences}}
\]

Precision was calculated as 95.68%. Graph of variation in precision can be seen on Figure- 4.

5.3 Recall

\[
\text{Recall} = \frac{\text{Total number of correct aligned sentences}}{\text{Total number of sentences in source}}
\]

Recall was calculated as 88.09%. The variation can be seen in the graph plotted between number of cases on x axis with corresponding percentage recall on y-axis. The graph for recall is given on Figure- 4.

5.4 Error Analysis

\[
\text{Error} = \frac{\text{Total number of not aligned English sentences}}{\text{Total number of English sentences}}
\]

Error rate was calculated as 11.90%. Taking the average of error percentage corresponding to each test case will give error count equal to 10.6%. The difference between this two results is due to the varying length of paragraphs.

Identified sources of errors:
- Due to not much difference between number of words of source and target language sentences merging not taking place.
- Due to lack of sufficient number of matching words from Anusaaraka translation and WordNet.
- Due to unavailability of corresponding sentence in target text (Incorrect corpus).
- Due to input to Anusaaraka in unrecognized format.

The sources of error in this system can be reduced in future by improving translation quality of Anusaaraka and by applying more heuristics with the fact that always perfectly aligned paragraphs should be given as input to this system.

5.5 F-Measure

\[
F\text{- Measure} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}
\]

F-Measure was calculated as 91.72.

6 Conclusion

We have developed a practical sentence aligner for English-Hindi bilingual corpora by combining state of the art MT, rich NLP tools and simple heuristics. This unique approach gave high recall and good accuracy even for distinct languages such as English and Hindi. The sentence aligned parallel corpora will facilitate the development of word alignment tools and can be further used to enhance the statistical MT results. Building of lexical resources such as bilingual dictionary, phrasal dictionary can be done in a cost effective manner. The future work includes improving the accuracy and precision of the system and using the same approach for other language pairs.

Acknowledgments

We are grateful to Prof. Vineet Chaitanya for his consistent guidance throughout the experiment. Special thanks to the consortium project for giving GyanNidhi corpus.

References

Gale and Church "A Program for Aligning Sentences in Bilingual Corpora" Computational Linguistics, also presented at ACL-91

Stanley Chen "Aligning Sentences in Bilingual Corpora Using lexical Information" ACL-93.
Church, Gale and Dagan "Aligning Parallel Texts: Do Methods Developed for English -French Generalize to Asian Languages?" Roeling 1993.


Arora.K.K. and Shukla.S.A.V.G.V.N. and Agrawal.SS "GyanNidhi: A Parallel Corpus for Indian Languages including Nepalii".

Kumano.A. and Hirakawa.H. "Building an MT dictionary from parallel texts based on linguistic and statistical information” Proceedings of the 15th conference on Computational linguistics-Volume1, Association for Computational Linguistics Morristown, NJ, USA.


Manning C. and Schtze H "Foundations of Statistical Natural Language Processing” MIT Press, Cambridge,Massachusetts.


Haruno M. and Yamazaki T "Highperformance bilingual text alignment using statistical and dictionary information.” Proceedings of the 34th conference of the Association for Computational Linguistics, pp.131- 138, Santa Cruz, California.


Dagan.I, Itai.A.and Schwall.U. "Two languages are more informative than one” Proceedings of the 29th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA


Hindi Wordnet source ”Hindi Wordnet source http://www.cfilt.iitb.ac.in/wordnet/webhwn/.

Anusaaraka source ”Anusaaraka source” http://ltrc.iiit.ac.in/amusaaraka.

Hindi Shallow Parser source ”Hindi Shallow Parser source” http://ltrc.iiit.ac.in/analyzer/

Brill Eric “Transformation Based ErrorDriven Learning and Natural Language Processing” A Case


Simard Michel, Plamondon Pierre ”Bilingual Sentence Alignment: Balancing Robustness and Accuracy” Universite de Montral.