International Journal of Management, IT & Engineering
Vol. 8 Issue 11, November 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

DEVELOPMENT OF ENGLISH –PUNJABI TRANSLITERATION USING UNICODE CHARACTER MAPPING AND PARALLEL CORPUS

Yogeshkumar MCA [*]

ABSTACT- In this research paper, we have worked upon the problem of "development of English-Punjabi parallel corpus and Unicode Character Mappingusing existing English-Punjabi machine transliteration system and using sentence alignment". The alignment is based on the length and location based technique. We will use English-Punjabi machine transliteration system. These tasks are need to English-Punjabi parallel corpus and Unicode Mapping. Sentence alignment is useful for developing English-Punjabi parallel corpus and English-Punjabi dictionary. The accuracy is basically depending upon the complexity of the corpus and correspondence mapping, 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 paper is to develop English-Punjabi parallel corpus and Unicode Mapping using latest and existing techniques and method with a high accuracy and time efficiency.

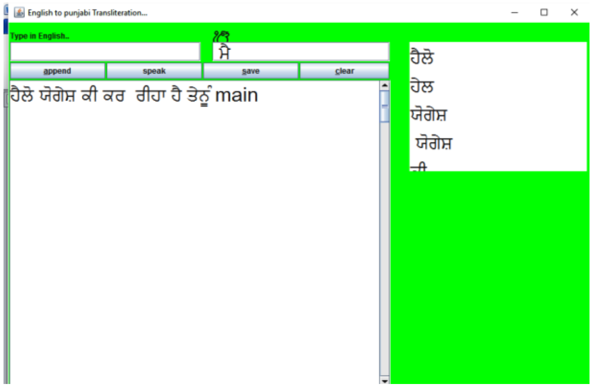
[*] **Assitant professor**

INTRODUCTION

A parallel corpus is a corpus that contains a collection of original texts in language L1 and their transliterations into a set of languages L2 ... Ln and An array of Punjabi Unicode . In most cases, parallel corpus contains data from only two languages where texts, sentences, and words are linked to each other. Alignment of corpus is mainly of three types: Sentence-wise,Paragraph-wise, Word-wise.

Sentence-wise: It is the identification of the corresponding sentencesin both halves of the parallel text. Alignments of parallel corpus at sentence level are prerequisite for many areas of linguistic research. During transliteration, sentences can be split, merged, deleted, inserted or changed in order. Mainly the shorter sentences are aligned with shorter sentences and longer sentences are aligned with longer sentences.

Word-wise: It is the identification of the corresponding words in both halves of the parallel text. Automatic alignment of word means that without the human interaction the parallel corpus should be aligned with the machine accurately. An example of an alignment between an English-English sentence pair, blue links indicates alignment of cognates



ਮੈ

ਹੈਲੋ ਯੋਗੇਸ਼ ਕੀ ਕਰ ਰੀਹਾ ਹੈ ਤੇਨੂੰ main

ਹੈਲੋ
ਹੇਲ
ਯੋਗੇਸ਼
ਯੋਗੇਸ਼

Fig 1 Transliteration from English-Punjabi

TECHNIQUES USED IN DEVELOPMENT OF ENGLISH-PUNJABI PARALLEL

- **Based on Length-** short sentences will be translated as short and long as long sentences.
- **Based on Location**-Beads of sentences in the two texts have same positions, these methods do not attempt to align beads of sentences but rather it align position using two parallel texts.

PROBLEM DEFINATION AND OBJECTIVE

The problem definition of our project is to “Development of English-Punjabi parallel corpus and Unicode Mapping by using existing English to Punjabi machinetransliteration system and using sentence alignment”. A collection of pairs of texts in two different languages where each member of pair is a transliteration of other is called a “Parallel Corpus”. The objective of our research is to Develop English-Punjabi Parallel corpus using existing English to Punjabi machine transliteration system. The development of English-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 EnglishPunjabi 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 Array of Punjabi Unicode for English to Punjabi machine transliteration system.

RELATED WORK

Researchers have worked for Punjabi(Indian language 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 transliterations of each other. A parallel text consists of a source language text & its transliteration into some target language.

D.Wu, (1994) has developed Chinese and English parallel corpora at 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 based on length methods give satisfactory result even between unrelated languages which is a surprising result. Next, it shows the effect of adding lexical cues to a based on length methods. According to these results, using lexical information increases accuracy of alignment from 86% to 92%.

- Brown et al., 1991 Similar approach as Gale and Church, in this sentence lengths are compared in terms of words rather than characters. Other difference in goal: Brown et al. didn’t want to align full articles but just a subset of the corpus relevant for further research.

Sheng et al., 1994 It uses the length of sentences as well as length of texts, the length of upper and lower part of the candidate sentences, and some information like that to emphasize the effect of location of sentences in the text. In this sense it can be said that it is a next step of pure based on length method.

International Journal of Computer Applications (0975 – 8887) Volume 5– No.9, August 2010 18 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 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 approximate the same amount of training data for each language span.

Kay &Roscheisen, 1993 Idea: Use word alignment to help regulate sentence alignment. Then use sentence alignment to refine word alignment. Method: 1. Begin with Istand Last of text as anchors 2. Form ancasting 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 span 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.

Haruno& Yamazaki, 1996 their technique is a variant of Kay &Roscheisen (1993) with the following differences: For structurally very different languages, function words impede alignment. They dispose 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 same word pairs.

Zhonghuaxiao, Tony McEnery (2002) has made 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. Our experience in collaborating with the Xara team also tells us that the cooperation between corpus creators and software 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.

METHODOLOGY

Existing system of Google transliteration did not provide better results for Punjabi font. In my paper I have rectified these things by using Punjabi Corpus and Unicode mapping which provide better results than Google translator.

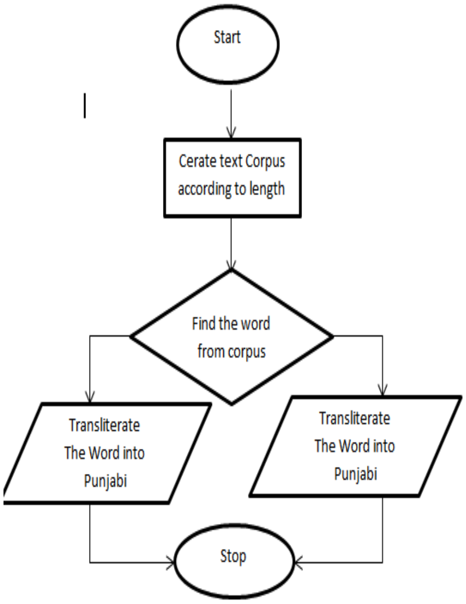

Start
Cerate text Corpus according to length
Find the word from corpus
Transliterate The Word into Punjabi
Transliterate The Word into Punjabi
Stop

Fig 2. Methodology

CONCLUSION

In existing system length of corpus word system was very small ranging from 1- 10 words. I have designed a corpus word system which can transliterate words upto 25 character long and also it provide accurate transliteration of English words into Punjabi. In its future we can increase the length of character more than 25.

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.

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

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

[6] 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).

Author's Detail:

Yogesh Kumar, MCA (ComputerApplication)Assistant Professor at Sachdeva Girls College Kharar Mohali (Punjab). I have done my research in Transliteration of English–Punjabi by using Unicode mapping and English-Punjabi corpus.