International Journal of Engineering, Science and Mathematics
Vol. 7 Issue 4, April 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@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

BuildingEnglish-Punjabi Parallel corpus for Machine Translation

Shishpal Jindal Ph.D. Research Scholar, IKG Punjab Technical University Kapurthala, Punjab, India

Vishal Goyal
Department of Computer Science, Punjabi University, Patiala, Punjab, India

Jaskarn Singh Bhullar
Department of Applied Sciences, MIMIT, Malout, Punjab, India

Abstract—**Objectives:**Parallel corpus is the key resource for English Punjabi machine translation. At wide level there is no availability of English-Punjabi Corpora. There is a primary requirement of parallel corpus for the training of statistical machine translation. **Methods/Analysis:**In this paper, our work focuses on building English-Punjabi corpus at large scale. It posed difficulties and the intensive labor to develop the corpus. We are intricate on the collection as well as the flow of work for the construction of parallel corpus. Now after getting the raw text, we need to refine the corpus in such a way that every source language sentence should have corresponding target language sentence. **Findings:** The paper attempts to explore existing tools as well as building new tools. One of the goals is alignment of bilingual corpus. The alignment algorithms are used to tune the sentences. The accuracy depends on the type of corpus. **Novelty/Improvement:** A cautious endeavor has been made to capture different types of texts.

Keywords—bilingual corpora, Machine-translation, English, Punjabi, NLP**.**

I. INTRODUCTION

Parallel corpus is the collection of text from two or more languages. It has great importance in natural language process (NLP). Corpora is a linguistics term, corpora relates to a collection of sentences, of a specificlanguage. A collection of texts of two languages is called parallel corpus [1-2]. In parallel corpus one is original language and the otheris the translation of original text. Parallel corpora plays very significant role in the field of machine translation, and linguistic studies.Electronic form of text came late, into Indian languages. The importance of corpus came into fore in ICT. With the incorporation of Information Technology in each and every field, the e-content of Indian languages gradually started growing. There have been attempts to develop parallel corpus in Indian languages [3]. The quality of results based on the number of sentences presented in the corpora. The alignment at sentence and word levels makes parallel corpora very useful from accuracy point of view. But, the parallel sentences of English-Punjabi language pair, is not available in bulk. Hence English sentences have to be collected and manually translated into Punjabi in order to create bilingual corpus [12-13]. In order to give considerable support, to research in related fields, we have started collecting and compiling the English-Punjabi corpus. So, far over 2.5 lakh sentences of English texts of different types and corresponding Punjabi texts has been collected and included into the corpus.The corpus has been aligned and verified manually at sentence level. The analysis of the translated results indicates the language specific differences between the language pair. We will show our work towards developing and utilizing the corpus.As English is a position restrictive language,the word order plays a very important role [4]. The sentence order in English is subject-verb-object. English language has very limited inflection of words.On the other hand, Punjabi follows SOV word order. This paper describes the collection of the sentences and its role to the function of statistical machine translation. The results and performance shows the challenges for SMT for English Punjabi language pair.

II. LITERATURE SURVEY

In general, the previous work on parallel corpus was focused on the multilingual corpus. There have been no public projects, which are based on the extraction of parallel corpus. Some of the projects build parallel corpus from Wikipedia [6]. There are two approaches to align the sentences to build the corpora (Dutch English). The lexicon was compared with the sentences to compute the lexical similarly between the actual and translated texts. Gale and Church (1991a) represented the co-occurrence of bilingual text and later (1991b), as Brown *et al*, (1991), they presented the correlation in length of characters and words, to compare equivalence of sentences in the given corpora [7]. Some other co-efficient were considered like intuition that words in different languages which are orthographically same and used for the alignment (Church 1993; McEnery and Oakes, 1995). The parallel corpus created by our approach is the improved form of previously mentioned flaws in English-Punjabi corpus. We deduce that the development of corpus by using existing techniques will very helpful to support the task of machine translation by reducing the dependency on human generated corpora. The air of parallel corpus to build statistical machine translation, if the corpora is alreadyexisting, then there is no significance to generate the parallel corpus [8]. Actual basic data is not

International Journal of Engineering, Science and Mathematics
Vol. 7 Issue 4, April 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@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

available for regional languages pairs. We take help from the existing techniques to extractparallel sentences from different resources, to generate the English-Punjabi corpus.

III. CORPUS COLLECTION

The acquisition of English-Punjabi text requires following five steps to generate bilingual corpus.

- Obtain the raw text from different resources.
- Extract the English and Punjabi chunks of sentences and words.
- Break the paragraph and large sentences into small sentences.
- Create the corpus according to SMT system.
- Sentence alignment of the corpus.

IV. RESOURCES OF CORPUS

Our goal is to produce good quality parallel corpus for English-Punjabi language pair. To achieve such corpus, we leveraged some sources of corpora. The English-Punjabi corpus is very less resourced language. Besides, most of most of the sentences in parallel corpus are not in the refined form, so it needs editing and proofing (Choudary, 2010). As the said corpus is not available as per requirement, so it has been developed from the scratch.We decided to collect English Punjabi text from wide range of resources.For the collection of corpus, we first tried to track the text online. But then we realized that most of the text is available in English Hindi. The Hindi text is translated into Punjabi to build the required corpora (English-Punjabi).The diversity of corpus is to be considered to introduce different types of genres and gets reflected in the corpus. We analyze the domain of text, before it is included into the corpus.Following are the resources from where the text is extracted.

Table 1. Resources of Corpus

Name of resource	Corpus code	No. of Sentences	Percentage
EMILLE	C1	160000	64
Gyan Nidhi	C2	15000	6
TOURISM	C3	7000	2.8
Health	C4	7000	2.8
BIBLE	C5	12000	4.8
GURU GRANTH SAHIB	C6	16000	6.4
PSEB Books	C7	1000	0.4
BILINGUAL NEWS PAPERS	C8	2000	0.8
NAMED ENTITY	C9	30000	12
TOTAL	C10	250000	100

(A) Gyan Nidhi (B) EMILLE (C) TOURISM (D) Health (E) BIBLE (F) GURU GRANTH SAHIB (G) PSEB Books (H) BILINGUAL NEWS PAPERS (I) NAMED ENTITY.

Total number of sentences as well as percentage of every resource is mentioned in Table 1. Unique code is given to every resource.

A. Emille

The EMILLE Corpus is developed as unit of a joint venture between the EMILLE framework of Lancter University and the CIIL, Mysore, India. The corpora has three parts: monolingual, parallel and annotated. The EMILLE monolingual corpora has 92,799,000 wordsin which text of spoken data of Bengali, Gujarati, Hindi, Punjabi and Urdu languages. The parallel corporahas 200,000 words of English and regional languages such as Hindi, Bengali, Punjabi, Gujarati and Urdu. We have isolated English Punjabi text from the above corpus.

B. Gyan Nidhi

Million pages' multilingual parallel text corpus in English and 11 Indian languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Marathi, Malayalam, Oriya, Punjabi, Tamil & Telugu) based on Unicode encoding. The Gyan Nidhi corpus contains the text in the form of books. In these books there were number of diagrams, figures, charts and other special symbols. These are removed from the text by using automated and manual tools. The text in gyan nidhi is in the form of paragraphs, that are converted into short sentences.

C. Tourism

The corpus is drawn from web covering tourism/travel and there are total 7000 sentences comprised of very simple, simple, complex and compound sentence structures. The sentence structures include relative clauses, complement clauses, finite and non-finite conjunctions. Text encoding is UTF-8. The corpus was available in English-Hindi language pair that is converted in the form English-Punjabi.

D. Health

English-Punjabi Parallel Health Text corpus is developed in Unicode under English to Indian Language Machine Translation (EILMT) Consortium. This corpus is created in excel format and size of the corpus is 7000 sentences.

E. Bible

The complete set of 66 books contains approximately 800k words in English. This might seem small compared to modern (parallel) corpora—like, for example the Canadian Hansards corpus (Germann, 2001) with nineteen million words, and the Europarl (60 million words on average each language); however it is much huge than individual literature: for example, the size of the average novel is about one hundred thousand words.Majority parallel corpora exist in a small number of languages or in common languages pairs [10]. It was available in English Hindi that is converted into English

International Journal of Engineering, Science and Mathematics
Vol. 7 Issue 4, April 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@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

Punjabi format by using Hindi to Punjabi machine translation system.

F. Sri Guru Granth Sahib

The Sri Guru Granth Sahib corpus is available in electronic form. It was not very difficult to create the English Punjabi parallel corpus. We eventually were able to create a second sample corpus of written modern Punjabi. This sample corpus consists of 16000 sentences.

G. PSEB Books

E-books Punjab School Education Board were download of some subjects, that was in PDF format. Sodhak typing pad, font converter and SpellChecker was used to create the English Punjabi text.

H. Bilingual News Paper

Bilingual newspaper is used in the development of corpora. This newspaper is also in English Hindi form. The files were in image format. The were images converted into text format and the Hindi text converted into Punjabi language.

I. Named Entities

Named entity recognition as entity identification, entity chunking and entity extraction is a subtask of machine translation. Named entities in text such as the names of persons,organizations,locations,expression of times, quantiles,monetary values. These entities are to be transliterated and not to be translated.

V. SYSTEMATIC WORKFLOW IN BUILDING OF PARALLEL CORPUS

To assist the building of parallel corpus, we developed the sequence of steps for the construction of parallel corpus. Following steps are to be applied on the collected corpus, before they entered into refined corpus.

A. Noise Removing

Majority of our data has been downloaded from Internet and therefore the data is in the form of HTML files, where many tags and other irrelevant information exists. So removing of tags and irrelevant information from the data is known as noise removing. After the removal of noise, text becomes a clean text [9].

B. Textual attribute Tagging

Textual attributes are tagged in the text. Our goal is semantic tagging of textual content with the help of Notepad++, to extract the text from the tagged files.

C. Text Alignment

Alignment tools are used to align the text at paragraph and sentence level. Hunalign sentence aligner is used to align bilingual text on the sentence level.

D. Human Verification

With intervention of Humans, the results of aligned text are verified and errors are corrected to produce the high accuracy of data.

E. Bilingual text

The verified bilingual textis stored into the spreadsheet. It is easy to track the sequence number of sentences.

F. Segmentation and POS tagging

Automatic tools are used for segmentation and tagging of English text.

G. Indexing the texts

Indexing of processed text isdone in this step.

VI. IDENTIFICATION OF PARALLEL DOCUMENTS

We did a content-based comparison of English-Punjabi documents that were collected from the different resources. The process of filtering isolates the document in different categories i.e. Hindi,Punjabi and Html files.

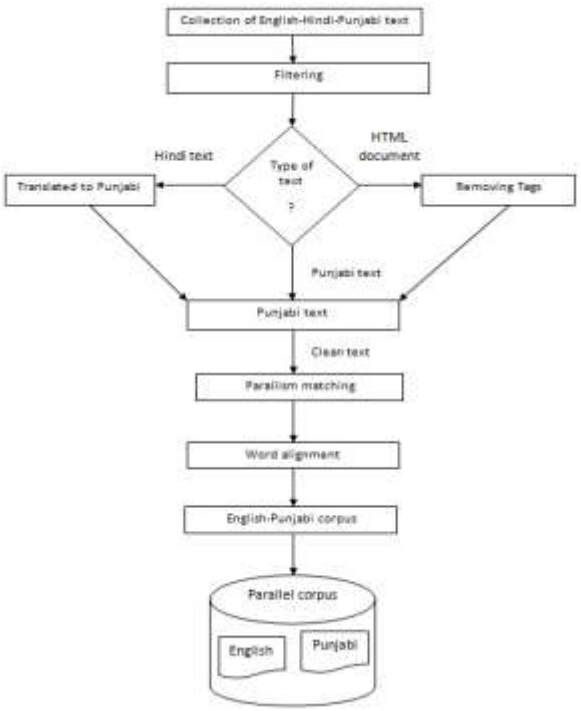

Figure 1. Architecture of Parallel corpus

VII ANALYSIS OF CORPORA RESOURCES

International Journal of Engineering, Science and Mathematics
Vol. 7 Issue 4, April 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@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

Gyan Nidhi corpus developed by C-DAC Pune, we purchased the corpus from C-DAC. The c0rpus was stored in 7 DVD's. It was available in English and another 18 Regional languages. We isolate the English language files and we search and match the corresponding Punjabi files, because all English files had not corresponding to Punjabi files. The gyan nidhi files contains the text of primary school level books. So it contains chatrs,pictures, symbols etc. So we convert the files into normal text mode. Some pictures of files were in HTML format, we write a program to convert the text into sentences.

The corpus of tourism and health contains so many special symbols, spaces, tags and it was available in English-Hindi language. So first of all we cleaned the text into normal form. Long sentences were trimed into shorter form. We had used notepad ++ to remove the noise from sentences. Then after cleaning the English-Hindi corpus, Hindi files were translated into Punjabi language with the help of Hindi to Punjabi statistical machine translation system, developed by Vishal Goyal and Ajit Kumar at Punjabi university, Patiala.

The EMILLE Corpora is developed with the joint venture between the Lancaster University, and the Central Institute of Indian Languages (CIIL), Mysore, India. The EMILEE corpora has three parts: monolingual, parallel and annotated corpora. In EMILEE fourteen monolingual corpora, in which written text and spoken data of fourteen Indian languages such as Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telugu and Urdu. The EMILLE monolingual corpora contain approximately 92,799,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. We have isolated English Punjabi text from the above corpus.

Bible corpus was downloaded from Internet. It was available in the form of chapters. The corpus was in English-Hindi language. Then again we convert Hindi text to Punjabi text by using Hindi to Punjabi statistical machine translation system. Some complicated and complex sentences were removed and longer sentences were converted into small sentences.

Guru Granth Sahib corpus was downloaded from Internet. It was available in English-Punjabi language. But corpus contains so many shaloks, shabad, hymns, and other unwanted text. The Punjabi part of the corpus was in Gurmukhi. Such Gurmukhi lipi was converted into plain Punjabi.

E-books of P.S.E.B. were downloaded from the Internet. The books were available in English and Punjabi medium. These books were in PDF. Format. The English medium books were easily convert from PDF format to word format but PDF copies of Punjabi medium books were not easily to convert. We used some of tools which was developed at Punjabi University, Patiala website like Sodhak, typing pad, font converter and spell checker etc. to convert these Punjabi files into the normal word format. The paragraphs of word were converted into sentences. Then it was stored into the spreadsheet.

Bilingual newspaper corpus was available in Corel draw files. These corel draw files were converted into word format with the help of corel draw software. The text was in English-Hindi format. Then, the Hindi text was translated into Punjabi text.

In Named Entities corpus, the names of persons, surnames, places, cities, states, countries, organizations, currencies, quantities, expression of time, etc. were included. These name entities were transliterated into Punjabi.

VIII CHALLENGES IN DEVELOPMENT OF PARALLEL CORPUS

There have been significant work in creating Parallel Corpora for related languages but creating Parallel Corpora for unrelated languages remain a challenge due to their word order differences among languages. The challenges become more acute if the languages in question are not related.

Following are some of the challenges to build the corpus

i. There is scarcity of resources for English-Punjabi text.
ii. Preprocessing the source text, a series of steps that would convert the texts to a standard format, reordering the source text, clean the texts,and define the word limit of sentence boundaries,
iii. Aligning the corpora at paragraph/sentence/word level
iv. Tagging the corpora for part of speech.
v. OOV(out of vocabulary)word becomes challenges in unrelated languages.

IX. CONCLUSIONS & FUTURE WORK

In this paper we discussed different sources of parallel corpus and its application in developing English-Punjabi statistical. It is very difficult to assess quality of MT without big size of parallel corpus.The use of MT system to translate a sentence from English into Punjabi can be judged from the quality of MT.The system output may be corrected as it differs from the ones by the human. The main obstacle was that parallel corpus was not available in sufficient quantities. As we saw most previous work has been conducted on a few manually constructed parallel corpus. This paper did not aim to test or analyze the statistical translation model. In this paper, however the process of constructing the parallel corpus did not take into consideration the domain to which the document belongs. The techniques presented in this paper searched only for the parallel pages that were good translation for each other. This technique would enrich the parallel corpus and make it bigger in size, but on the other side, it would have many false

parallel documents that in turn would result in worse translation quality.But how the parallel corpus could be effectively used in the proposed system is still being further investigated. This paper focuses on developing English-Punjabi parallel corpus and still the work is incomplete. We hope we could continue to push forward our effort to develop the corpus and the effective tools to analyze the entire system.

ACKNOWLEDGMENT

My thanks go to Dr. Jaskarn S Bhullar, MIMIT, Malout and Dr. Vishal Goyal, Punjabi University, Patiala. They made their contribution to the work presented in the paper.We have shown how the source text created and how the corpora in target language have been translated.The process of corpus development has been though labor intensive.

REFERENCES

[1] P. Baker, A. Hardie, T. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizauskas, O. Hamza, D. Maynard, V. Tablan, C. Ursu, B. D. Jayaram, M. Leisher, "Corpus linguistics and South Asian languages: corpus creation and tool development", *Literary Linguist. Comput*. Vo. 19, pp. 509–524, 2004.

[2] G. N. Jha, "The TDIL program and the Indian language corpora initiative (ILCI)", *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association*, 2010.

[3] N. Choudhary, "Web-drawn Corpus for Indian languages: a case of Hindi", *Proceedings of the ICISIL*, vol. 139, pp. 218–223. 2011.

[4] M. Shrivastava, P. Bhattacharyya, "Hindi POS tagger using naive stemming: harnessing morphological information without extensive Linguistic knowledge", *Proceedings of the International Conference on NLP* (ICON08), 2008.

[5] S. Dandapat, S. Sarkar, A. Basu, "Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor resource scenario", *Proceedings of the Association for Computational Linguistic*, pp 221–224, 2007.

[6] A. Bharati, D. M. Sharma, L. Bai, R. Sangal, "Anncorra: Annotating Corpora", LTRC, IIIT, Hyderabad, 2006.

[7] S. Baskaran, K. Bali, M. Choudhury, T. Bhattacharya, P. Bhattacharyya, G. N. Jha, S. Rajendran, K. Saravanan, L. Sobha, B. M. Subbarao, "A Commonparts-of-speech tag set framework for Indian languages", *Proceedings of the 6th International Language Resources and Evaluation* (LREC'08), 2008.

[8] V. Goyal, G. S. Lehal, "Hindi morphological analyzer and generator", *Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology*, 2008.

[9] T. Bögel, M. Butt, A. Hautli, S. Sulger, "Developing a finite-state morphological analyzer for Urdu and Hindi", *Proceedings of the 6th International Workshop on Finite-State Methods and Natural Language Processing*, 2007.

[10] V. Goyal and G. S. Lehal, "N-Grams Based Word Sense Disambiguation: A Case Study of Hindi to Punjabi Machine Translation System", *International Journal of Translation*, Vol. 23(1), pp. 99-113, 2011.

[11] V. Goyal and G. S. Lehal, "Advances in Machine Translation Systems", *Language In India*, Vol. 9, pp. 138-150, 2010.

[12] V. Goyal and G. S. Lahal, "Hindi Morphological Analyzer and Generator", *IEEE Computer Society Press*, Washington, DC, USA 1156-1159, 2008.

[13] P. Brown, S. A. D. Pietra, V. J. D. Pietra, R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, Vol. 19 (2), pp. 263-311, 1993.

[14] V. B. Dang, and B. Ho, "Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining", *Proceedings of the International Conference on Innovation and Vision for the Future*, pp..261-266, 2007.

[15] www.sikhiwiki.org/index.php/Guru_Granth_Sahib

[16] www.pseb.ac.in

[17] www.christos-c.com/bible

[18] www.tdil.mit.gov.in