LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 10 : 11 November 2010

ISSN 1930-2940

Development of Punjabi-Hindi Aligned Parallel Corpus from Web Using Machine Translation

Gurpreet Singh Josan, Ph. D.
Jagroop Kaur, M. Tech.

Abstract

Aligned parallel corpus plays a vital role for research in various automatic NLP tasks. A constantly increasing resource for collecting parallel text is the World Wide Web. This paper discusses a novel approach for collecting parallel text for language pair Punjabi-Hindi. We use Machine Translation and DOM for finding parallel text from internet. The collected text is of heterogeneous nature and is aligned at word level with high precision. The approach discussed in this paper guarantees high quality parallel data in short time span.

1. Introduction

Recent advancements in natural language processing are largely based on statistical approaches. The parallel corpus plays a vital role in statistical approaches as it allows empirical studies for various applications of NLP as language studies, machine translation, cross language information retrieval, bi-lingual lexicon development etc. Parallel corpus is a collection of original texts translated to another language where the texts, paragraphs, and sentences down to word level are typically linked to each other.

There exists multi language parallel corpus like Europarl, Bible, and OPUS etc. as well as bi lingual parallel corpus like ISJ-ELAN Sloveign English, English Chinese, English Norwegian parallel corpus etc. English enjoys the privileges when it came to the creation of parallel corpora. Most of the time, it is one of the two languages in the pair. Also the size of available corpus is limited. Another constraint is the limited domain. Most of existing corpora are developed from either government documents or from Newswire texts. There is a scarcity of parallel corpora for any other language pair excluding English particularly among Indian

languages. The problem is a big barrier in the development of NLP applications involving Indian Languages.

World Wide Web is a constantly evolving source of a parallel text. Electronically accessible information is available on the web and is increasing day by day. The web mining seems to be a promising and can be used for building parallel corpora for the under privileged and minority languages. Collecting parallel corpus particularly for resources starved languages from the internet is among the challenging problems in NLP tasks. This is not a trivial task at all for the huge network makes the process very labor intensive. Besides there are the chances that useful documents are mixed up with garbage and high quality translations are mixed up with garbage.

Therefore, scientists have designed several systems to automate this construction process. The idea leads to the development of software for automatic discovering parallel text on World Wide Web such as BITS (Xiaoyi and Liberman, 1999), PTMiner (Chen and Nie, 2000), and Strand (Resnik, 1998; Resnik and smith, 2003) etc. This paper describes a technique for automatic generation of parallel corpora for Punjabi and Hindi. We will try to utilize best possible techniques available and supplement these techniques with additional resources. We will show why the already present systems are not suitable for our work and then we describe how a machine translation system helps in identifying and then aligning the parallel corpus obtained from the web.

2. Existing systems

(Resnik, 1998) proposed a simple method based on the anchor tag. A simple query is posted to Altavista to locate the pages that point to a pair of pages which contain an anchor text indicating the language of its parallel text. This is the case for an Index.html file which contains pointers to two parallel texts anchored as “English version” and “French version”. However this simple method can only catch a small part of all the parallel pages. A lot of other parallel pages do not satisfy this condition.

PTMiner (Chen and Nie, 2000) uses the method described by Resnik and also employed file name matching. File name matching is based on the fact that the translated version has same resemblance in file name like same file name in respective language name folder or same filename with respective language suffixes e.g.

> …./hindi/file.htm and …../Punjabi/file.htm Or …./abc/file_h.html and …./abc/file_p.html

The outstanding feature of PTMiner is the ability to effectively reject false pairs prior to downloading them. (Chen et.al. 2000) uses parallel text identification system (PTI) which includes content analyzer module in addition to above mentioned techniques. This module measures the semantic similarity by using bilingual dictionary. BITS (Xiaoyi et.al, 1999) provides a different approach. All pages from a specified domain are crawled exhaustively. Their language is determined by a language detector and all possible combinations of these pages (a full cross product) have to be examined to find matches. In this proposal, a bilingual

dictionary has been used to perform matching at word level between parallel documents. This approach is easy to understand yet very time-consuming.

STRAND (Resnik, 1998; Resnik and smith, 2003) has a similar approach to PTMiner except that it handles the case where URL-matching requires multiple substitutions. Structural filtering with a tuning parameter optimized by using Machine Learning gives it the ability not to examine all possible combinations like BITS. (Resnik and smith, 2003) also proposes a content-based matching method as in (Xiaoyi and Liberman, 1999) but similarity is measured in a different way.

The parallel text identification system was developed by (Jisong, Chau and Yeh, 2004) for collecting parallel corpus from the web. A filename comparison module and a content analysis module are used to measure the semantic similarity between two pairs. They report recall rate of 0.96 and precision rate of 0.93. Another Automatic Acquisition of Chinese-English Parallel Corpus from the Web was performed by (Ying, Wu, Gao, and Vines, 2005). They used various features for candidate selection like anchor text, image alt attribute text etc. Extractions of candidate pair is done by pattern matching and edit distance similarity measure. They use KNN classifier for parallel text validation. (John Fry, 2005) also described a method of collecting parallel data from RSS feed.

2.1 DOM Tree Alignment Model

The Document Object Model is a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. It defines the logical structure of documents and the way a document is accessed and manipulated. (Lei, Cheng, Ming and Gao 2006) described a DOM based model for extracting parallel data. They claim the precision of system by using DOM to be 97.2%. Reduced bandwidth cost and improved mining throughput are some other benefits of their approach.

3. The Approach

For building Punjabi Hindi parallel corpus, the potential source is a news website http://www.webdunia.com/ which is published in eight languages besides Punjabi and Hindi. It was observed that when presenting the same content in two different languages, authors exhibit a very strong tendency to use the same document structure (e.g., Figure 1)

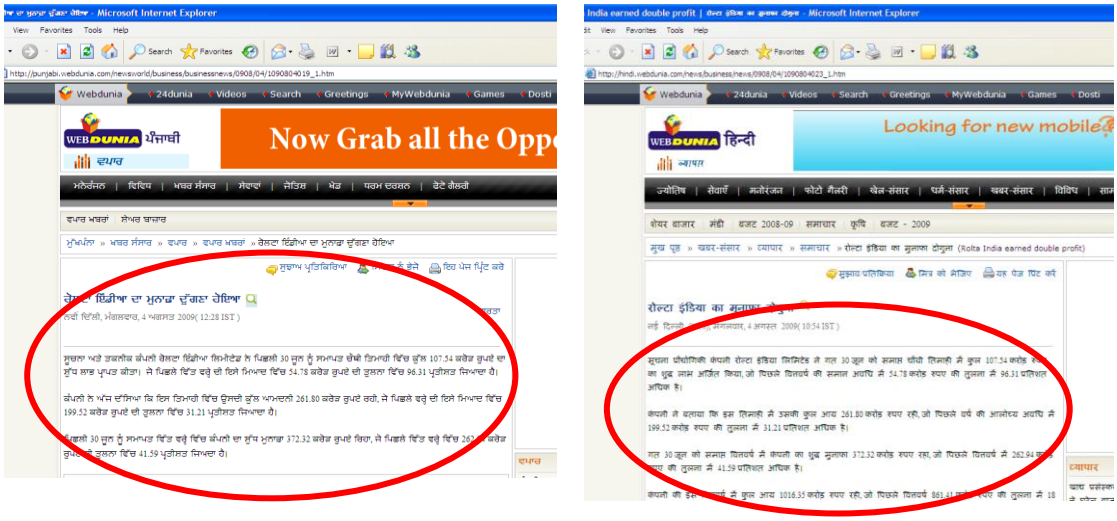

Figure 1 (a and b) Web Pages Showing same contents and similar page structure in Punjabi and Hindi (Similar contents circled)

Although the website contains lot of text in both languages, the collection of parallel text is not straightforward. The above discussed techniques are non-effective for the task in hand due to various reasons. First, the two versions of website don't use the anchor tag to point the other version. Second, the bilingual websites hosted by Webdunia uses varied naming schemes. E.g. a news item "ਆਤਮਘਾਤੀ ਹਮਲਿਆਂ ਨਾਲ ਦਹਿਲਿਆ ਕਾਬੁਲ" in Punjabi has file name http://punjabi.webdunia.com/newsworld/news/ international/0902/11/1090211025_1.htm whereas its Hindi version "आत्मघाती हमलों के साथ दहला काबुल" has filename http://hindi.webdunia.com/ news/news/international/0902/11/1090211133_1.htm.

We can see that the website uses numeric figures to identify the files which has no correspondence with other version. This means we can't use the file name for identifying the potential targets as done by PTMiner. Similarly the approaches used by PTI system and BITS are too much time consuming. Structural filtering in STRAND (Resnik and Smith, 2003) cannot be applied since all pages from a news website share the same structure.

The current system for collecting parallel text in Punjabi and Hindi is designed by keeping the above mentioned drawbacks. The system is shown in figure 2. It takes advantage of the machine translation system for converting source language text in target language text and then searches the web by posting the translated text to Google search engine. The result is filtered for the text retrieved from only webdunia host which is hosting Hindi version. Following are the implementation steps:

1. Select the source text from http://punjabi.webdunia.com/
2. Retrieve all the anchor tags.
3. For each anchor tag, retrieve the text between start and end anchor tag.
4. Translate the retrieved text using Punjabi to Hindi Machine Translation System.
5. Post the translated result to a search engine (Google in our case).
6. From the result obtained from the search engine, select only those which are retrieved from webdunia domain.

7. From the retrieved urls, fetch the page.
8. Perform sentence alignment (Described below)
9. Perform word alignment. (Described below)

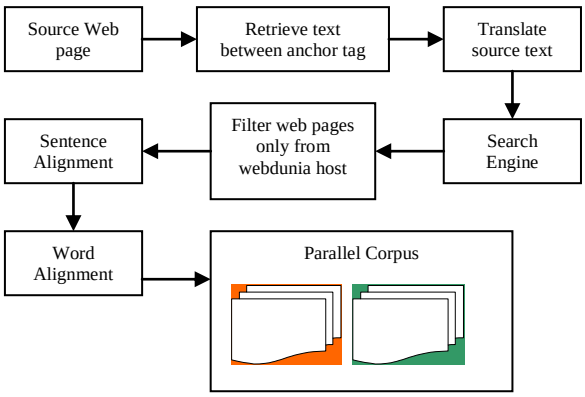

Figure 2 Modules in parallel corpus developer Machine Translation System

Machine translation system plays a pivotal role for collecting parallel corpus. The efficiency of system is limited by the accuracy of MT system. There is also a possibility that an MT system provide different wording for the same source concept. The MT system developed by (Josan and Lehal, 2008) is used for machine translation purpose. The following shows some output of Punjabi-Hindi translation:

Punjabi text: ਪੰਜਾਬੀ ਮੇਰੀ ਮਾਂ ਬੋਲੀ ਏ। Translated Hindi Text: पंजाबी मेरी मातृभाषा है।

TT: pañjābī mērī māṃ bōlī hai.
G: Punjabi my mother tongue is.
E: Punjabi is my mother tongue.

4. Sentence alignment

There are number of sentence aligning algorithms. Some are based on the sentence lengths like (Brown, 1991; Gale and Church, 1993). Chen (1993) did considerable amount of work on English-French corpus using lexical information. Bharati et.al. 2002, describe an algorithm for aligning sentences with their translations in a bilingual corpus using lexical information of the languages with a precision of 94.3%. Singh and Husain, 2005 has evaluated several algorithms for sentence alignment and suggested some guidelines for English Hindi sentence alignment.

This is quite obvious that alignment algorithms that use lexical information offer a potential for high accuracy on any corpus. We also tried to do sentence alignment using machine translation and document object model. Each element of HTML page is treated as one paragraph. For increasing the algorithm speed, we select only leaf nodes of DOM tree as these nodes contain the data that we are interested in. We found that the website designers use the same class name for the tags in corresponding files that contains parallel data. This helps us to quickly find the parallel text. If class name is not available then for each leaf node

tag in source file, the contents are translated and matched with the corresponding tags in the target file.

The possible parallel paragraphs can be identified by using a matching score function. The scores are obtained by matching the tokens of translated text and target language text. Different scoring functions can be used to calculate the score of match. The function used is same as that employed by Bharati et.al. 2002:

$$score(S,T) = \frac{N}{M}$$

Where N=Number of matching tokens
M=Maximum_of(source tokens, target tokens)
S: Source sentence T: Target sentence

The paragraphs are marked parallel if more than 70% of tokens match with the target text.

5. Word Alignment

For each aligned paragraph, the tokens of each sentence are translated and matched with the tokens of target sentence. The matching tokens are marked as aligned words. The tokens whose translated match is not found are marked as candidate tokens for word alignment as shown in following example:

Table 1 Word Alignment example

Punjabi text	ਵਿਅਕਤੀਗਤ	ਸੂਚਨਾਵਾਂ	ਵਿੱਕਰੀ	ਲਈ	ਇੰਟਰਨੈਟ	'ਤੇ
Translated Text	व्यक्तिगत	सूचनाएँ	बिक्री	के लिए	इंटरनेट	पर
Actual Hindi text	व्यक्तिगत	सूचनाएँ	बिक्री	हेतु	इंटरनेट	पर

For the sentence in above example, more than 70% tokens of translated text are matched with actual text. So this sentence is marked as parallel text. The matching tokens are marked as aligned words as shown by arrows in the table. As shown in above table for the word ਲਈ the translation produced is के लिए but target text contains हेतु. So ਲਈ and हेतु are marked as candidate tokens. Algorithms may be developed further to check whether हेतु can be a target of ਲਈ or not.

6. Results and Discussion

We manually examine 200 randomly picked pairs. The system gives 96.7% sentence level precision i.e. 96.7% of selected sentence pairs are actually parallel. At word level the figure is 95.5%. To find out the quality of the mined parallel corpus, 2000 sentence pairs were randomly taken from results and evaluated manually by two persons on the scale of three points as follow

Scale	Degree of Parallelness	Example	
1	Exact parallel	ਮਾਈਕ੍ਰੋਸੌਫਟ ਨੇ ਆਪਣੇ ਸਭ ਤੋਂ ਲੋਕਪ੍ਰਿਆ ਮੈਸੇਂਜਰ, ਐਮਐਸਐਨ ਮੈਸੇਂਜਰ ਨੂੰ ਬੰਦ ਕਰਨ ਦਾ ਫੈਸਲਾ ਕੀਤਾ ਹੈ।	माइक्रोसॉफ्ट ने अपने सबसे लोकप्रिय मैसेंजर, एमएसएन मैसेंजर को बंद करने का फैसला किया है।
2	Roughly parallel	ਜ਼ਿਆਦਾ ਪਾਣੀ ਪੀਣ ਨਾਲ ਜ਼ਹਿਰੀਲੇ ਤੱਤ ਸਰੀਰ ਵਿਚੋਂ ਬਾਹਰ ਨਿਕਲ ਜਾਂਦੇ ਹਨ।	प्रत्येक दिन 8-10 गिलास साफ पानी पीने से शरीर में रहने वाले जहरीले पदार्थ बाहर निकल जाते हैं।
3	Not parallel	ਖੂਬ ਪਾਣੀ ਪੀਣ ਨਾਲ ਜਵਾਨ ਅਤੇ ਫੁਰਤੀਲਾ ਦਿਸਿਆ ਜਾ ਸਕਦਾ ਹੈ।	भोजन के तुरन्त बाद पानी पीने से शरीर मोटा होता है।

Table 2 Quality analysis scale

The results are as follow:

	Exact parallel	Roughly parallel	Not parallel
No. of Sentences	1634	258	108

Table 3 Quality Analysis Results

One drawback of our method is that it is dependent upon the quality of translation system. If the translation system produced the wrong translation or the translation considerably different from the text available in Hindi sites then the probable good candidate document may get skipped. E.g. in following sentence

News heading in Punjabi site:

ਜੈਕਸਨ ਦੇ ਡਾਕਟਰ ਦੇ ਕਲੀਨਿਕ 'ਤੇ ਛਾਪਾ

{Raid on clinic of Jackson's doctor}

After Translation:

जैक्सन के डाक्टर के क्लीनिक पर छापा

Actual Text in Hindi News Site:

माइकल जैक्सन के डॉक्टर के कार्यालय पर छापा

{Raid on office of Michael Jackson's doctor}

In this sentence, Punjabi version uses the word ਕਲੀਨਿਕ where as English version use the word कार्यालय which can never be translated by the machine. Also note that Hindi version contains full name i.e. माइकल जैक्सन while Punjabi version contains only last name i.e. ਜੈਕਸਨ. This leads to the failure of search engine to get the target document from the internet.

7. Conclusion

Our experiment shows how internet can be helpful to quickly assemble a parallel corpus. In the case of our Punjabi-Hindi corpus, we supplemented the algorithm to perform sentence and word level alignments based on DOM on the fly. We collected 6,129 article pairs in a short time. Although the figure is not much attractive but we are sure to assemble more parallel data in future. The main features of our system are that the data collected is from variety of news and articles making it a heterogeneous collection, has less noise, and we get word level alignment in a short time span. The quality and rate of growth of our system are stable.

==========

References

- Anil Kumar Singh and Samar Husain. 2005. "Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs". IN proceedings of ACL 2005 Workshop on Parallel Text. Ann Arbor, Michigan. June 2005.

- Bharati Akshar, Sriram V, Vamshi Krishna A, Rajeev Sangal, Sushma Bendre. 2002. "An Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information", Published in the proceedings of ICON-2002: International Conference on Natural Language Processing, Mumbai, 18-21 Dec 2002.

- Brown P,J.Lai and R.Mercer. 1991. "Aligning Sentences in Parallel Corpora" 47th Annual meeting for the Association of Computational Linguistics.

- Chen Jiang and Nie Jian-yun. 2000. "Parallel web text mining for cross language IR" In Recherche d'Informations Assist´ee par Ordinateur (RIAO), pages 62–77, Paris, April.

- Chen Stanley. 1993. "Aligning Sentences in Bilingual Corpora Using lexical Information", Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, 9.16. Columbus, OH.

- Gale, William A and Church, Kenneth W. 1991. "A Program for Aligning Sentences in Bilingual Corpora." Proceedings of 29th Annual Meeting of the Association for Computational Linguistics, 177.184. Berkeley, CA.

- Jisong Chen, Chau, R. and Yeh, C.-H. 2004. "Discovering Parallel Text from the World Wide Web", In Proc. Australasian Workshop on Data Mining and Web

Intelligence (DMWI2004), Dunedin, New Zealand. CRPIT, 32. Purvis, M., Ed. ACS. 157-161.

- John Fry. 2005. "Assembling a parallel corpus from RSS news feeds", in Proceedings of the Workshop on Example-Based Machine Translation, MT Summit X, Phuket, Thailand, September 2005.

- Josan, Gurpreet Singh and Lehal, Gurpreet Singh. 2008. "A Punjabi to Hindi Machine Translation System", In proceedings of In proceedings of International Conference on COLING 2008 at University of Manchester, 18-22 Aug., 2008 pp157-160.

- Lei Shi, Cheng Niu, Ming Zhou and Jianfeng Gao. 2006. "A DOM tree alignment model for mining parallel data from the web". In proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia pp 489 - 496

- Resnik P. and Smith N. A. 2003. "The Web as a Parallel Corpus" Computational Linguistics, 2003, 29(3):349–380.

- Resnik Philip. "Parallel strands: A preliminary investigation into mining the Web for bilingual text" in Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529, Langhorne, PA, October 28-31.

- Xiaoyi Ma, Liberman Mark. "BITS: A method for bilingual text search over the web" Machine Translation Summit VII, September, 1999.

- Ying Zhang, Wu, K., Gao, J. F., and Vines, P. 2006. "Automatic Acquisition of Chinese-English Parallel Corpus from the Web". In Proceedings of ECIR-06, 28th European Conference on Information Retrieval. Imperial College London April 2006 pp 420-431

Gurpreet Singh Josan, Ph.D.
Rayat & Bahra Institute of Engineering & Biotechnology
Sahauran
Mohali
Punjab, India
josangurpreet@rediffmail.com

Jagroop Kaur, M.Tech.
University College of Engineering
Punjabi University
Patiala
Punjab, India.
jagroop_80@rediffmail.com