International Journal of Engineering & Technology, 7 (2) (2018) 690-693

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.8892
Research paper

Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System

Kamal Deep [1] *, Dr. Ajit Kumar [2], Dr. Vishal Goyal [2]

[1] Research Scholar, Punjabi University, Patiala
[2] Assistant Professor, Multani Mal Modi College, Patiala
[2]Associate Professor, Punjabi University, Patiala
**Corresponding author E-mail: kamal.1cse@gmail.com*

Abstract

This paper describes the creation process and statistics of Punjabi English (PunEng) parallel corpus. Parallel corpus is the main requirement to develop statistical machine translation as well as neural machine translation. Until now, we do not have any availability of PunEng parallel corpus. In this paper, we have shown difficulties and intensive labor to develop parallel corpus. Methods used for collecting data and the results are discussed, errors during the process of collecting data and how to handle these errors will be described.

Keywords: *English; Machine Translation; Parallel Corpus; Punjabi; Puneng Corpus*

1. Introduction

The quality of statistical machine translation or neural machine translation systems is strongly related to the amount of parallel text available for the language pairs [1]. However, most language pairs have little or no readily available bilingual training data available. Punjabi and English are worldwide languages. Punjabi is an Indo-Aryan language spoken by over 100 million native speakers worldwide, ranking as the 10th most widely spoken the language in the world [W1]. In contrast, English is spoken by around 125 million people in India, of which a very small fraction are native speakers [2]. So, there is a large requirement for digital communication in Punjabi and interfacing with the rest of the word via English. Hence there is a need of Punjabi to English Machine Translation system. Parallel corpus development is a first step here for the development of Machine Translation system. We have collected Hindi English corpus and Punjabi English corpus from various available resources. Then by applying different techniques, we have collected 266019 parallel sentences of Punjabi and English.
This paper is organized as: chapter 2 describes various sources for parallel corpus. Chapter 3 discusses various corpus processing techniques and chapter 4 tells the corpus statistics. Chapter 5 is conclusion.

2. Data sources for parallel corpus

On the part of building parallel corpora, we have used these approaches (a) human translation (b) extraction from an existing comparable corpus either from the web or from another digital media. A second translator can verify the human translation. However, there is no end to the verification. Since five different human translators may translate a source sentence in five different ways. Be it an ambiguous sentence, or be it pragmatic, or it requires discourse resolution, the meaning of the source sentence should be fully conveyed to the target sentence. We have collected parallel corpora from various sources online. The sources are:

2.1. Emille

The emille Corpus has been constructed as part of a collaborative venture between the EMILLE project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India [W2]. This corpus is distributed by the European Language Resources Association. EMILLE distributes corpus in three forms: monolingual corpora, parallel corpora and annotated corpora. It contains corpus for fourteen South Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu. EMILLE corpus contained Punjabi corpus extracted from various online newspapers: Daily Ajit Jalandhar, Sanjh Savera, Eh Din and Nagara [W3].They also collected Punjabi text from Shree Guru Granth Sahib.35000 parallel sentences of Punjabi and English corpus have been taken from EMILLE corpus.

2.2. TDIL

The Ministry of Electronics and Information Technology, India initiated the TDIL (Technology Development for Indian Languages) with the objective of developing Information Processing Tools [W4]. TDIL has created various Linguistic resources like Named Entity resources for various languages, Speech corpora, and Text Corpora. Text corpora contain monolingual as well as a bilingual corpus for various languages like Punjabi, Hindi, Gujarati and Marathi etc. Bilingual corpus is available for different domains like health, tourism, and entrainment etc. From TDIL we got Punjabi English parallel tagged corpora of 25000 sentences each in domain of tourism and health. So our first work was to remove tags form both English and Punjabi parallel corpus as we require untagged corpus for training.Table1 shows the tagged corpus of English and Punjabi.

Table 1: TDIL Tagged Corpus of Punjabi and English

Punjabi	English
ਦੰਦਾਂ\N_NN ਨਾਲ\PSP ਤੁਹਾਡਾ\PR_PRP ਆਤਮਵਿਸ਼ਵਾਸ\N_NN ਵੀ\RP_RPD ਵੱਧਦਾ\V_VM_VF ਹੈ\V_VAUX |\RD_PUNC	Your\PRP$ self-confidence\NN also\RB increases\VBZ with\IN teeth\NNS.\.
ਦੰਦਾਂ\N_NN ਨੂੰ\PSP ਠੀਕ\JJ ਤਰ੍ਹਾਂ\PSP ਸਾਫ਼\JJ ਕਰੋ\V_VM_VNF |\RD_PUNC	Clean\NNP your\PRP$ teeth\NNS properly\RB.\.
ਪ੍ਰਕਿਰਤਿਕ\JJ ਸੁੰਦਰਤਾ\N_NN ਨਾਲ\PSP ਭਰਪੂਰ\JJ ਸ਼ਿਮਲਾ\N_NNP ਹਿਮਾਚਲ\N_NNP ਪ੍ਰਦੇਸ਼\N_NN ਦਾ\PSP ਸਰਤਾਜ\JJ ਹੈ\V_VAUX |\RD_PUNC	Shimla\NNP provides\VBZ a\DT unique\JJ experience\NN for\IN the\DT tourists\NNS.\.
ਦੇਸ਼\N_NN ਆਜ਼ਾਦ\N_NN ਹੋਣ\V_VM_VNF ਤੱਕ\PSP ਸ਼ਿਮਲਾ\N_NNP ਪੰਜਾਬ\N_NN ਰਾਜ\N_NN ਦਾ\PSP ਹਿੱਸਾ\N_NN ਸੀ\V_VAUX |\RD_PUNC	On\IN country\NN becoming\VBG independent\JJ Shimla\NNP was\VBD part\NN of\IN the\DT Punjab\NNP province\NN.\.

After removing all the tags, we manually checked the corpus and found many mistakes. The types of mistakes found in the corpus are shown in Table 2.

Table 2: Errors in TDIL Corpus of Health and Tourism

English	Punjabi	Remarks
Use obtain for it.	ਦੱਬੀ ਹੋਈ ਨੱਕ ਦੇ ਉਭਾਰ ਦੇ ਲਈ ਅੱਜਕਲ੍ਹ ਪੋਰੇਕਸ ਇੰਪਲਾਂਟ ਦਾ ਉਪਯੋਗ ਕੀਤਾ ਜਾ ਰਿਹਾ ਹੈ |	Punjabi Sentence is not a translation of English sentence
	ਸਿਕੱਮ ਪ੍ਰਕ੍ਰਿਤੀ ਪ੍ਰੇਮੀਆਂ ਦਾ ਪਸੰਦੀਦਾ ਸਥਾਨ ਤਾਂ ਹੈ ਹੀ , ਟ੍ਰੈਕਿੰਗ ਦੇ ਲਈ ਰੋਮਾਂਚ ਭਰੇ ਅਨੇਕ ਖੇਤਰਾਂ ਦੇ ਕਾਰਣ ਟ੍ਰੇਕਰਸ ਨੂੰ ਵੀ ਵਿਸ਼ੇਸ਼ ਰੂਪ ਨਾਲ਼ ਆਕਰਸ਼ਿਤ ਕਰਦਾ ਹੈ |	In English Sentence, we got empty sentence whereas Punjabi side has some text written.
Here ``` rupa ''' stands for silver	ਇੱਥੇ ਰੂਪ ਦਾ ਮਤਲਬ ਚਾਂਦੀ ਤੋਂ ਹੈ	Not exact Translation
Complete it	ਜਿੰਨ੍ਹਾਂ ਲੋਕਾਂ ਨੂੰ ਨਕਸੀਰ ਫੁੱਟਣ ਦੀ ਸ਼ਕਾਇਤ ਹੈ ਯਾਨੀ ਜਿੰਨ੍ਹਾਂ ਦੀ ਨੱਕ ਵਿੱਚੋਂ ਖ਼ੂਨ ਆਉਂਦਾ ਹੈ , ਉਹਨਾਂ ਦੇ ਲਈ ਵੀ ਰੋਜ਼ ਆਂਵਲਾ ਖਾਣਾ ਬੇਹੱਦ ਲਾਭਦਾਇਕ ਹੈ |	In English, we got only “complete it” sentence.
Avoid oil , chilli	ਤੇਲ.ਖਟਿਆਈ , ਮਿਰਚ ਤੋਂ ਪ੍ਰਹੇਜ਼ ਕਰੋ |	Not exact Translation

2.3. WMT

WMT (Workshop on Machine Translation) is a workshop on an area of Statistical Machine Translation [W5]. The organizers of the workshop provide training and testing data for the development of Machine Translation system. A parallel corpus of 155000 sentences of English and Hindi has been taken from the workshop site. As we needed English Punjabi Parallel Corpora, so Hindi text is converted to Punjabi by using the hi2puSMT system. After translation to Punjabi from Hindi of WMT data we manually checked the English and Punjabi text of WMT data by copying whole data to excel. A lot of Punjabi sentences are not an exact translation of English sentences. Table 3 shows mistakes of WMT corpus. We removed all these types of sentences. At last, we got corpora of 143500 sentences from WMT.

Table 3: Errors in WMT Corpus

English	Punjabi	Remarks
%s and %s	ਅੰਦਰ	Not a sentence
%s:%d: Parse error: %s	:: ਵਿਆਖਿਆ ਖਾਮੀਂ :	Contains special characters
Invite them home. <s> Let them have somewhere they can be together.	ਉਨ੍ਹਾਂ ਨੂੰ ਕੋਈ ਅਜਿਹੀ ਜਗ੍ਹਾ ਦਿਓ ਜਿੱਥੇ ਉਹ ਇਕੱਠੇ ਹੋ ਸਕਣ	Contain tags on the English text
Or	or	Punjabi and English sentence is same
{0} joined the network {1}	{ 0 } подключился к сети { 1 }	Punjabi sentence is not an exact translation of each other.
<s> Strikeout</s>	ਹਟਾਏਂ</s>	Contains tags on both English and Punjabi side

2.4. Brills bilingual newspaper

Brills is English Hindi Bilingual Newspaper published from Bathinda (Punjab) in India. It publishes a newspaper weekly that contains parallel news in English and Hindi. We got a soft copy of 89 bilingual newspapers in the form of cdr (CorelDraw) files from them. We have extracted text from these files by using java program developed by us for this purpose. Then extracted text was aligned by using the alignment tool. We have extracted 10006 parallel sentences of English and Hindi form Brills Bilingual Newspaper.

2.5. Gyan nidhi

C-DAC Pune [W8] developed Gyan Nidhi corpus. This corpus was available in English and another 18 Regional languages. We got a corpus in forms of [7] DVDs from C-DAC Pune. From each DVD we copied the English and Punjabi corpus folder. Each folder contained data in form of HTML files. Each HTML file has text, charts, pictures, and symbols. Figure 1 and Figure 2 shows the extracts of English and Punjabi HTMLfile.

Fig. 1: Gyan Nidhi English HTML File.

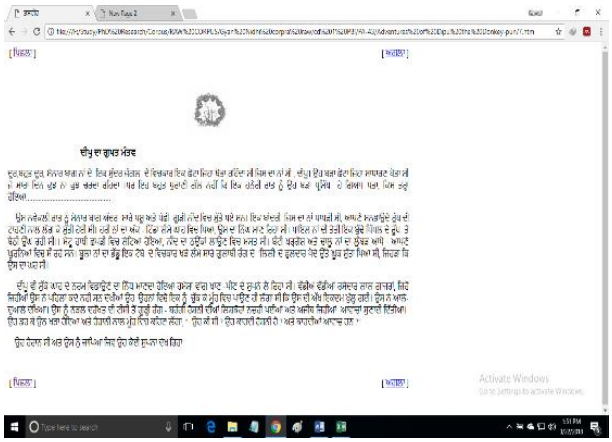

Fig. 2: Gyan Nidhi Punjabi HTML File.

By cleaning this corpus we have extracted 10000 parallel sentences of English and Punjabi. We manually checked the corpus and found that data extracted is actually a comparable corpus and not parallel corpus. Table 4 shows the Gyan Nidhi Corpus.

Table 4: Gyan Nidhi Corpus

English	Punjabi	Remarks
"Let's peep into his room and see what's up.	ਚਲੋ ਦਾਦਾ ਜੀ ਦੇ ਕਮਰੇ ਵਿਚ ਝਾਕ ਕੇ ਦੇਖਦੇ ਹਾਂ	his"in English is translated to ਦਾਦਾ ਜੀ
Dipu, of course, was very cozy in his soft bed of hay. He was dreaming as usual of food	ਦੀਪੂ ਵੀ ਸੁੱਕੇ ਘਾਹ ਦੇ ਨਰਮ ਵਿਛਾਉਣੇ ਦਾ ਨਿੱਧ ਮਾਣਦਾ ਹੋਇਆ ਹਮੇਸ਼ਾ ਵਾਂਗ ਖਾਣ -ਪੀਣ ਦੇ ਸੁਪਨੇ ਲੈ ਰਿਹਾ ਸੀ।	Punjabi's one line text is translated into two-line text in English.
"Now tell me what I must do."	ਹੁਣ ਮੈਨੂੰ ਦੱਸ ਕਿ ਮੈ ਕੀ ਮਦਦ ਕਰਾ ?	Sentences are not an exact translation of each other.
"What about Chalu, the fox?" asked Papri.	"ਚਾਲੂ ਲੂੰਬੜ ਬਾਰੇ ਕੀ ਰਾਇ ਹੈ ?" ਪਾਪੜੀ ਨੇ ਪੁੱਛਿਆ।	Good translation of each other.

2.6. News.2012.en.shuffled

From Ninth Workshop on SMT [W5] we download the compressed file of monolingual data of English 2012. A single text file contains approximate 50 Lakh English sentences in separate lines. A number of words in each line vary from 5 to 50. By developing a code using Java, we have extracted all those lines which contain maximum 20 words in each line of text file. B using Google translator [W6], we have translated 20012 sentences from English to Punjabi.

2.7. Wikipedia

We manually copied 1100 English sentences from Wikipedia [W7]. These sentences are converted to Punjabi manually.

3. Corpus processing

Data we got from section 2.2(TDIL) was also manually checked. There were approximately 600 sentences out of 50K corpus on English side that contains "Complete it." Only. Some sentences were not a translation of each other and in some sentences, Punjabi is written but corresponding English was not written. Approximate 1K sentences are of this type. So whole health and tourism data is manually checked sentence by sentence. After doing all processing we got a parallel corpus of 47500 sentences from TDIL of Health and Tourism domain. Figure 3 is flowchart that shows different processing steps done on TDIL Corpus.

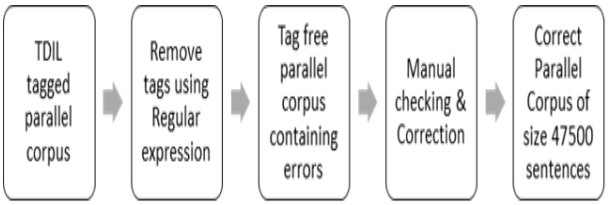

Fig. 3: Processing Of TDIL Corpus.

We also got Hindi-English Parallel Corpus from WMT (section 2.3) and Brills Bilingual News Paper (section 2.4). So to create Punjabi-English Parallel corpus from Hindi-English Parallel corpus we have to translate Hindi or English text to Punjabi. If we did translation work manually it took a lot of time. To convert Hindi text to Punjabi we have basically three systems Google Translate [W6], hi2pulearnPunjabi [3] and hi2puSMT system. These three systems were compared by Dr. Ajit and the hi2puSMT system performs better than other two systems. We have also used the hi2puSMT system for translation of Hindi text to Punjabi text. So in this way, we translated the 165000 lakh sentence from Hindi to Punjabi. Figure 4 is flowchart that shows different processing steps done on WMT/Brills Bilingual Corpus.

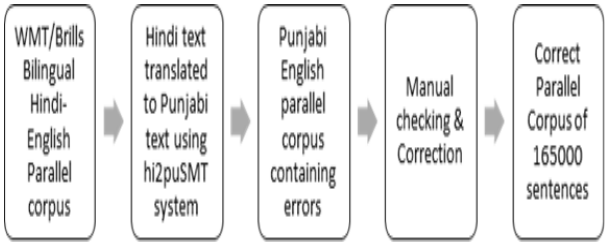

Fig. 4: Processing of WMT/Brills Bilingual Corpus.

After translation to Punjabi from Hindi of WMT data we manually checked the English and Punjabi text of WMT data by copying whole data to excel. Many Punjabi sentences are not an exact translation of English sentences. We removed all these types of sentences also. At last, we got corpora of 143500 sentences from WMT.

Section 2.5(Gyan Nidhi) provided data in form of excel sheets after copied from HTML pages. We manually checked this comparable corpus to make it parallel to use in the task of Machine translation. Therefore, after doing changes in sentences and deletion of sentences if they are not parallel we got 8900 sentences for the parallel corpus. Figure 5 is flowchart that shows different processing steps done on Gyan Nidhi Corpus.

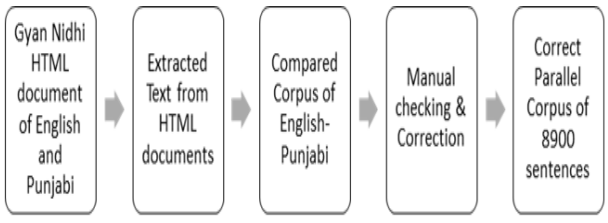

Fig. 5: Processing Steps on Gyan Nidhi Corpus.

Data from section 2.6(News.2012.en.shuffled) and 2.7(Wikipedia) was in English. We translated it to Punjabi by use of Google Translator [W6]. This data was again checked and we got approximate 21112 sentences of Punjabi and English from both corpora.

4. Corpus Statistics

Now we summarize the overall statics of Punjabi English corpora after doing all processing in form of table 5.

Table 5: Punjabi English Corpus Statistics

Source	Sentences(Parallel)	Tokens in English	Tokens in Punjabi
EMILLE	35000	0.71M	0.77M
TDIL	47500	0.80M	0.84M
WMT	143500	1.51M	1.59M
Brills Bilingual News-paper	10006	0.21M	0.23M
GyanNidhi	8900	0.18M	0.20M
News.2012.en.shuffled	20012	0.30M	0.36M
Wikipedia	1101	0.02M	0.02M
TOTAL	266019	3.97M	4.28M

5. Conclusion

We have presented Punjabi English (PunEng) parallel corpus that will be used to implement statistical and neural machine translation system in future. We have discussed different sources for the development of English Punjabi parallel corpus. Verification is essential requirement to ensure a quality of corpus. In future, we are also planning to increase the size of the corpus from additional sources like Newspapers, Indian Government websites.

References

[1] M. Post, C. Callison-Burch, and M. Osborne, "Constructing parallel corpora for six Indian languages via crowdsourcing," Wmt-2012, pp. 401–409, 2012.
[2] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The IIT Bombay English-Hindi Parallel Corpus," pp. 2–5, 2017.
[3] V. Goyal and G. S. Lehal, "Hindi to Punjabi machine translation system," Commun. Comput. Inf. Sci., vol. 139 CCIS, no. 1, pp. 236–241, 2011.

Webliography

[W1]https://en.wikipedia.org/wiki/Punjabi_language
[W2] http://www.lancaster.ac.uk/fass/projects/corpus/emille/
[W3]http://www.lancaster.ac.uk/fass/projects/corpus/emille/MAUAL.htm
[W4]http://tdildc.in/index.php?option=com_download&task=fsearch&lang=en&limitstart=15&limit=5
[W5]http://www.statmt.org/wmt16/translation-task.html
[W6]https://translate.google.com/
[W7]https://www.wikipedia.org/
[W8]http://tdildc.in/index.php?option=com_download&task=showresourceDetails&toolid=281&lang=en