International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Parallel Corpora : A Much-Needed Linguistic Resource for Low Computational Resource Languages

Preeti Dubey
Assistant Professor, Department of Computer Science, J&K Higher Education Department, India

ABSTRACT

Natural language Processing (NLP) is one of the upcoming research areas of computer science. There are many applications of NLP, but in the last decade, most of the effort in this field is inclined towards machine translation. A lot of work is available for the machine translation of English and Hindi. Some work is also undertaken for the translation of Indian languages, therefore; there has been a revolutionary research in development of text in machine readable form. Currently efforts are being made for developing large parallel corpora for most Indian languages, which is a much-needed linguistic resource for the development of Statistical Machine Translation systems. This paper introduces the concept of parallel corpus, its need and application in natural language processing. The various projects undertaken for the development of parallel corpus, followed by tools where parallel corpus is applied is also presented. The need of development of this resource for languages with low computational resources is also discussed.

Keywords : Text Corpus, Speech Corpus, Parallel Corpora, Natural Language Processing, Low Resource Languages

I. INTRODUCTION

Machine Translation Systems are in great demand and are widely in use. For the past few years, a number of Machine Translation Systems has been developed for Indian as well as foreign languages. The efficiency of a machine translation system depends upon the accuracy rate of the output produced by the system. Therefore, machine translation is not mere dictionary based substitution of words of one natural language into another natural language, but it needs to preserve the meaning of the sentences just like a human translator. There are many available approaches that can be used for machine translation. Some famous approaches are: Direct, Indirect and statistical. Recently machine translation systems are also being developed based on deep learning methods. The statistical approach of MT is widely used as most systems developed using this method have highly accurate results.

II. THE STATISTICAL APPROACH

of machine translation is being widely used for the purpose of achieving efficient outputs. These systems require a large parallel corpus and the working is based on statistical methods like the Bayes' Theorem. The text to be translated is matched with that in the corpus and translation is done with the text has maximum frequency. Some statistical machine translation systems that display highly accurate results have been developed for the following language pairs: Hindi-Punjabi, Punjabi-English, English-Urdu, Telugu, Gujarati-English, Bengali -

English etc. As read in the literature, the SMT output is coarse due to lack of corpora for Indian languages or due to small size of the corpus. As studied by NJ Khan et.al. [6], the results of SMT system that takes the Indian language (Hindi, Urdu, Bengali, Tamil, Malayalam, Telugu) sentences as input and it generates corresponding closest translation in English. The translation of over 800 sentences were evaluated using automatic evaluation metric i.e. BLEU evaluation. The reported average BLEU score was 10% to 20% for all the languages. It is concluded by the authors in their study that the quality of translation is directly dependent on the scope and quality of parallel language corpora.

Statistical methods are not only being used for the development of machine translation systems but also for the evaluation of machine translation systems. Evaluation of the output produced by the machine translators is very important. Earlier the evaluation of these systems was completely manual i.e. the evaluation was done by linguists manually. Therefore, it was time consuming and a cumbersome process. Presently many statistical evaluation tools are available. Some widely used automatic evaluation tools are: BLUE, NIST etc.

The major requirement of any statistical tool is the parallel corpus. The accuracy any statistical tool whether a statistical machine translation system or a statistical evaluation system depends on the size of the corpus used by it.

III. CORPUS

A corpus is a collection of text or phrases of a language that can be used as a sample of the language. Corpus can be text as well as spoken. Collection of spoken/speech corpus is difficult as compared to text corpus. Corpus can also be a part of a larger corpora, such a corpus is called sub corpora. Sub corpora can also be domain specific for example corpora containing only technical text or corpora for the medical domain, tourism etc.

A parallel corpus is a collection of texts, translated into one or more languages. If it involves two languages such that one of the corpora is an exact translation of the other, then it is referred as a bilingual corpus If some corpora involves more than one language such that one of the corpora is an exact translation of the more than one language, it is called multilingual parallel corpora.

IV. NEED FOR PARALLEL CORPUS

To enhance research on computational linguistics, there is a great need to generate linguistic resources which can further be used for developing tools that can be used for languages that are computationally low-resourced. Dogri is one such language which has how computational resources. It is a language used in the state of Jammu and Kashmir. It is a constitutional language of India. Presently, there is no work done so far related to the technological development of Dogri and which is the need of the hour. Only one tool that is the **Hindi to Dogri machine translation system** developed by the author in 2014 is available for the automatic translation of Hindi Text into Dogri text.

V. CHARACTERISTICS OF CORPUS

I. The corpus should be as large as possible, since the accuracy of the system developed depends on the size/quantity of the corpus used.
II. It should have a variety of text/speech samples. The efficiency of the system also depends on the variety of samples in the corpus. Therefore, the quality of the corpus must be varied.

VI. INITIATIVES TAKEN FOR THE DEVELOPMENT OF PARALLEL CORPUS FOR INDIAN LANGUAGES

- **GyanNidhi Parallel Text Corpus**

It contains million pages multilingual parallel text corpus in English and 11 Indian languages. It is a useful resource that can be used for as improving translation system, and also be useful for other applications such as spell checkers dictionaries. Kiran Pala, Sriram Chaudary, Lakshmi Narayana kodavali and Keshav Singhal (2008) have worked on the alignment of English to Hindi texts in Gyan Nidhi parallel corpus at sentence level.

- **ILCI (Indian Languages Corpora Initiative)**

This project is funded by Technology Development for Indian Languages (TDIL) unit of Ministry of Communication and Information Technology (MCIT) for building parallel corpus for major Indian languages including English.

The project is aimed at building parallel corpora for Hindi (SL). It is focused on two domains namely: health and tourism.

- **EMILLE Corpus**

The EMILLE Corpus is a collaborative effort by the EMILLE Project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. The EMILLE/CIIL Corpus (ELRA-W0037) is distributed free of charge for use in non-profit-making research only.

- **TIDES**

It is a Hindi-English corpus which was originally collected for the DARPA-TIDES surprise language contest in 2002. It was later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008. It contains 50K sentence pairs taken mainly from news articles.

- **WMT (Workshop on Machine Translation)**

In 2014, WMT introduced English-Hindi as an experimental, low resource language pair.

- **The Hindi-Punjabi parallel corpus**:

was developed using the existing Hindi to Punjabi machine translation system developed by Vishal Goyal. Vishal Goyal and Pardeep Kumar (2010) have contributed by developing the parallel corpus for this language pair.

VII. CONCLUSION & FUTURE SCOPE

In current situation of NLP, research is progressing for Indian languages that have the required linguist resources for their automization, whereas the computationally low resources languages are still struggling Low resourced languages are the languages for which the computational resources required for the automatic translation of two languages are not available. Computational resources like machine readable dictionary, corpora etc are very important for the development of NLP tools. It is very challenging in terms of time and money to start from scratch. Dogri is one such low resourced language. The only NLP tool for this language is the HINDI-Dogri machine translation system developed by the author. The author is also working on the development of this resource using the existing HINDI-DOGRI machine translation system. to make way in this field. The plight is that a computer researcher has a financial constraint to develop the linguist resources and on the other hand, a linguist lacks the computational knowledge. The state of art is that the research on NLP for these computationally low resourced languages can progress only if the required resources are developed, enabling the regional languages researchers to find flaws in the

available methods and develop new techniques/algorithms. Therefore, the development of these resources must be encouraged for the processing of every low resourced Indian language.

VIII. REFERENCES

[1]. Akshar Bharati, Dipti Misra Sharma, Rajeev Sangal et al., (15th December, 2006), AnnCorra: Annotating Corpora, Guidelines for POS and Chunk Annotation for Indian Languages. Retrievedfrom http://researchweb.iiit.ac.in/~rashid.ahmedpg08/ilmtdocs/chunk-posann-guidelines-15-Dec-06.pdf (15th December, 2006) AnnCorra: Annotating Corpora, Guidelines for POS and Chunk Annotation for Indian Languages. Retrievedfrom http://researchweb.iiit.ac.in/~rashid.ahmedpg08/ilmtdocs/chunk-pos-ann-guidelines-15-Dec-06.pdf

[2]. Ben Langmead. (n.d.) Hidden Markov Models. Retrieved from http://www.cs.jhu.edu/~langmea/resources/lecture_notes/hidden_markov_models.pdf

[3]. PAN Localization. (n.d.). Retrieved from http://www.panl10n.net/english /Outputs%20 Phase%202/CCs/ Nepal /MPP/Papers/2008/Report% 20on% 20Nepali%20Computational%20Grammar.pdf

[4]. Chirag Patel et. al., Part-Of- Speech Tagging for Gujarati Using Conditional Random Fields, Proc. Of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 2008, pp.117-122.

[5]. NJ KHAN,et.al, ' machine translation approaches and survey for indian languages' https://arxiv.org/ftp/arxiv/papers/1701/1701.04290.pdf

[6]. Mutatis Iqbal, et.Al ' English to Kashmiri machine translation system, International journal of Advance Research in Computer Science & technology (IJARCST 2015),vol:3, issue2 (Apr. - Jun. 2015), ISSN : 2347 - 8446 (Online) ISSN : 2347 - 9817 (Print)

[7]. Raghavendra Udupa U, et. Al, " An English-Hindi Statistical Machine Translation System", Part of the Lecture Notes in Computer Science book series (LNCS, volume 3248), LNAI 3248, pp. 254–262, 2005. https://link.springer.com/chapter/10.1007/978-3-540-30211-7_27

[8]. Tej Bahadur Shai et al. 2013. Support Vector Machines based Part of Speech Tagging for Nepali Text, International Journal of Computer Applications, May 2013, Vol: 70-No. 24, pp. 0975-8887.

[9]. Prajadip Sinha et al. Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach, International Journal of Emerging Technology and Advanced Engineering.2015 Vol 5(5).

[10]. Antony P J et al. 2011.Parts of Speech Tagging for Indian Languages: A Literature Survey, International Journal of Computer Applications, 2011, Vol. 34(8), pp. 0975-8887.

[11]. Amruta Godase, "MACHINE TRANSLATION DEVELOPMENT FOR INDIAN LANGUAGES AND ITS APPROACHES", International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015, ISSN: 2278-1307