Punjabi to Urdu Machine Translation System

Nitin Bansal[1], Ajit Kumar[2]

[1] Department of Computer Science, Punjabi University, Patiala, India

[2] Associate Professor, Multani Mal Modi College, Patiala, India

E-mail: [1] profnitinbansal@gmail.com, [2] ajit8671@gmail.com

Abstract

Development of Machine Translation System (MTS) for any language pair is a challenging task for several reasons. Lack of lexical resources for any language is one of the major issues that arise while developing MTS using that language. For example, during the development of Punjabi to Urdu MTS, many issues were recognized while preparing lexical resources for both the languages. Since there is no machine readable dictionary available for Punjabi to Urdu which can be directly used for translation; however various dictionaries are available to explain the meaning of the word. Along with this, handling of OOV (out of vocabulary words), handling of multiple sense Punjabi word in Urdu, identification of proper nouns, identification of collocations in the source sentence i.e. Punjabi sentences in our case, are the issues which we are facing during development of this system. Since MTSs are in great demand from the last one decade and are being widely used in applications such as in case of smart phones. Therefore, development of such a system becomes more demanding and more user friendly. Their usage is mainly in large scale translations, automated translations; act as an instrument to bridge a digital divide.

1 Introduction

Due to the availability of many regional languages in India, machine translation in India has enormous scope. Human and machine translation have their share of challenges. Scientifically and philosophically, machine translation results can be applied to various areas such as artificial intelligence, linguistics, and the philosophy of language. Various approaches are required in machine translation to make communication possible among two languages. These approaches can be rule-based, corpus-based, hybrid or neural-based. Here, hybrid approach is a combination of two approaches i.e. rule-based and corpus-based mainly. The quality of machine translation systems can be measured mainly using Bi-lingual Evaluation Study (BLEU), where it produces a score between 0 and 1.

Among various regional languages in India, we have chosen Punjabi and Urdu for developing Punjabi to Urdu Machine Translation System (PUMTS). Punjabi is the mother tongue of our state, Punjab, where it was used as an official language in government offices. Urdu was also being used as an official language in Punjab, before independence. Thus, PUMTS helps us to make Punjabi understandable to Urdu communities who still want to be in touch with earlier Punjab. These two languages in India, are taken as resource-poor languages, because parallel corpus on language pairs is not available. Thus it became a challenging task for us to develop parallel corpus on this language pair. Further, it also describes types of MTSs being developed with Indian and non-Indian perspective.

2 Methodology

An introduction to Punjabi and Urdu languages help in understanding about history and close proximity among this language pair. Since word-order of this language pair is same but writing order is different from each other i.e. Punjabi can be written from left-to-right and Urdu from right-to-left. Mapping among characters of language pairs has also been studied during the development of PUMTS. The implementation of our methodology for the development of PUMTS, where the architecture followed during the development has been documented. We have proposed three approaches

to develop bilingual parallel corpus for Punjabi and Urdu languages. But BLEU score suggested for one final approach for corpus development, results in higher accuracy. All the algorithms which were developed during the development of PUMTS, followed the final corpus approach. Lastly, Punjabi to Urdu machine transliteration system to handle Out-of-Vocabulary Words (OOV) words has also been designed and developed, which is working as web-based nowadays. This system has been designed in two phases i.e. first on a web-based platform using ASP.Net and secondly, it has been designed for PUMTS, to handle OOV words during machine translation, using MOSES platform.

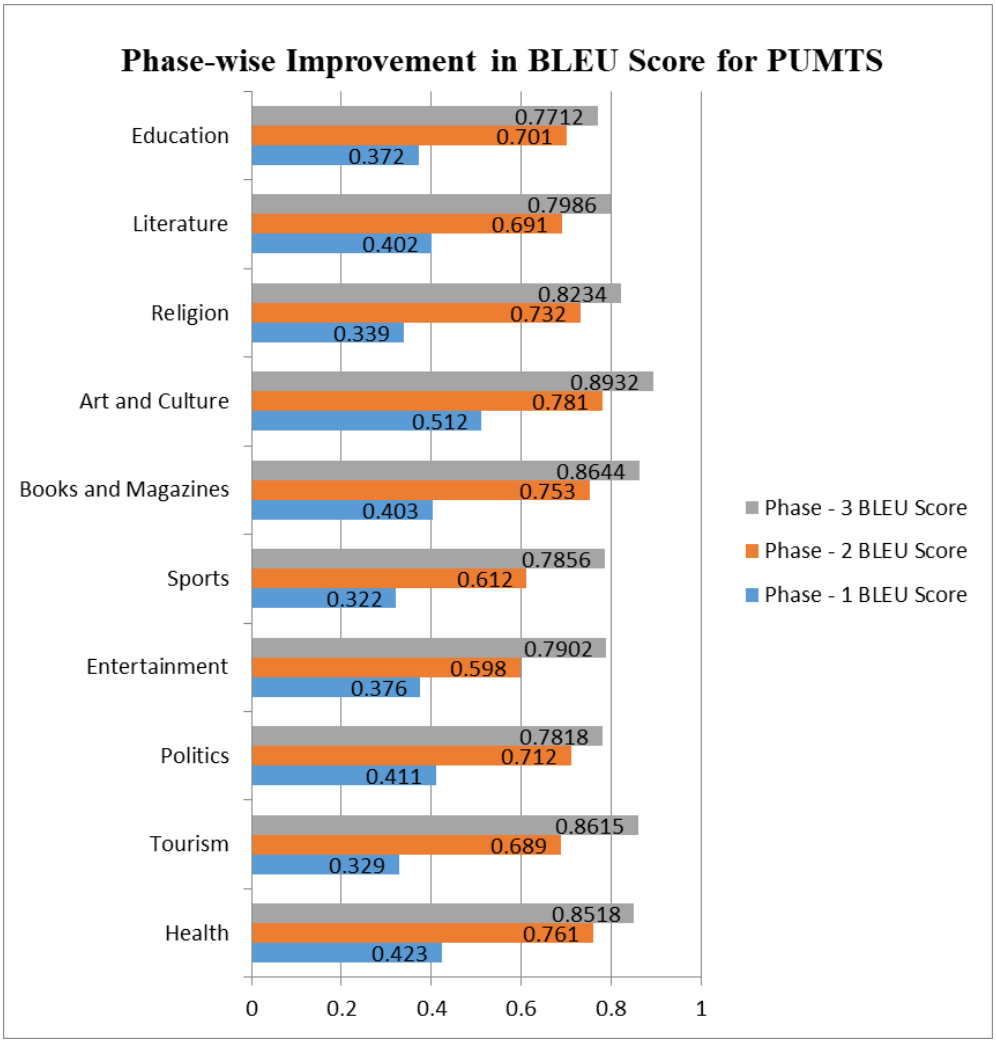

Chart 1: Phase-wise improvement in BLEU score for PUMTS

3 Results and Discussion

Various results had been evaluated by starting from 10000 parallel sentences to 1 lakh parallel sentences after including pre-processing and post-processing modules. The results have been compared with Google translator so as to keep the accuracy comparable and required improvisation can be included in PUMTS. Human evaluation has also been conducted where our evaluators are well known to both the languages. Accuracy has been tested using standard automated metric methodologies i.e. BLEU and NIST, on PUMTS and Google translator. Data domains followed during the development of parallel corpus are politics, sports, health, tourism, entertainment, books &

magazines, education, arts & culture, religion, and literature.

Since, human evaluation is still considered the most reliable and efficient method to test the system's accuracy. However, this is impracticable in today's circumstances. Thus, we have used automatic evaluation with BLEU and NIST to quickly and inexpensively evaluate the impact of new ideas, algorithms, and data sets. During the evaluation of PUMTS, a sufficient bilingual parallel corpus in Punjabi-Urdu language pair (more than 1 lakh parallel sentences) has been used on MOSES, and automated standard metric scores have been generated. Various methods had been applied to increase the system's accuracy, like the order of languages has been changed during the testing to analyze which one gives better results. Moreover, the PUMTS system has also been checked with the Google translator output, where we have found that our system output performs better than Google translator with an accuracy of about 82%. Following chart representation helps us to get an idea where PUMTS generates better results domain-wise.

As shown in chart 1, the development of PUMTS has been started from 10,000 parallel sentences, and the MOSES system has been set-up for this purpose to regularly test the accuracy of this data. Therefore, phase-wise testing and the recording of BLEU and NIST scores has been performed. The second phase has been tested on 50,000 sentences, and after that, final evaluation has been performed on more than 1,00,000 sentences. We can observe from the above chart; there was a sharp increase in accuracy when the number of sentences had been increased from 10,000 to 50,000 sentences. It has also been observed that the increase in size from 50,000 to 1,00,000 results in increments of accuracy at a slower rate, which is due to the handling of OOV words and increments on corpus size, gives more chances of meaningful sentences too.

References

Thomas D. Hedden, 1992-2010, *Machine Translation: A brief Introduction*, http://ice.he.net/~hedden/intro_mt.html

P Koehn, H Huang, et al., 2007, *Moses: Open Source Toolkit for Statistical Machine Translation*. ACL Demos, 2007.

Shahid Aasim Ali and Malik Muhammad Kamran, 2010, *Development of parallel corpus and English to Urdu Statistical Machine Translation*, International Journal of Engineering and Technology, PP. 31-33, Vil 10 No 5, October 2010.

Ajit Kumar and Vishal Goyal, 2011, *Comparative analysis of tools available for developing statistical approach based machine translation system*, in proceedings of International conference ICISIL 2011, Patiala (Punjab), India, PP. 254-260, March9-11.

Tajinder Singh Sani, 2011, *Word Disambiguation in Shahmukhi to Gurmukhi Transliteration*, Processing of the 9th Wordshop on Asian Language Resources, Chiang Mai, Thailand, pages: 79-87, November 12 and 13.

Gurpreet Singh Lehal and Tejinder Singh Saini, 2012, *Development of a Complete Urdu-Hindi Transliteration System*, Proceedings of COLING 2012: Posters, PP. 643-652, COLING 2012, Mumbai.

Arif Tasleem et al, *An analysis of challenge in English and Urdu machine translation*, National conference on Recent Innovations and Advancements in Information Technology (RIAIT 2014), ISBN 978-93-5212-284-4

Ajit Kumar and Vishal Goyal, 2015, *Statistical Post Editing System (SPES) applied to Hindi-Punjabi PB-SMT system*, Indian Journal of Science and Technology", Vol 8(27).

Zakir H. Mohamed and Nagnoor M. Shafeen, 2017, *A brief study of challenges in machine Translation*, International journal of computer Science Issues, PP. 54-57, Vol 14 No 2.