International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792


# A Baseline English-Garhwali Statistical Machine Translation System

Arushi Uniyal
*Research Scholar, Center for Linguistics, Jawaharlal Nehru University, New Delhi, India*

***Abstract*: This paper focusses on the development of a baseline statistical machine translation system for the language pair English and Garhwali. This is the first attempt to generate machine translation technology for Garhwali which is one of the lesser-known and less resourced languages of India. The paper discusses the resource development process, the methodology for system development and the evaluation score of the baseline system.**

***Keywords*: Statistical Machine Translation, Garhwali, English**

## 1. Introduction

Machine Translation (MT) is considered as one of the important applications from the field of Natural Language Processing and MT developments are especially relevant in a multilingual country like India with 22 official langauges, almost 2000 dialects, 5 language families and 12 scripts (Garje and Kharate, 2013). With institutionalization of TDIL [1] (Mission for the Technology Development in Indian Languages), research and development in Language technology for Indian langauges are on the rise. But with the diverse linguistic situation found in India, only the major languages tend to get research focus. Through the development of a statistical machine translation (SMT) system for Garhwali (a Central Pahari Language spoken in the Garhwal region of Uttarakhand), an effort has been made to contribute and encourage language technology and language engineering for the lesser known languages of India.

## 2. Resource development

The fundamental requirement for a statistical machine translation system development is a consolidated corpus reserve. An English-Garhwali SMT system required Garhwali monolingual corpus and English-Garhwali parallel corpus. Since Garhwali language, like many other Indian languages, follows an oral tradition and has been passed on from generations to generations, there is not much considerable written corpus available. Hence, corpus collection and corpus development for Garhwali involved a time-consuming and elaborate undertaking.

### *A. Monolingual corpus creation*

A baseline SMT system requires a minimum of 10,000 monolingual Garhwali sentences (Koehn et al, 2007). The sentences were collected through web crawling, Bible translations in Garhwali and a Garhwali magazine through optical character recognition (Uniyal, 2018).

### *B. Parallel corpus creation*

Along with monolingual corpus, developing an SMT system also requires a minimum of 10,000 parallel sentences (sentence -aligned English-Garhwali translated text) (Koehn et al, 2007). The source language (SL) is English and the sentences to be translated were taken from the ILCI [2] Project; the target language (TL) is Garhwali in which the translations were made. Translations followed certain general principles like maintaining the syntactic structure of the target language and semantics of the source language. Also, it was ensured that correspondences to every linguistic unit (like punctuations, spaces etc.) were made (Uniyal, 2018).

## 3. System development

There are several approaches to machine translation, for example; rule-based, phrase-based, knowledge-based, transfer-based, interlingua-based etc. Statistical-based MT development is one of the popular approaches which is based on machine learning algorithms. Among Indian languages, major developments using statistical approach are EILMT (English to Indian Languages Machine Translation), ANUVAADAK, Google Translate etc.

There are multiple SMT system development platforms and toolkits that are available for training and developing a statistical machine translation system. The English-Garhwali SMT system has been deployed using the MOSES [3] toolkit. Moses consists of two important components – the training pipeline and the decoder (with some other complimentary tools and utilities). The major steps that were involved in the system training, after corpus preparation, are discussed below:

Tokenizing the sentences: The sentences in both monolingual and parallel corpus were tokenized (segmented) to maintain by using a JAVA command in MOSES.

*Filtering the corpus:* The long and complex sentences were filtered out so that the system has a standard length of a sentence. This is also done because long and complex sentences are difficult to be processed at the initial phase of system training.

*Language Model Training:* There are multiple language

models available compatible with MOSES, the present system uses the KenLM [4] language model.

Training of Translation System: Training the system requires GIZA++ to be installed first as it generates a directory in which translation commands and logs can be stored. Parallel sentences, which are 10,000 each in size were uploaded for system training.

*Tuning the System:* A small set of parallel sentences are required for tuning purpose. These sentences act as a reference inventory for the target language to derive contextual and syntactic information. A set of 500 sentences were assigned as tuning data.

With the 'tuning' phase, the baseline English-Garhwali SMT system is ready for the evaluation phase.

## 4. Evaluation

Evaluation (also known as the testing phase) is performed to accesses the translation results of the newly developed SMT system. The machine generated output is compared to human translation which are in the reference inventory. There are several automatic evaluation metrics available for evaluating the translation output; some of the major ones are METEOR, NIST, TER, WER etc. The evaluation metric used in this study is BLEU which exhibits higher similarity to human evaluation judgments (Graham and Baldwin, 2014). The score generated by the BLEU metric for the baseline English-Garhwali Statistical Machine Translation System is 10.37.

## 5. Conclusion

Considering that this is only the first phase of preparing an SMT system for the English-Garhwali language pair, the BLEU score of 10.37 is encouraging. The translation output exhibits several cases of translation divergence (mapping errors in the parallel data) and these cases encourage linguistic studies that can propose certain mapping rules for better translations. A baseline system is hoped to provide a jumpstart to a more consolidated system development. Hence, the next phase to the baseline English-Garhwali SMT system is a more sophisticated and competent SMT system which is developed with more training and reference corpus and have some linguistic rules plugged in to generate better translation results. This system development is also hoped to mainstream a lesser known language like Garhwali and increase the reach and services of language technology.

## References

[1] http://www.tdil.meity.gov.in/
[2] http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci
[3] https://kheafield.com/code/kenlm/
[4] https://en.wikipedia.org/wiki/BLEU
[5] Badodekar, S. (2003). Translation Resources, Services and Tools for Indian Languages. Computer Science and Engineering Department, Indian Institute of Technology, Mumbai, 400019, India.
[6] Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on systran's rule-based translation system. In Proceedings of the Second Workshop on SMT, 220-223. Dwivedi, S. K., & Sukhadeve, P. P. (2010). Machine Translation System in Indian Perspectives. Journal of Computer Science, 6(10), 1111-1116.
[7] Garje, G. V., and Kharate GK. (2013). Survey of Machine Translation in India. Department of Computer Engineering and Information Technology PVG's College of Engineering and Technology, Pune, India.
[8] Graham, Y., and Baldwin, T. (2014). Testing for Significance of Increased Correlation with Human Judgment. Department of Computing and Information Systems. The University of Melbourne.
[9] Koehn, P., Bertoldi F. and Moran C., Federico M., Cowan B., Hoang H., Birch A., Zens R.,Constantin A., Dyer C., Shen B., Bojar O.,Herbst E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation.
[10] Naskar, S., and Bandyopadhyay, S. (2005). Use of Machine Translation in India: Current Status. In Proceedings of MT SUMMIT X; September 13-15, 2005, Phuket, Thailand.
[11] Nainwani, Pinkey, 2015. Challenges in Automatic Translations of Natural Languages - a study of English-Sindhi Divergence, Jawaharlal Nehru University, New Delhi.