ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 1, pp: (127-132), Month: January - March 2019, Available at: www.researchpublish.com

English-Garhwali SMT System – Development and Evaluation

Arushi Uniyal

Jawaharlal Nehru University
Author Email id:*arushi.uniyal@gmail.com*

Abstract: **The present paper introduces language technology efforts being made for Garhwali, one of the lesser-known and under-researched languages of India. Language technology developments have been initiated for Garhwali through the development of an English-Garhwali statistical machine translation system. This paper presents the development of the English-Garhwali SMT system and discusses evaluation studies carried out to assess the translation results of the system. The development of a machine translation system is hoped to encourage further language engineering efforts in Garhwali and the evaluation study is aimed to identify translation errors that can then be resolved to increase system efficiency.**

Keywords: **Statistical Machine Translation, English-Garhwali, Evaluation.**

1. INTRODUCTION

Machine Translation (MT) in Indian languages has experienced significant research and development initiatives in the past few years and the noteworthy among these are AnglaBharati, Anuvaadak, Anusaarka, MaTra, Mantra, Sampark, Transmuter, UNL-based MT, Shakti and many others. There are different approaches to machine translation like Rule-Based, Interlingua-Based, Transfer-Based, Knowledge-Based etc. but Statistical Machine Translation (SMT) is observed to be the most sought-after approach to building a machine translation system and this is because SMT systems are corpus-based and statistical information drawn from a large corpus has seen to result in robust processing systems (Hutchins and Somers, 1992). This is especially true for less-researched languages like many Indian languages and an obvious research approach for a lesser known language like Garhwali, a central Pahari language spoken in the Garhwal region of the state of Uttarakhand. Developing a machine translation system for the language pair with English and Garhwali is an effort to firstly, digitalize Garhwali language and create language resources for future research and secondly, encourage more language engineering and language technology for Garhwali speakers.

The aim of MT systems is to produce translation results which are syntactically and semantically close to human translations and require minimum human post-processing. A competent machine translation system produces output that is understandable, acceptable and of good quality (Kalyani and Sajja, 2015). Hence, development of an MT system is followed by an evaluation process which comprises of identification of translation errors that need to be resolved to increase the performance of the MT system. The present paper discusses the design and development of an English-Garhwali statistical machine translation system followed by the evaluation process which includes discussion on the nature of errors observed in the system.

2. DEVELOPMENT OF AN ENGLISH-GARHWALI MACHINE TRANSLATION SYSTEM

Research and development of SMT systems is primarily dependent on the available corpus reserves as a consolidated monolingual and parallel (sentence aligned bitext) corpus is necessary for system training. The training of SMT systems is facilitated by different toolkits and on different platforms, the two training platforms that have been used to develop the present English-Garhwali SMT system are Moses and Microsoft Translator Hub (MTHub). In this section, there will be a discussion on the nature of corpus that has been developed and used for training purposes and a brief presentation of the two MT systems.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 1, pp: (127-132), Month: January - March 2019, Available at: www.researchpublish.com

(i) Corpus Creation

Developing an English-Garhwali SMT system first required a time-consuming and elaborate undertaking of creation of a corpus reserve for Garhwali as it is a less-resourced language without any substantial amount of text available. The monolingual corpus reserve in Garhwali comprised of 40,000 sentences which were general-domain and were collected from a monthly Garhwali magazine (using Optical Character Recognition), translations of Bible excerpts and some internet blogs (through web crawling). The parallel corpus was developed by manual translation of general-domain English sentences taken from the ILCI (Indian Languages Corpora Initiative) Project, which is a language technology initiative for Indian languages supported by the Ministry of Electronics and Information (MeitY), Government of India. Translations were guided by certain general principles to maintain the semantics of the source language English and the syntax of the target language Garhwali (Uniyal, 2018). The parallel dataset was a corpus size of 30,000 sentences and both monolingual and parallel data comprised of simple as well as complex sentence types to represent all types of sentence constructions in both these languages.

(ii) Moses-based English-Garhwali SMT

The introduction of Moses as an SMT toolkit in 2007 (Koehn et al, 2007) can be considered a milestone in the field of Natural Language Processing as it was an open source toolkit that presented possibilities for training new SMT systems. The most important feature of Moses is that it allows the users to make modifications in the internal architecture of the training system in accordance with the nature of the training corpus and the requirements of the MT system that is being developed. There are two major components in the Moses architecture - the decoder and the training pipeline and the system development steps include processes like tokenization, truecasing, filtering the corpus, training on the language model, tuning and testing.

The English-Garhwali SMT system was trained on the Moses platform which has been especially designed for Indian languages, developed at the School for Sanskrit and Indic Studies, Jawaharlal Nehru University. Following is a screenshot of the first stage of training of the English-Garhwali Translator –

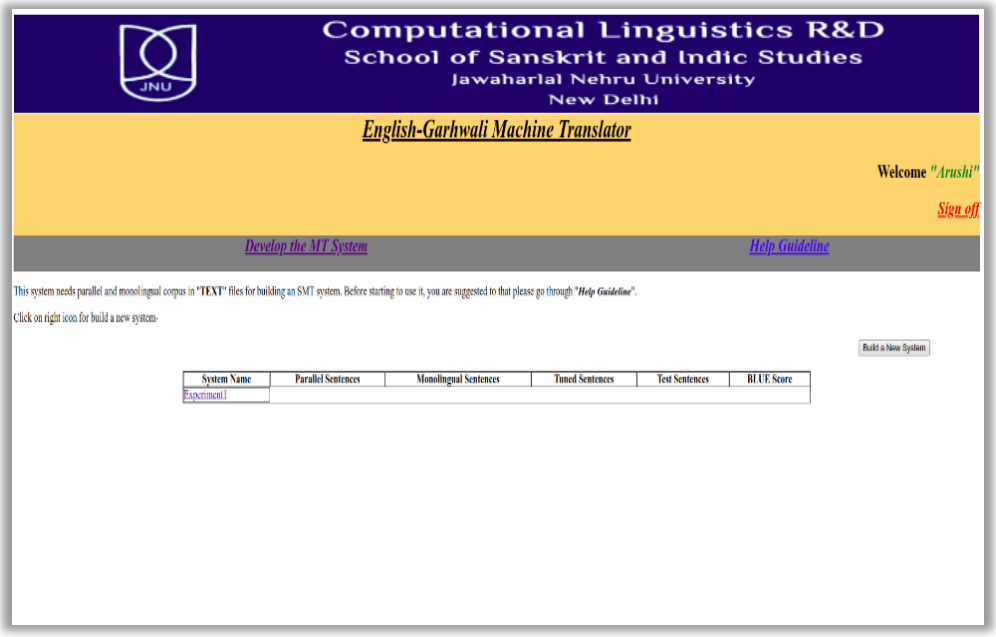

Fig 1: Moses-based English-Garhwali SMT system

(iii) Development of English-Garhwali SMT system on MTHub

MTHub was developed by Microsoft Corporation with the objective to enable users to design and develop customized MT systems which can be trained either on the pre-built language models of the already existing language pairs, in the system memory; or on completely new language pair(s) as well. A user can create a translation project on MTHub and train monolingual and parallel data to develop an MT system for a language pair which can be either general or domain specific. Following is the snapshot of the English-Garhwali MT project that was created and trained on MTHub –

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 1, pp: (127-132), Month: January - March 2019, Available at: www.researchpublish.com

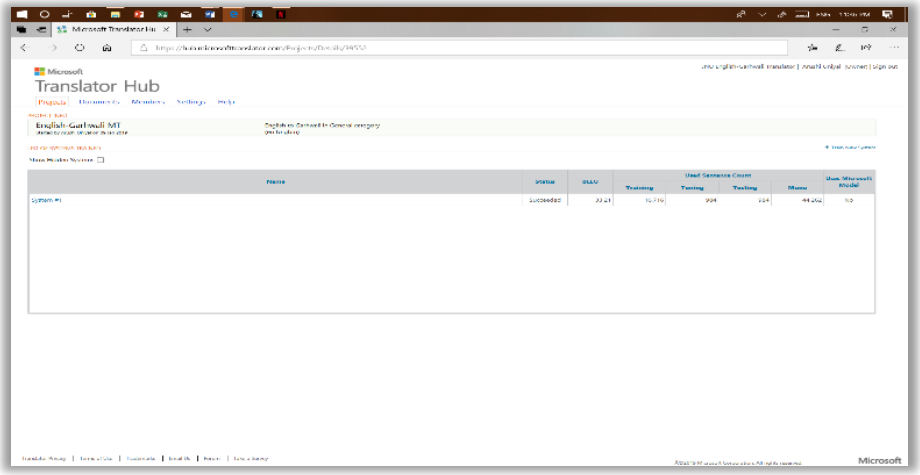

Fig 2: MTHub-based English-Garhwali SMT system

3. EVALUATION OF THE ENGLISH-GARHWALI MACHINE TRANSLATION SYSTEM

The competency of a machine translation system can be determined by different evaluation approaches which can be either categorized under an automatic evaluation approach or a human evaluation approach. The present English-Garhwali SMT system has been evaluated through both these approaches to check its efficiency and the following sections discuss and illustrate these results in detail –

A. Automatic Evaluation

Automatic Evaluation is based on different metrics and algorithms that are developed to access translation results of an MT system. The advantages of using automated evaluation is that it is faster, labor-efficient and unbiased (as compared to human evaluation) but the disadvantages are that the processing of semantics is not satisfactory in an automatic evaluation process.

There are several metrics that can be employed for evaluation of MT systems, the major ones are - Metric for Evaluation of Translation with Explicit Reordering (METEOR), National Institute of Standards and Techniques (NIST), Translation error Rate (TER), Precision and Recall, Word Error Rate (WER) etc. The present study, however, uses Bilingual Evaluation Understudy (BLEU) as it presents higher similarity to human evaluation judgements (Graham and Baldwin, 2014).

The BLEU scores are automatically generated at the end of system training in both Moses and MTHub platforms. The BLEU score for Moses-based system was recorded at 35.52 and for MTHub-based system at 33.21, below is the graphical representation of the BLEU scores for the two English-Garhwali systems –

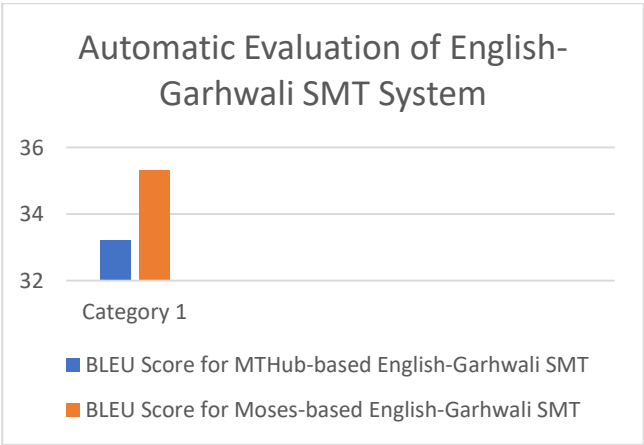

Fig 3: Automatic Evaluation of English-Garhwali SMT System

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 1, pp: (127-132), Month: January - March 2019, Available at: www.researchpublish.com

B. Human Evaluation

Human Evaluation is the manual assessment performed by a bilingual speaker by assigning scores to the translated sentences. The advantages of using the human evaluation approach is that the role of context in determining meaning to the sentence is accurately processed. The disadvantage, however, is that manual evaluation is time-consuming and highly subjective. The major human evaluation metrics are Ranking, Intra-Evaluator Agreement, Fluency, Adequacy etc. The present study uses Fluency and Adequacy metric for manual evaluation of the two English-Garhwali MT systems.

Human Evaluation was conducted on 500 translated sentences each from the two English-Garhwali SMT systems and these translations were scored by three human evaluators for Adequacy and Frequency. The scoring schema to be followed for Frequency was –

Table 1: Scoring Schema for Frequency

Parameters	**Scale**
Perfect Translation	5
Good Translation	4
Non-Native Translation	3
Disfluent Translation	2
Incomprehensible Translation	1

The scoring schema to be followed for Adequacy was –

Table 2: Scoring Schema for Adequacy

Parameters	**Scale**
All Meaning	5
Most Meaning	4
Much Meaning	3
Little Meaning	2
None	1

Cumulative scores were generated using the formula = Sum total of scores/Number of sentences.

The cumulative scores for 500 sentences for Moses-based system was Adequacy at 65.37 and Fluency at 54.24; and for MTHub-based system, Adequacy at 52.64 and Fluency at 48.10. Following is the graphical representation of the Adequacy and Fluency scores for the two English-Garhwali systems –

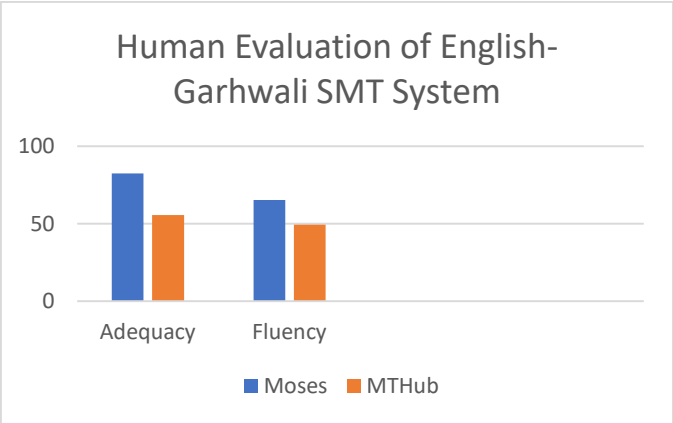

Fig 3: Human Evaluation of English-Garhwali SMT System

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 1, pp: (127-132), Month: January - March 2019, Available at: www.researchpublish.com

4. ERROR ANALYSIS

This section presents the nature of translation errors that were observed in the output of the English-Garhwali SMT system due to 'decoding' issues.

(i) Translation Errors due to punctuation marks(s)

English sentence – This is 2.5 kilograms rice.

Garhwali translation - यो 2.

ITRANS - Yo dwi
Gloss - This two

The period marker (.) was observed to generate a significant number of translation errors and this was because while in English a period marker is used as a full stop and as a decimal point; in Garhwali the system was, at times (like in the presented example), processing it only as a full stop marker and hence was breaking the sentence to two. It is to be noted that this problem was encountered only in the MTHub-based system and not in the Moses-based system.

(ii) Translation Errors due to Transliteration

English sentence - The leaves are flexible.

Garhwali translation - पत्ती flexible हूंदि ।

ITRANS - patti phaleksibal hundi
Gloss - leaves flexible are

In this type of error, the translation system produces the accurate structure for the target language but is unable to transliterate a unit and produces it in the source language format. This type of errors is observed in both the MT system although the frequency of this type of error is relatively low.

(iii) Translation Error due to lack of lexicon

English Sentence - This is everlasting flower.

Garhwali translation - यो एवरलास्टिंग फ्लावर च ।

ITRANS - yo aivarlAstiiMg phalAvar cha
Gloss - this everlasting flower is

In the presented example, the English word 'everlasting' did not find an equivalent in the Garhwali lexicon and hence was merely transliterated but not translated. Such errors occur because of culture specific vocabulary and hence any language pair has a fairly large set of such words, hence both these systems present a considerable set of such words.

(iv) Translation Error due to gender agreement

English sentence – He is going.

Garhwali Translation - श्या जाणी च

ITRANS - ShyA jANi cha
Gloss - she going is

As can be observed in the following example, the translation system does not process gender agreement properly at times, but this is a rare occurrence in the two MT systems.

5. CONCLUSION

The present paper was an effort to introduce the two English-Garhwali statistical machine translation systems, developed on Moses and MTHub respectively, presented in section 2. This was followed by an assessment of the translation performance of these two systems through automatic and human evaluation metrics in section 3. Section 4 was a discussion on the nature of translation errors that were observed due to system 'decoding' issues.

This work is hoped to encourage resolutions for the discrepancies of the present English-Garhwali SMT system which could result in a better translation system for this language pair. Also, it is seen as a starting point for research and development in both theoretical and applied language studies for a less researched language like Garhwali.

REFERENCES

[1] 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.

[2] Garje, GV., & 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.

[3] Graham, Y., & Baldwin, T. (2014). Testing for Significance of Increased Correlation with Human Judgment. Department of Computing and Information Systems. The University of Melbourne.

[4] Hutchins, WJ & Somers, HL. (1992). An Introduction to Machine Translation. London Academic Press.

[5] Kalyani. A & Sajja. P (2015). A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies. International Journal of Computer Science.

[6] Koehn, P., Bertoldi F. & 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.

[7] Naskar, S., & Bandyopadhyay, S. (2005). Use of Machine Translation in India: Current Status. In Proceedings of MT SUMMIT X; September 13-15, 2005, Phuket, Thailand.

[8] Nainwani, Pinkey, 2015. Challenges in Automatic Translations of Natural Languages - a study of English-Sindhi Divergence, Jawaharlal Nehru University, New Delhi.

[9] Uniyal, A (2018). Developing Garhwali Corpus for Statistical Machine Translation, Center for Linguistics, Jawaharlal Nehru University.