Sādhanā (2022)47:106
https://doi.org/10.1007/s12046-022-01855-w

Explicitly unsupervised statistical machine translation analysis on five Indian languages using automatic evaluation metrics

SHEFALI SAXENA*, SHWETA CHAUHAN, PARAS ARORA and PHILEMON DANIEL

Electronics and Communication Department, National Institute of Technology, Hamirpur, HP, India
e-mail: shefali@nith.ac.in; shweta@nith.ac.in; parastakkar88@gmail.com; phildani7@nith.ac.in

MS received 3 November 2021; revised 7 February 2022; accepted 11 February 2022

Abstract. This letter, presents the compendium of eight unsupervised Machine Translation (MT) systems built from monolingual corpus of five Indian languages from the Indo-Aryan and Dravidian language families. Recent research has demonstrated outstanding results in completely unsupervised training of Phrase-based Statistical MT (PBSMT) systems using innovative and designs that rely solely on monolingual datasets. However, prior research has shown that Unsupervised Statistical MT (USMT) outperforms Unsupervised Neural MT (UNMT), particularly for language pairings that are not closely related. The purpose of this work is to investigate the architecture of the USMT system utilizing only monolingual dataset using four different Indian morphologically rich languages and one low-resource endangered Kangri language. The experimental results analysis are evaluated using different natural language toolkit tokenizers and analyzed for different language pair using various fully automatic MT evaluation metrics for different iterations.

Keywords. Machine translation; statistical machine translation; low-resource language; evaluation.

1. Introduction

Machine Translation (MT) is a well-explored research area, still the development of MT for Indic language is in a nascent stage. India is one of the greatest examples of a multilingual and multicultural country as it provides room to a significant number of related languages. According to a recent study of United Nations Educational, Scientific, and Cultural Organization, there are 197 Indian languages in peril, seven of which are Himachal language, with Kangri being one of them [1]. Kangri is an Indo-Aryan language that is spoken throughout North Asia, including today's Himachal Pradesh.

Traditional human translation can provide high-quality results, but it is inefficient, expensive, and requires more effort, and cannot fulfil a wide range of translation requirements [2]. To achieve good quality of MT, the standard Statistical MT (SMT) [2] necessitates a massive quantity of bilingual text in both the source as well as at the destination language.

1.1 *Motivation, challenges and contribution*

In recent years, the unsupervised MT job has been quite effective. The majority of us are fluent in our own language but struggle to comprehend other languages. As a result, translation is critical in both speech and text. A large volume of parallel data was required for text translation. Even though significant research has been done on Indian languages, there are still a number of low-resource Indian languages where the MT problem has yet to be tackled.

Kangri (ISO 639-3xnr), a low-resource Indian language of the Indo-Aryan language family; dataset has been developed, as described in [1]. When translating from a high-resource language pair to a Low Resource Language (LRL) pair, MT is always a challenge. Another difficulty is that Indian languages are extremely agglutinative and morphologically rich, resulting in linguistic diversity and changes in word forms, making automated n-grams metrics matching like BLEU ineffective [3].

To the best of our knowledge, this is the first analysis of the Unsupervised SMT (USMT) model for Indic languages. This work focuses on the analysis of text MT tasks, particularly for Indic languages, and, secondarily, for definitely endangered LRL using monolingual data alone. In order to illustrate the efficacy of USMT system, eight translation tasks [English (En), Hindi (Hi), Tamil (Ta), and Telugu (Te)], including a definitely endangered low-resource kangri (Kn) language (i.e., English-Kangri and Hindi-Kangri language pairs), using monolingual data only to address parallel data scarcity and investigate the effectiveness/ ineffectiveness of the method. Various automatic MT evaluation metrics were used to study the MT tasks, including National Institute of Standards and Technology (NIST) [4], Translation Error Rate (TER) [5], Word Error

*For correspondence

Published online: 02 June 2022

Rate (WER) [6], and Match Error Rate (MER) [7], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [8] and BiLingual Evaluation Understudy (BLEU) [9]. To investigate the morphologically rich languages, various natural language toolkit tokenizers have combined with fully automatic MT evaluation metrics. The experimental result demonstrate that a state-of-the-art fully USMT system based on monolingual dataset is advantageous for morphologically languages, particularly for Indic and a Low resource languages.

2. Theory and model

The standard SMT system has three main models: language model (LM), translation model and decoder. The basic principle for this approach is based on the Bayesian's theorem stating as with the sentence d in the source language (SL) of the input, the system finds a sentence c in target language (TL) to get the maximum value of $P\left(\frac{c}{d}\right)$ is evaluated as (eq. 1)

$$argmax_{e}p(c|d) = argmax_{e}p(d|c)p(c) \tag{1}$$

where, $P\left(\frac{c}{d}\right)$ is conditional probability of translating an SL sentence into the TL sentence. The TM is determined using an aligned bilingual corpus and checking how the output in the TL matches the input in the SL.

The systematic diagram of the model is shown in figure 1. The whole process of USMT system can be summarized as follows. This [10] uses the standard methods as a baseline system. However, it replaces the bilingual corpus by generating the n-gram embeddings for the L1 and L2 monolingual corpus. Through self-learning [11], the learnt word as well as phrase patterns were map to a cross-lingual space. The generated mapped phrase is combined with a LM and a distance-based distortion model to create a typical phrase-based SMT system. An iterative minimum error rate training variation is employed to tune the model's weights in an unsupervised way, and then the entire system is further improved by iterative back-translation.

3. Results and discussions

3.1 *Dataset*

For Indic languages, the IndicNLP toolkit has been used, and for English, the mosesdecoder toolkit has been used. The mosesdecoder truecaser.perl and clean-corpus.perl has been used for truecasing and cleaning the corpus.

The IndiCorp corpus for English, Hindi, Tamil and Telugu [12] and the kangri corpus [1] were taken for training; and for testing WMT [13] English, Hindi, Tamil and Telugu test corpus has been used. The information of the monolingual corpus in the preprocessing phase is displayed in table 1.

To reduce the accidental errors, all baseline system are run for eight MT tasks for five different language pairs and for five iterations to get the best iteration for the USMT model. The evaluation of the trained corpus has been evaluated on the En, Hi, Kn, Ta, and Te and the dataset based on BLEU, NIST, TER, WER, METEOR, and MER.

3.2 *Evaluation criteria*

The standard SMT model require a significant amount of bilingual data to attain satisfactory translations. The availability of the parallel corpus for low resource language for high-quality translation is limited. Therefore, based on the current research line, this work shows the analysis of the SMT model for major Indic languages and Kangri morphologically rich and low resource endangered language. The testing of the SMT system is performed on sentences other than the training corpus.

Figure 2 depicts the increase in BLEU points for En-hi as the corpus size is varied.

The Natural Language Toolkit is popularly used toolkit for NLP applications. Tokenizers can be chosen from a variety of parameters in the nltk.tokenize module. Six word tokenizers are used in this paper, including sentence, blankline, whitespace, punctuation-based, treebank, and Multiword expression tokenizers. Above of all the tokenizers, it was discovered that the sentence tokenizer performs best for all of the evaluation metrics. It has been discovered that the METEOR is only effective when using the Blankline tokenizer.

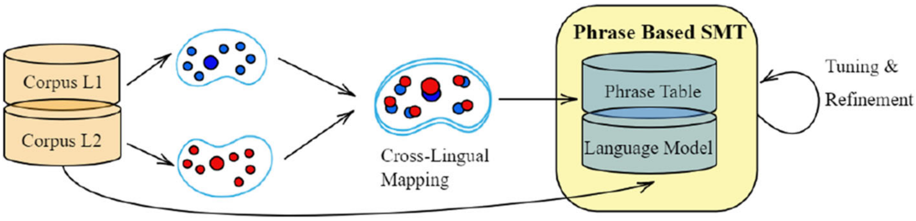


Figure 1. Systematic diagram of USMT model.

Table 1. Statistics of the corpus used.

		English	Hindi	Tamil	Telugu	Kangri
Script		Latin	Devnagri	Brahmi	Brahmi	Tapri (now Devnagri)
Alphabet Size		52	46	30	56	58
Mono lingual	Sentences	54.3M	63.1M	31.5M	674M	1.81M
	Words	1.22B	1.86B	582M	674M	22.7M
Test	Sentences	1000	1000	1000	1000	500

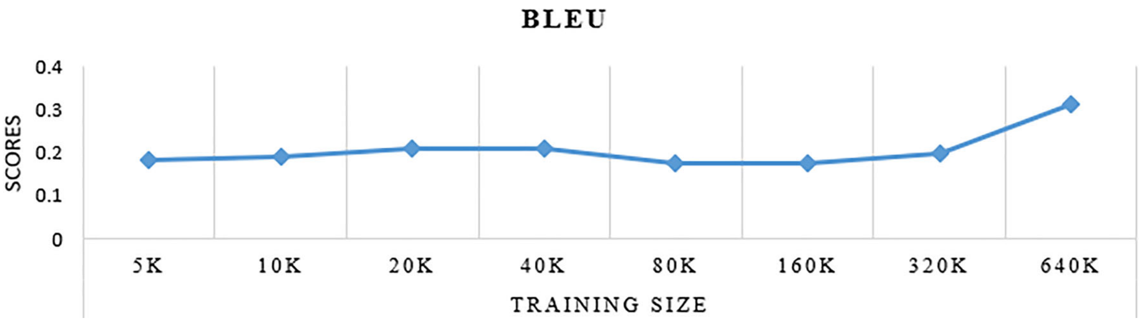
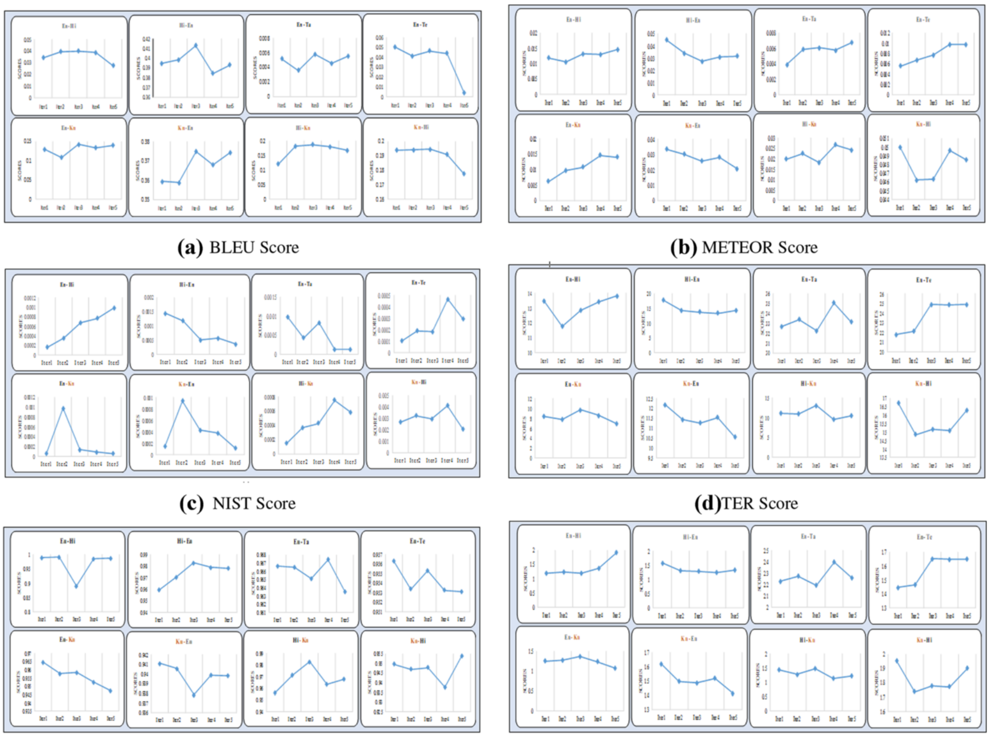


Figure 2. Variation in the BLEU assessment metric with the English-Hindi varying training size.



(a) BLEU Score

(b) METEOR Score

(c) NIST Score

(d)TER Score

(e) MER Score

(f) WER Score

Figure 3. The score for eight translation jobs in seven directions, i.e., En→Ta, En→Te, En↔Hi, Kn↔Hi, Kn↔En, on six different assessment criteria, up to five iterations with sentence tokenizers. (**a**) BLEU Score (**b**) METEOR Score (**c**) NIST Score (**d**) TER Score (**e**) MER Score (**f**) WER Score. The Kangri (Kn) endangered LRL's output translation is marked with orange colour.

The results are shown in figure 3, indicating variations for different types of datasets or experiments on different tasks for various iterations using sentence tokenizer. Based on a large number of experiments, USMT shows the best results with the third Iterative Back Translation for most translation tasks. These experiments show that the state-of-the-art USMT system works best for the same domain languages. Though automatic evaluation of MT systems has proved to be a miracle for the MT evaluation field, one cannot directly depend on these scoring techniques, as Indic languages are morphologically rich languages. MT Evaluation techniques suffer from several deficiencies or we can say a large number of issues are concerned with the MT evaluation algorithm [7].

- The first issue with morphologically rich languages is that the same sentence can have multiple translations.
- The second issue is that most of the evaluation systems focus on partial aspects of the quality of translations (e.g., Lexical and fluency) but not on the overall quality of the translations.
- The third is that most algorithms work only for specific domains, or we can say that they are domain-specific or domain-dependent.
- The fourth issue is the difficulty to decide the quality of the translation, and there are no fixed standard measures for comparing two systems.

We evaluated the system's performance in a variety of languages as well as on LRL, where data quality is often poor. The studies conducted for each language pair by keeping the source and target domains separate and testing the system's performance in all testing scenarios.

As training synthetic parallel data for the refining phases, [10] recommend using mixed back-and-forth translations. It provides a new feature for the induced phrase table based on the orthographic similarity between the source and target phrases, among other changes. The resultant USMT system excels European language pairings. There were no results for genuinely distant and LRL pairings, or Asian languages, in their study. We intend to study interesting reordering approaches in the future to improve the USMT architecture.

4. Conclusion

The work presented an USMT system on low resource settings for eight MT tasks with four Indian languages and one definitely endangered low resource Kangri language. The experimental results were analyzed for the USMT system for different morphologically rich language pairs including different iterations. We showed the comparison by using various fully automatic MT evaluation metrics using different tokenizer, for different iterations to get clear understanding of USMT model on morphologically richer languages. With the output translation quality, it is observed that the translation involving Indo-Aryan languages produces a good accuracy level. The translation results for English to Indic and vice-versa are also reported.

In our future work, we plan to extend this preliminary unsupervised setting to improve the translation quality by incorporating linguistic features and devising an improved initialization step and better using the synthetic data suitable for the Indic and Kangri language pair.

References

[1] Chauhan S, Saxena S and Daniel P 2021 Monolingual and Parallel Corpora for Kangri Low Resource Language. arXiv preprint arXiv:2103.11596

[2] Koehn P, Och F J and Marcu D 2003 Statistical phrase-based translation. University of Southern California Marina Del Rey Information Sciences Inst

[3] Ananthakrishnan R and Pushpak Bhattacharyya and M Sasikumar and Ritesh M Shah 2007 Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU

[4] Przybocki M, Peterson K, Bronsart S and Sanders G 2009 The NIST 2008 Metrics for machine translation challenge–overview, methodology, metrics and results. *Machine Translation*, 23(2): 71–103

[5] Snover M, Dorr B, Schwartz R, Micciulla L and Makhoul J 2006 A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* , pp. 223–231

[6] Su K Y, Wu M W and Chang J S 1992 A new quantitative quality measure for machine translation systems. In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*

[7] Malik P and Baghel A S 2018 A summary and comparative study of different metrics for machine translation evaluation. In: *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 55–60. IEEE

[8] Banerjee S and Lavie A 2005 METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72

[9] Papineni K, Roukos S, Ward T and Zhu W J 2002 Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318

[10] Artetxe M, Labaka G and Agirre E 2018 Unsupervised statistical machine translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3632–3642

[11] Artetxe M, Labaka G and Agirre E 2018 A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:1805.06297

[12] Kakwani D, Kunchukuttan A, Golla S, Gokul N C, Bhattacharyya A, Khapra M M and Kumar P 2020

IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. Findings of the Association for Computational Linguistics: EMNLP 2020, Nov, 2020 Association for Computational Linguistics, pp. 4948–4961

[13] Post M, Callison-Burch C and Osborne M 2012 Constructing parallel corpora for six Indian languages via crowdsourcing. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 401–409. Association for Computational Linguistics