System Description of BV-SLP for Sindhi-English Machine Translation in MultiIndic22MT 2024 Shared Task

Nisheeth Joshi[1#], Pragya Katyayan[2*], Palak Arora[1ǂ], Bharti Nathani[4]**
[1]Speech and Language Processing Lab, Banasthali Vidyapith, Rajasthan, India
[2]School of Computer Science, University of Petroleum and Energy Studies, Uttrakhand, India
[#]nisheeth.joshi@rediffmail.com, [*]pragya.katyayan21@gmail.com, [ǂ]palak.arora.pa55@gmail.com,
[**]nbharti@banasthali.in

Abstract

This paper presents our machine translation system that was developed for the WAT2024 MultiIndic MT shared task. We built our system for the Sindhi-English language pair. We developed two MT systems. The first system was our baseline system where Sindhi was translated into English. In the second system, we used Hindi as a pivot for the translation of text. In both the cases, we had identified the name entities and translated them into English as a preprocessing step. Once this was done, the standard NMT process was followed to train and generate MT outputs for the task. The systems were tested on the hidden dataset of the shared task

1 Introduction

This paper presents the system description of our neural machine translation system developed for the MultiIndic shared task organized at WMT 2024. We collected around two lac English-Hindi parallel corpus from Press Information Bureau's website[1] which had collection of news articles in English as well as in Hindi and then translated it into Sindhi (in Devanagari script). Thus, two NMT systems were trained on using this corpus. The first system was the baseline system which was trained using the Sindhi-English language pair. The second system had two NMT systems, Sindhi-Hindi and Hindi-English. This system used Hindi as the pivot language for translation.

2 System Overview

2.1 Preprocessing

Here, we did tokenization of text and also performed spelling correction. Then named entities from Sindhi text were extracted using the Bi-LSTM Sindhi POS tagger that was developed in-house (Nathani et al. 2023). The identified named entities were then classified into MUC-6 category (Grishman et al. 1996) through a rule-based approach. These tagged named entities were searched in a knowledge base which had translations of Sindhi/Hindi Organization and Location named entities in English. We extracted the named entities from the Sindhi/Hindi corpus using a rule-based NER system. Once the named entities were extracted, they were searched in a knowledge base that had translations of these named entities in English (Organization and Location names). If they were found then the same were replaced in the Sindhi/Hindi Corpus. In cases where the named entity translations were not present in the knowledge base, then they were transliterated and were replaced in Sindhi/Hindi corpus. This became our Named Entity Translation module which identified the named entities and accordingly translated/transliterated them into English (Sharma et al. 2023; Joshi & Katyayan 2023). The work of this module is shown in Figure 1.

2.2 Byte Pair Encoding

Here the source and the target corpus were divided into smaller units known as subwords. This task was performed to convert the words into smaller basic units which helped neural MT models in better handling of out of vocabulary (OOV) words.

2.3 Training of the Model

In training both systems we applied the same steps. For system 1 which was the baseline system, we had only one named entity translation module (Sindhi-English) while for the system 2 the named entity translation module performed Hindi-English translation/transliteration. The process followed

[1] https://pib.gov.in/

Proceedings of the Ninth Conference on Machine Translation, pages 793–796

was; the POS tagging of Sindhi sentence was performed and NER was performed using a rule-based module. The identified named entities were translated as explained in the previous section. This produced an augmented corpus-based source sentence. For example, let us consider a Sindhi sentence, "**निशीथ जोशी नई दिल्ली जे इंदिरा गांधी अंतर्राष्ट्रीय हवाई अड्डे खां जयपुर जो सफर करे रहियो आहे।**" Here "**निशीथ जोशी** (Person)", "**नई दिल्ली** (location)", "**जयपुर** (location)", and "**इंदिरा गांधी अंतर्राष्ट्रीय हवाई अड्डे** (organization)" are named entities. Among these since "**निशीथ जोशी**" and "**जयपुर**" were not available in the knowledge base, so they were transliterated to "Nisheeth Joshi" and "Jaipur" respectively. The rest of the named entities had their categories in the knowledge base; thus, they were looked up in a sequential manner. "**नई दिल्ली**" was not found and was transliterated to "New Delhi", similarly "**इंदिरा गांधी अंतर्राष्ट्रीय हवाई अड्डे**" was translated to Indra Gandhi International Airport". The entire training corpus was augmented using this methodology. Figure 2 shows the working of the entire system.

The hyperparameters used in training both the systems are shown in table 1.

Parameter	Value
No. of Encoding Layers	6
No. of Decoding Layers	6
Early Stopping	
metric	bleu
min_improvement	0.2
steps	6
Optimizer	Adam
beta_1	0.8
beta_2	0.998
learning_rate	1.0
droupout	0.25
Regularization	
type	l1_l2
scale	1e-4
Minimum_learning_rate	0.00001
Max_steps	1000000

Tabel 1: Hyperparameters Used in Training NMT Models

3 Evaluation

We participated in the shared task using the hidden corpus and submitted the outputs for both the systems viz baseline and pivot MT systems. The results of the same are shown in table 2.

System	BLEU	chrF	chrF++
System 1	19.4	44.6	43
System 2	20	44.7	43.2

Tabel 2: Evaluation Results

The baseline system which translated Sindhi text into English had a BLEU score (Papineni et al. 2002) of 19.4, chrF score (Popović 2015) of 44.6 and chrF++ score (Popović 2017) of 43. From a human annotators perspective, this system produced fluent translations but in some cases lacked the desired quality. The second system which used Hindi as a pivot language (where Sindhi was translated into Hindi and then this Hindi translation was translated into English) produced slightly better results. Its BLEU score was 20, chrF score was 44.7 and chrF++ score was 43.2. This system generated translation which had improved adequacy and fluency scores.

Acknowledgement

This work is supported by the funding received from the Ministry of Electronics and Information Technology, Government of India for the project "English to Indian Languages and vice versa Machine Translation System" under National Language Translation Mission (NLTM): Bhashini through administrative approval no. 11(1)/2022-HCC(TDIL) Part 5 and funding received from Anusandhan National Research Foundation (previously, Science Engineering and Research Board), Government of India through grant number SPG/2021/003306 for project entitled, "Development of Sindhi-Hindi and Hindi-Sindhi Machine Assisted Translation System".

References

Grishman, R., & Sundheim, B. M. 1996. *Design of the MUC-6 evaluation. In TIPSTER TEXT PROGRAM PHASE II*: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996 (pp. 413-422).

Joshi, N., & Katyayan, P. 2023. Improving English-Bharti Braille Machine Translation Through Proper Name Entity Translation. In ICIDSSD 2022: Proceedings of the 3rd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2022, 24-25 March 2022,

New Delhi, India (p. 168). European Alliance for Innovation.

Nathani, B., Arora, P., Joshi, N., Katyayan, P., Rathore, S. S., & Dadlani, C. P. 2023. *Sindhi POS Tagger Using LSTM and Pre-Trained Word Embeddings*. In XVIII International Conference on Data Science and Intelligent Analysis of Information (pp. 37-45). Cham: Springer Nature Switzerland.

Papineni, K., Roukos, S., Ward, T., & 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).

Popović, M. 2015. *chrF: character n-gram F-score for automatic MT evaluation*. In Proceedings of the tenth workshop on statistical machine translation (pp. 392-395).

Popović, M. 2017. *chrF++: words helping character n-grams*. In Proceedings of the second conference on machine translation (pp. 612-618).

Sharma, R., Katyayan, P., & Joshi, N. 2023. *Improving the quality of neural machine translation through proper translation of name entities*. In 2023 6th International Conference on Information Systems and Computer Networks (ISCON) (pp. 1-4). IEEE.

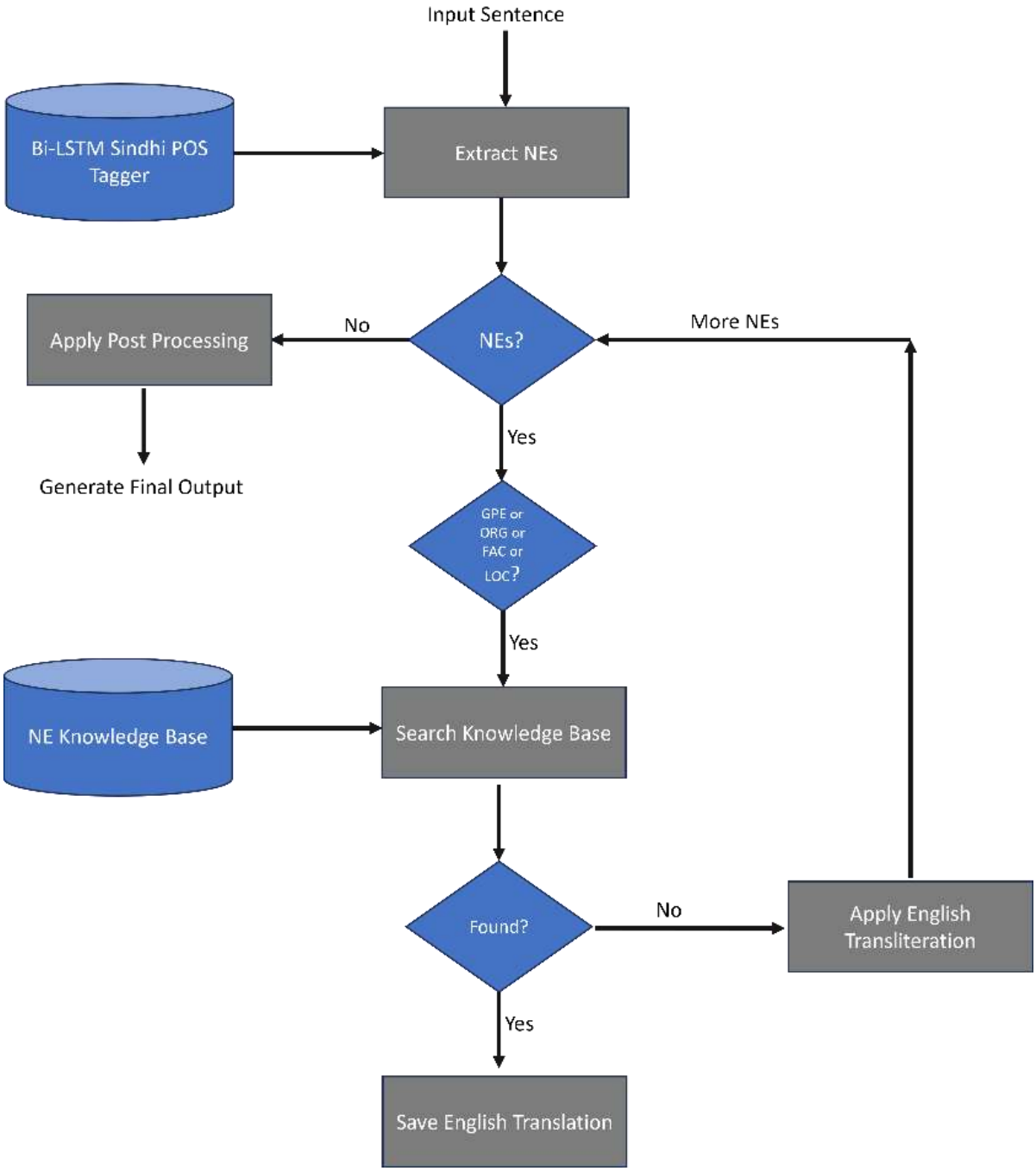

Figure 1: Named Entity Translation Module

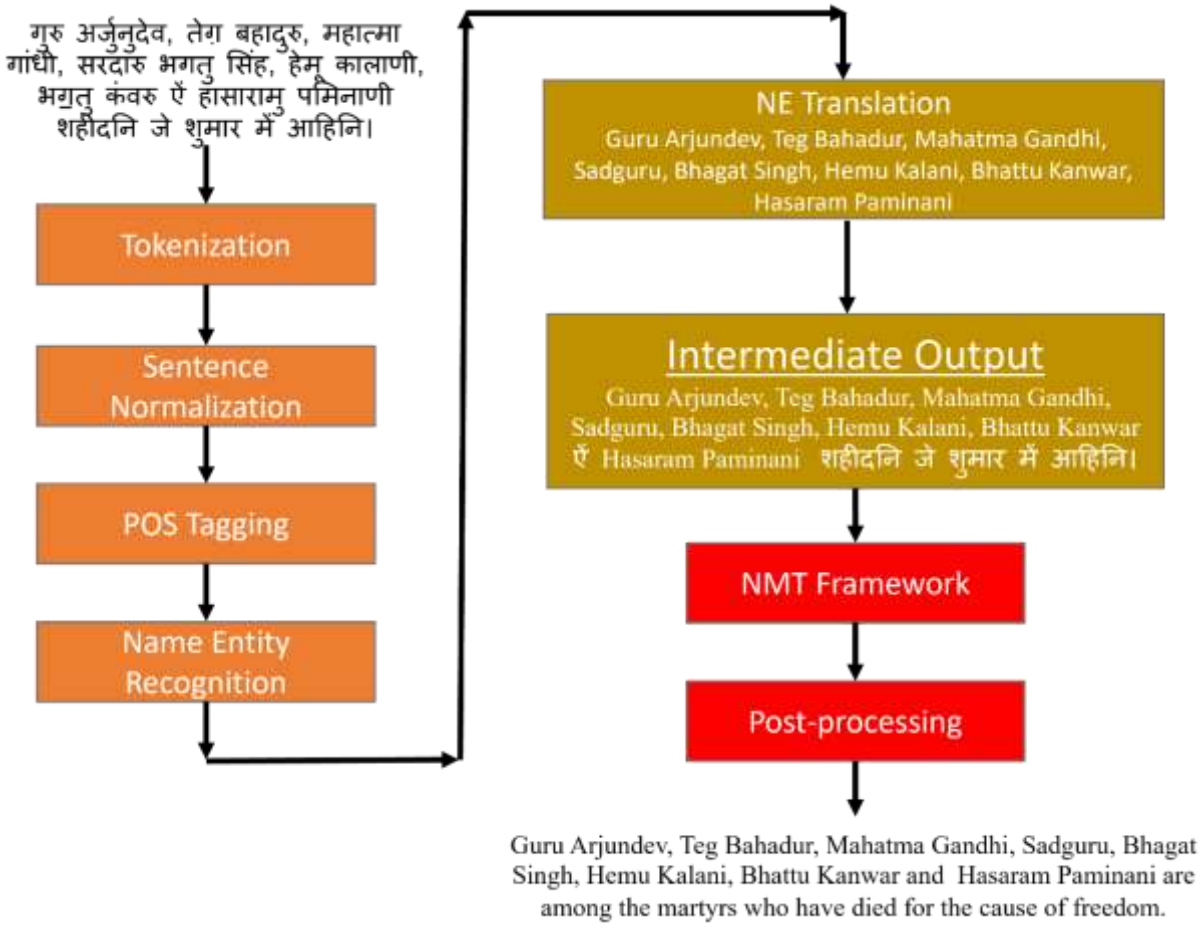

Figure 2: BV-SLP NMT Approach