Punjabi to English Bidirectional NMT System

Kamal Deep
Department of Computer Science
Punjabi University, Punjab, India
kamal.1cse@gmail.com

Ajit Kumar
Department of Computer Science
Multani Mal Modi College, Punjab, India
ajit8671@gmail.com

Vishal Goyal
Department of Computer Science
Punjabi University, Punjab, India
vishal.pup@gmail.com

Abstract

Machine Translation is ongoing research for last few decades. Today, Corpus-based Machine Translation systems are very popular. Statistical Machine Translation and Neural Machine Translation are based on the parallel corpus. In this research, the Punjabi to English Bidirectional Neural Machine Translation system is developed. To improve the accuracy of the Neural Machine Translation system, Word Embedding and Byte Pair Encoding is used. The claimed BLEU score is 38.30 for Punjabi to English Neural Machine Translation system and 36.96 for English to Punjabi Neural Machine Translation system.

1 Introduction

Machine Translation (MT) is a popular topic in Natural Language Processing (NLP). MT system takes the source language text as input and translates it into target-language text(Banik et al., 2019). Various approaches have been developed for MT systems, for example, Rule-based, Example-based, Statistical-based, Neural Network-based, and Hybrid-based(Mall and Jaiswal, 2018). Among all these approaches, Statistical-based and Neural Network-based approaches are most popular in the community of MT researchers. Statistical and Neural Network-based approaches are data-driven(Mahata et al., 2018). Both need a parallel corpus for training and validation(Khan Jadoon et al., 2017). Due to this, the accuracy of these systems is higher than the Rule-based system.

The Neural Machine Translation (NMT) is a trending approach these days(Pathak et al., 2018). Deep learning is a fast expanding approach to machine learning and has demonstrated excellent performance when applied to a range of tasks such as speech generation, DNA prediction, NLP, image recognition, and MT, etc. In this NLP tools demonstration, Punjabi to English bidirectional NMT system is showcased.

The NMT system is based on the sequence to sequence architecture. The sequence to sequence architecture converts one sequence into another sequence(Sutskever et al., 2011). For example: in MT sequence to sequence, architecture converts source text (Punjabi) sequence to target text (English) sequence. The NMT system uses the encoder and decoder to convert input text into a fixed-size vector and generates output from this encoded vector. This Encoder-decoder framework is based on the Recurrent Neural Network (RNN)(Wołk and Marasek, 2015)(Goyal and Misra Sharma, 2019). This basic encoder-decoder framework is suitable for short sentences only and does not work well in the case of long sentences. The use of attention mechanisms with the encoder-decoder framework is a solution for that. In the attention mechanism, attention is paid to sub-parts of sentences during translation.

2 Corpus Development

For this demonstration, the Punjabi-English corpus is prepared by collecting from the various online resources. Different processing steps have been done on the corpus to make it clean and useful for the training. The parallel corpus of 259623 sentences is used for training,

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 7–9
Patna, India, December 18 - 21, 2020. ©2019 NLP Association of India (NLPAI)

development, and testing the system. This parallel corpus is divided into training (256787 sentences), development (1418 sentences), and testing (1418 sentences) sets after shuffling the whole corpus using python code.

3 Pre-processing of Corpus

Pre-processing is the primary step in the development of the MT system. Various steps have been performed in the pre-processing phase: Tokenization of Punjabi and English text, lowercasing of English text, removing of contraction in English text and cleaning of long sentences (# of tokens more than 40).

4 Methodology

To develop the Punjabi to English Bidirectional NMT system, the OpenNMT toolkit(Klein et al., 2017) is used. OpenNMT is an open-source ecosystem for neural sequence learning and NMT. Two models are developed: one for translation of Punjabi to English and the second for translation of English to Punjabi. The Punjabi vocabulary size of 75332 words and English vocabulary size of 93458 words is developed in the pre-processing step of training the NMT system. For all models, the batch size of 32 and 25 epochs for training is fixed. For the encoder, BiLSTM is used, and LSTM is used for the decoder. The number of hidden layers is set to four in both encode and decoder. The number of units is set to 500 cells for each layer. BPE(Banar et al., 2020) is used to reduce the vocabulary size as the NMT suffers from the fixed vocabulary size. The Punjabi vocabulary size after BPE is 29500 words and English vocabulary size after BPE is 28879 words. “General” is used as an attention function.

By using Python and Flask, a web-based interface is also developed for Punjabi to English bidirectional NMT system. This interface uses the two models at the backend to translate the Punjabi text to English Text and to translate English text to Punjabi text. The user enters input in the given text area and selects the appropriate NMT model from the dropdown and then clicks on the submit button. The input is pre-processed, and then the NMT model translates the text into the target text.

Model	BLEU score
Punjabi to English NMT model	38.30
English to Punjabi NMT model	36.96

Table 1: BLEU score of both models

5 Results

Both proposed models are evaluated by using the BLEU score(Snover et al., 2006). The BLEU score obtained at all epochs is recorded in a table for both models. Table 1 shows the BLEU score of both models. The best BLEU sore claimed is 38.30 for Punjabi to English Neural Machine Translation system and 36.96 for English to Punjabi Neural Machine Translation system.

References

Nikolay Banar, Walter Daelemans, and Mike Kestemont. 2020. Character-level Transformer-based Neural Machine Translation, arXiv: 2005.11239.

Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya, Siddhartha Bhattacharyya, and Jan Platos. 2019. Statistical-based system combination approach to gain advantages over different machine translation systems. *Heliyon*, 5(9):e02504.

Vikrant Goyal and Dipti Misra Sharma. 2019. LTRC-MT Simple & Effective Hindi-English Neural Machine Translation Systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation,Hong Kong, China*, pages 137–140.

Nadeem Khan Jadoon, Waqas Anwar, Usama Ijaz Bajwa, and Farooq Ahmad. 2017. Statistical machine translation of Indian languages: a survey. *Neural Computing and Applications*, 31(7):2455–2467.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush, Josep Crego, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source Toolkit for Neural Machine Translation. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*:67–72.

Sainik Kumar Mahata, Soumil Mandal, Dipankar Das, and Sivaji Bandyopadhyay. 2018. SMT vs NMT: A Comparison over Hindi & Bengali Simple Sentences. In *International Conference on Natural Language Processing*, number December, pages 175–182.

Shachi Mall and Umesh Chandra Jaiswal. 2018. Survey: Machine Translation for Indian Language. *International Journal of Applied Engineering Research*, 13(1):202–209.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2018. English–Mizo Machine Translation using neural and statistical approaches. *Neural Computing and Applications*, 31(11):7615–7631.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*:223–231.

Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating Text with Recurrent Neural Networks. *Proceedings of the 28th International Conference on Machine Learning*, 131(1):1017–1024.

Krzysztof Wołk and Krzysztof Marasek. 2015. Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts. *International Conference on Project MANagement*, 64:2–9.