NLPRL Odia-English: Indic Language Neural Machine Translation System

Rupjyoti Baruah, Rajesh Kumar Mundotiya
Department of Computer Science & Engineering
Indian Institute Technology (BHU)
Varanasi, India
{rupjyotibaruah.rs.cse18, rajeshkm.rs.cse16}@iitbhu.ac.in

Abstract

In this paper, we (team name is NLPRL) describe our systems that were submitted to the translation shared tasks at WAT 2020. We submitted two systems Odia to English and English to Odia translation under Indic tasks. We presented the Neural Machine Translation (NMT) system based on the Transformer approach using a byte-level version of BPE, which requires a base vocabulary of size 256 only. Experiments show the BLEU score of English to Odia 1.34 and Odia to English 11.33 on the benchmark test data.

1 Introduction

In this paper, we (team name is NLPRL) describe the system that we develop as part of our participation in the Workshop on Asian Translation (WAT) 2020 (Nakazawa et al., 2019, 2020) for the Odia-English language pairs. For the first time, Odia-English language pair is adopted for a translation task in the WAT 2020. The shared task organizers provide Odia-English (OdiEnCorp 2.0) (Parida et al., 2020) parallel corpus that contains train, valid, and test data collected by researchers at UFAL (Institute of Formal and Applied Linguistics). The Odia (old name is Oriya) language is one of the official languages of Odisha, situated in the eastern part of India. It belongs to the Eastern Indo-Aryan group, which is a branch of Indu-European languages. It comes under one of the 22 official languages recognized by the Government of India.

In this experiment, we use byte-level subwords, specifically byte-level BPE (BBPE) (Wang et al., 2020) tokenizer, which is more efficient than pure bytes, worked on the Transformer architecture. The primary motivation was to try out this model on Odia-English parallel data provided by the organizers. We presented the translation results by using an automatic evaluation server prepared by shared task technical collaborators.

The article is structured as follows. In Section 2, we discuss our system description that covering the statistics of the dataset and its preparation with the experimental setup. Section 3 describes results and analysis. Finally, We conclude in Section 4 with the conclusion and future work.

2 Related Study

ANUVADAKSH[1], an expert in English to Indian Languages Machine Translation System (EILMT), allows translation of the text from English to eight Indian languages. The EILMT tool is a hybrid system that has been trained with the data in the three different domains of health, tourism, and agriculture. The EILMT can translate English to Odia language in all three domains [2]. The machine translation system, OMTrans [3] as English to Oriya by Utkal University, Bhuvaneshwar using school book sentences (product information only). The OMTrans system translates text from English to Oriya based on grammar and semantics of the language.

There are many challenges for Neural Machine Translation (NMT) than other approaches mentioned in different literature (Koehn and Knowles, 2017). Most state-of-the-art NMT systems achieve outstanding results based on large parallel corpora only. It has been outperforming the traditional methods for machine translation, such as rule-based and statistical-based approaches. However, for the low resource languages, various techniques have been proposed for NMT, such as meta-learning (Gu et al., 2018), transfer learning (Zoph et al., 2016), which produced promising results. The two main strategies, namely pivot-based and zero-shot machine translation, were used to build NMT models without direct parallel data.

[1] https://cdac.in/index.aspx?id=mc_mat_anuvadaksha
[2] http://eilmt.rb-aai.in/
[3] http://www.ilts-utkal.org/omt.htm

Proceedings of the 7th Workshop on Asian Translation, pages 118–121

The performance of NMT on the low resource languages mainly suffers due to the Out-Of-Vocabulary (OOV), the words not in the trained NMT model's vocabulary.. NMT models typically operate with a fixed vocabulary; however, the translation is an open-vocabulary problem. Several approaches have been proposed to reduce the issue of OOV. Byte-Pair-Encoding (BPE) (Sennrich et al., 2016) enables NMT model translation on open-vocabulary by encoding rare and unknown words as a sequence of subword units. Wang et al. (2020) proposed **B**BPE, which builds a **B**yte-level subword vocabulary for machine translation.

3 System Description

This section covers a dataset, preprocessing, and the experimental setup required for our systems.

3.1 Dataset

To the best of our knowledge, this is the first time the language pair (Odia-English) is running as a shared task in any machine translation competition. OdiEnCorp 2.0 parallel corpus covers many domains, namely the Bible, different literature, Government policies, general conversation, and many topics in Wiki data. The training data includes 69370 lines of parallel corpora. The validation and test data contain 13544 and 14344 lines of parallel sentences, respectively, for both directional pairs (Odia-English).

3.2 Preprocessing

The standard fairseq preprocess[4] script has been exploited for performing the tokenization that uses MOSES tokenizer. The lowercasing operation has performed over the input sentences before performing the tokenization. These tokens are further utilized for performing the BPE encoding because of BBPE encoding composed on BPE. The size of BBPE is 2048.

3.3 Experimental Setup

The Transformer translation model proposed by (Vaswani et al., 2017), which relies on self-attention mechanisms by handling long-term dependencies, has achieved state-of-the-art performance in recent NMT tasks. The vocabulary of the source and target language have been shared during the training of the model. The model uses 0.3 as a dropout to regularize the model. The model trained by Adam optimizer (Kingma and Ba, 2014) with 0.001 as the initial learning rate, which further gradually decays after each 4000.

We conduct experiments using fairseq (Ott et al., 2019) library, a sequence modeling toolkit to train our model with the same value for remaining parameters and hyper-parameters as mentioned in the original paper.

[4] https://github.com/pytorch/fairseq

4 Result and Analysis

The organizers have evaluated the submitted predictions on the test set (of OdiEnCorp 2.0 parallel corpus) of both directions pair in terms of BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) and Adequacy-Fluency Metrics (AM-FM) (Banchs et al., 2015). Table 1 and Table 2 depict the BLEU score and RIBES of the NLPRL and Organizer's model, trained on both directions, respectively. The obtained scores by baseline model and our model have referred to as Organizer and NLPRL in the following tables. From these tables, we have inferred that our model performs better for Odia→English, compared to the baseline. The Figure 1 also indicates that the NLPRL model obtained adequate and fluent results for Odia→English, while vice-versa, it is not true.

Pair	Organizer	NLPRL
Odia→English	8.93	11.33
English→Odia	5.49	1.34

Table 1: BLEU score of Odia-English

Pair	Organizer	NLPRL
Odia→English	.349459	.462557
English→Odia	.326116	.288520

Table 2: RIBES score of Odia-English

5 Conclusion and Future Work

In this paper, we report our submitted system scores based on the official scores released by the WAT 2020 shared translation task. We train our system for Odia-to-English and English-to-Odia language pairs using the Transformer-based neural machine translation using a byte-level version of BPE. WAT 2020 depicts three types of scores, namely BLEU, RIBES, and AM-FM. The BLEU score, a standard

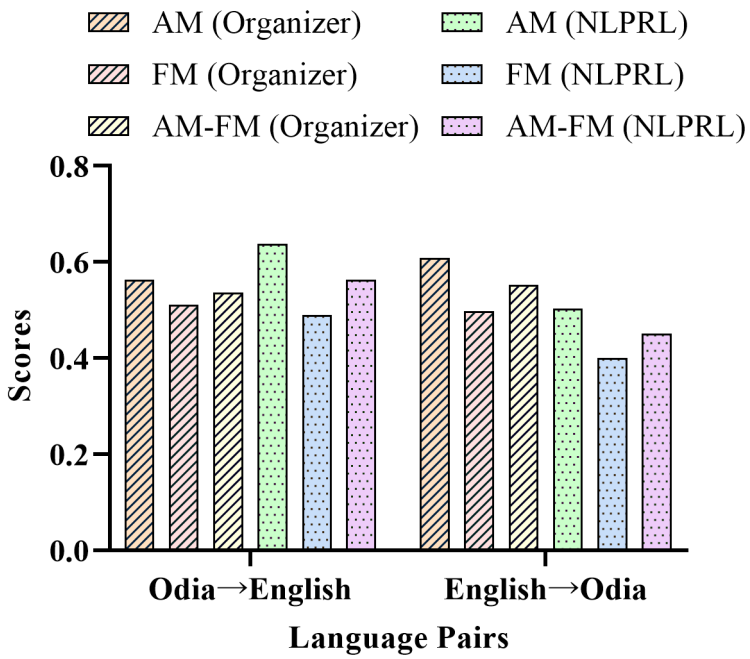

Figure 1: Comparison of average AM, FM and AM-FM scores of the Organizer and NLPRL's model

metric for machine translation evaluation of English to Odia and Odia to English on the benchmark test data, is 1.34 and 11.33, respectively. We found that our model performs better for Odia→English, compared to the Organizer's model.

As the next step, we would like to improve the evaluation score using a multi-lingual NMT model. Furthermore, using monolingual source and target language can improve the translation score of both directions of the language pair.

Acknowledgments

The support and the resources provided by PARAM Shivay Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Technology, Varanasi are gratefully acknowledged.

References

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-Fluency Metrics: Evaluating MT in the Continuous Space Model Framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):472–482.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2019. Overview of the 6th Workshop on Asian Translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik, and Debasish Kumar Mallick. 2020. OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural Machine Translation with Byte-Level Subwords. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9154–9160. AAAI Press.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.