The ADAPT Centre's Participation in WAT 2020 English-to-Odia Translation Task

Prashanth Nayak, Rejwanul Haque and Andy Way
The ADAPT Centre, School of Computing
Dublin City University, Dublin, Ireland
firstname.lastname@adaptcentre.ie

Abstract

This paper describes the ADAPT Centre submissions to WAT 2020 for the English-to-Odia translation task. We present the approaches that we followed to try to build competitive machine translation (MT) systems for English-to-Odia. Our approaches include monolingual data selection for creating synthetic data and identifying optimal sets of hyperparameters for Transformer in a low-resource scenario. Our best MT system produces 4.96 BLEU points on the evaluation test set in the English-to-Odia translation task.

1 Introduction

The ADAPT Centre participated in the English-to-Odia shared task at the 7th Workshop on Asian translation (WAT 2020) (Nakazawa et al., 2020).[1] This paper presents the approaches we adopted in order to try to build competitive MT systems for this translation task. We also discuss methods that did not work for us. Our NMT systems are state-of-the-art Transformer models (Vaswani et al., 2017).

This paper is organized as follows. Section 2 presents our approaches. We describe the resources we utilized for training in Section 3. Section 4 presents the results obtained, and Section 5 concludes our work with avenues for future work.

2 Our Approaches

2.1 Data Augmentation

Neural MT (NMT) (Vaswani et al., 2017) has made considerable progress in recent years, outperforming the previous state-of-the-art statistical MT in many translation tasks, particularly when there are large volumes of parallel corpora available. Building NMT systems for under-resourced languages still poses a challenge despite recent successes (Nakazawa et al., 2019).

As for the task in which we are participating (English-to-Odia), the parallel data that the task organisers provided is relatively small. The organisers also provided us with monolingual data. We made use of monolingual data in training in order to improve our baseline models. The use of synthetic data to improve NMT systems is a well-accepted and popular method, especially in low-resource scenarios (Sennrich et al., 2016a). We did not blindly use all sentences of the monolingual data; instead, we select those sentences that are similar in terms of style and domain to the sentences of the parallel data. In order to select the sentences which are similar to those of the parallel data, we use perplexity scores of the monolingual sentences according to the in-domain language model (Axelrod et al., 2011; Toral, 2013; Haque et al., 2020; Nayak et al., 2020; Parthasarathy et al., 2020). The selected monolingual sentences are then back-translated to form synthetic training data.

2.2 Hyperparameters Search

We conducted a series of experiments to find the best hyperparameters for Transformer as far as low-resource translation is concerned. For our experiments we primarily used those hyperparameters that are commonly used for low-resource scenarios (Sennrich and Zhang, 2019). Additionally, we varied a handful of parameters to see how the MT systems would perform, e.g. encoder and decoder layer sizes. We applied Byte-Pair Encoding (BPE) word segmentation (Sennrich et al., 2016b) both individually and jointly to the source and target language corpora. Since BPE when applied individually worked better for us, we stick to this setup for our system building. We found that the following hyperparameters provided us with the best result in this low-resource scenario: (i) the

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html

Proceedings of the 7th Workshop on Asian Translation, pages 114–117
December 4, 2020. ©2020 Association for Computational Linguistics

	Sentences	Words	
		English	Odia
Training	69,370	1,340,137	1,164,441
Development	13,544	158,166	140,554
Test	14,344	186,164	164,670

Table 1: Statistics of the training, development and test sets.

Monolingual-Corpus	Sentences	Words
OpusNlp	30k	1,003,211
OSCAR	284K	14,938,567
AI4Bharat-IndicNLP	3.5M	53,694,876

Table 2: Statistics of the monolingual corpora.

number of BPE merge operations: 32,000 (ii) the sizes of the encoder and decoder layers: 4 and 6, respectively, and (iii) the learning-rate: 0.02.

3 Data Used

We made use of both the parallel and monolingual data that were provided by the WAT 2020 task organisers.[2] Additionally, we used external monolingual data for system building. The statistics of the parallel and monolingual corpora (OpusNlp,[3] OSCAR[4] and AI4Bharat-IndicNLP)[5] are shown in Tables 1 and 2, respectively. In order to remove noisy sentences from the corpus, we used a language identifier CLD2[6] with a confidence of 95.

4 Results and Discussion

We used the state-of-the-art Transformer model in order to prepare our MT systems. For system building, we used the OpenNMT toolkit (Klein et al., 2017). In order to evaluate our MT systems, we used the widely-used evaluation metric, BLEU (Papineni et al., 2002).

4.1 The Baseline MT System

We made use of the parallel corpus in order to build our baseline NMT system. The original parallel data includes many duplicate entries. There were also many overlapping entries in the training, development and test sets. The duplicate entries from the training set were removed accordingly. Then we built an MT system on deduplicated training data. From now on, we call this MT system Base. We obtained the BLEU score to evaluate Base on the test set and report the score in Table 4. Note that we built all our MT systems following the best hyperparameters setup described in Section 2.2.

	BLEU
Base	6.11
Base + 1M	5.03

Table 3: The BLEU scores of the Odia-to-English MT systems.

	BLEU
Base	4.96
Base + 1M	3.53

Table 4: The BLEU scores of the English-to-Odia MT systems.

4.2 Using Monolingual Data

As mentioned above, since the parallel corpus is small in size, we made use of monolingual data to improve Base following the method presented in Section 2.1. For this, we built an Odia-to-English MT system and used it to translate our Odia monolingual sentences. The BLEU score of the Odia-to-English MT system (cf. Base) is shown in Table 3.

The quality of synthetic parallel data is crucial for training or fine-tuning an NMT system. As can be seen from Table 3, since our Odia-to-English baseline MT system (i.e. Base) is also not good in quality, we tried to improve it so that we can have a better quality synthetic parallel corpus. Therefore, in addition to the parallel corpus, we used a synthetic corpus of one million sentence-pairs for training. However, we can see from Table 3 that using synthetic data causes to deteriorate the Odia-to-English MT system's performance. As a result, we used our best Odia-to-English MT system, Base, for translating the Odia monolingual sentences.

The score of the English-to-Odia MT system built on training data composed of the authentic and synthetic parallel data is shown in Table 4. We see that adding synthetic data (one million sentence-

[2] https://github.com/shantipriyap/Odia-NLP-Resource-Catalog

[3] https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses

[4] https://oscar-corpus.com/

[5] https://github.com/ai4bharat-indicnlp/indicnlp_corpus

[6] https://github.com/CLD2Owners/cld2

pairs) to the original parallel data does not help in this case either.

5 Conclusions

This paper presents the ADAPT Centre system description for the WAT 2020 English-to-Odia translation shared task. Our best MT model, a Transformer model prepared using an optimal set of hyperparameters, obtain 4.96 BLEU points on the evaluation test set. We selected those monolingual sentences from a large monolingual data that are similar in terms of style and domain to the sentences of the parallel corpus. We then created a synthetic parallel corpus by translating the selected Odia monolingual sentences to English. We fine-tuned our baseline MT system on the training data that combines of the synthetic and original parallel corpora. This strategy did not work for us since using synthetic data causes to deteriorate the performance of the English-to-Odia MT system.

As for future work, we aim to explore transfer learning and using data of other related languages in order to improve translation of the English-to-Odia MT system.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077 and 18/CRT/6224 .

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. The ADAPT system description for the STAPLE 2020 English-to-Portuguese translation task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 144–152, Online. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

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.

Prashanth Nayak, Rejwanul Haque, and Andy Way. 2020. The adapt's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (Biomedical)*, Punta Cana, Dominican Republic.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Venkatesh Balavadhani Parthasarathy, Akshai Ramesh, Rejwanul Haque, and Andy Way. 2020. The ADAPT system description for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (News))*, Punta Cana, Dominican Republic.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. In *Proceedings of the second workshop on hybrid approaches to translation*, pages 8–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.