Adobe AMPS's Submission for Very Low Resource Supervised Translation Task at WMT20

Keshaw Singh
AI/ML Platform & Solutions, Adobe Inc.
Bengaluru, India
kessingh@adobe.com

Abstract

In this paper, we describe our systems submitted to the very low resource supervised translation task at WMT20. We participate in both translation directions for Upper Sorbian-German language pair. Our primary submission is a subword-level Transformer-based neural machine translation model trained on original training bitext. We also conduct several experiments with backtranslation using limited monolingual data in our post-submission work and include our results for the same. In one such experiment, we observe jumps of up to 2.6 BLEU points over the primary system by pretraining on a synthetic, backtranslated corpus followed by fine-tuning on the original parallel training data.

1 Introduction

This paper describes our submissions to the shared task on Very Low Resource Supervised Machine Translation at WMT 2020. The task involved a single language pair: Upper Sorbian-German. We submit supervised neural machine translation (NMT) systems for both translation directions, Upper Sorbian→German and German→Upper Sorbian.

NMT models (Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014a) have achieved state-of-the-art performance on benchmark datasets for multiple language pairs. A big advantage of such systems over phrase-based statistical machine translation (PBSMT) (Koehn et al., 2003) models is that they can be trained end-to-end. The bulk of the development, however, has been limited to a handful of high-resource language pairs. The primary reason is that training a well-performing NMT system requires a large amount of parallel training data, which means a lot of equivalent investment in terms of resources. Koehn and Knowles (2017) show that when compared to PBSMT approaches, NMT models need more training data to achieve the same level of performance.[1] One of the most popular ways to increase the amount of parallel training data for supervised training is backtranslation (Sennrich et al., 2016a). We utilize this approach to improve upon the performance of our baseline models.

All of our systems follow the Transformer architecture (Vaswani et al., 2017). Our primary system is a supervised NMT model trained on the original training bitext. We also report our results on experiments with backtranslation, which were completed post the shared task and hence not a part of our primary submissions. We use the backtranslated data in two distinct ways - as a standalone parallel corpus, and to create a combined parallel corpus by mixing in a 1:1 ratio with the provided training data. We also report the performance of fine-tuned models originally trained only on the backtranslated data. In the following sections, we begin by briefly describing the Transformer architecture and backtranslation. We then discuss our experimental setup as well as our experiments with backtranslation. We conclude with a discussion of our results and possible future work.

2 Related Work

The Transformer model is the dominant architecture within current NMT models due to its superior performance on several language pairs. While still a sequence-to-sequence (Sutskever et al., 2014) model composed of an encoder and a decoder, Transformer models are highly parallelizable thanks to being composed purely of feed-forward and self-attention layers rather than recurrent layers (Hochreiter and Schmidhuber, 1997; Cho et al., 2014b). The reader is encouraged to read the original paper (Vaswani et al., 2017) to gain a deeper understanding of the model. We adopt the Transformer base architecture available under the

[1] As measured by BLEU score (Papineni et al., 2002).

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1144–1149

fairseq[2] (Ott et al., 2019) library for all our models.

However, NMT models are known to be data-hungry (Koehn and Knowles, 2017); their performance improves sharply with the availability of more parallel training data. Except for a few language pairs (e.g. English-German), most have little to no such data available. On the other hand, a far greater number of languages have a decent amount of monolingual data available online (e.g. Wikipedia).

To address this issue of lack of parallel data, Sennrich et al. (2016a) introduced the concept of back-translation. It involves creating a synthetic parallel corpus by translating sentences from the target-side monolingual data to the source language and making corresponding pairs. A baseline target→source model (PBSMT or NMT), trained with limited data, is generally used for this purpose. It enables the use of large corpora of monolingual data for several languages, the size of which is typically orders of magnitude larger than any corresponding bitext available. What is notable is that only the source-side data is synthetic in such a scenario and the target-side still corresponds to original monolingual data.

Some studies (Poncelas et al., 2018; Popel, 2018) have investigated the effects of varying the amount of backtranslated data as a proportion of the total training corpus, including training only on the synthetic dataset as a standalone corpus. We follow some of the related experiments conducted by Kocmi and Bojar (2019) on Gujarati-English (another low-resource pair) with a few exceptions. Besides, we also report performance when pretraining solely on the synthetic corpus following by fine-tuning on either original or mixed data. While not quite the same, one could think of this approach as having some similarities with transfer learning (Zoph et al., 2016) as well as domain adaptation (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016) for machine translation. There has also been work on using sampling (Edunov et al., 2018) for generating backtranslations, but we stick to using beam search in this work.

3 Experimental Setup

3.1 Dataset

We used the complete parallel training corpus for our primary systems. In addition, we also made use of monolingual data from each language for two purposes - learning Byte Pair Encodings (BPE) (Sennrich et al., 2016b) and backtranslation. For Upper Sorbian (hsb), we used the monolingual corpora provided by the Sorbian Institute and by the Witaj Sprachzentrum. To control the quality of the backtranslated data, we chose not to use the data scraped from the web. For the German (de) side, we made use of the News Crawl[3] 2009 dataset, as it is large enough to satisfy the requirements for our experiments.

3.2 Data Preprocessing

Source	No. of sentences
hsb-de, bitext	58,389
hsb, monolingual	540,994
de, monolingual	2,000,000

Table 1: Processed training data.

Moses toolkit (Koehn et al., 2007) was used for tokenization and punctuation normalization for all data. Before doing any additional preprocessing, we learned separate truecaser models using the toolkit. For this purpose, we took first 500K sentences from each of the monolingual corpora and aggregated them with the corresponding portion from the training bitext. After tokenizing and truecasing, we joined the parallel training corpus with the same monolingual data. We learned joint BPE[4] with 32K merge operations over this corpus and applied them to the parallel training data to get vocabularies for each language. Additionally, we used the `clean-corpus-n.perl` script within Moses to filter out sentences from the parallel corpus with more than 250 subwords as well as sentence length ratio over 1.5 in either direction. Final corpus statistics are presented in Table 1.

3.3 Training

Our primary system is a Transformer base model, trained on the parallel training corpus for both translation directions till 60 epochs. We keep most of the hyperparameters to their default values in fairseq. More precisely, we chose Adam (Kingma and Ba, 2015) as the optimizer and Adam betas were set to 0.9 and 0.98, respectively. The maximum number of tokens in each batch was set to 4096. Learning rate was set to 0.0005, with an inverse squared

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

[3] http://data.statmt.org/news-crawl/de/

[4] https://github.com/glample/fastBPE

root decay schedule and 4000 steps of warmup updates. Label smoothing was set to 0.1 and dropout to 0.3. Label-smoothed cross-entropy was used as the training criterion.

We trained all our models for a fixed number of epochs, determined separately for each system, and chose the last checkpoint for reporting BLEU (Papineni et al., 2002) scores on the test sets.

All training was done using a single NVIDIA P100 GPU. Due to the small amount of parallel training data, each epoch of training took about 90 seconds on average for the primary system.

4 Additional Backtranslation Experiments

In this section, we report our post-submission work on using monolingual data for backtranslation. We took the raw monolingual data that we describe in Section 3.1 and backtranslated with our primary submission models for the respective translation directions, i.e., hsb→de for Upper Sorbian data and de→hsb for German data. We used `fairseq-generate` function with a beam size of 5 for this purpose. Once again, we limited the number of subwords in each sentence to 250. Finally, we took all sentence pairs for backtranslated Upper Sorbian corpus and the first two million sentence pairs for the German corpus. Table 1 indicates the size of the backtranslated corpora by original language. For further experiments, we name the datasets as follows:

- *auth*: Processed original training data.
- *synth*: Backtranslated de→hsb and hsb→de corpora.
- *mixed*: Augmented training data obtained by mixing *auth* with a portion of *synth* in 1:1 ratio, providing a total of 116,778 sentence pairs.

We define the following systems for making use of the backtranslated data. Note that the first system only differs from the primary system in the number of training epochs completed.

- *auth-from-scratch*: This system has the same settings as the primary system. It was trained on the *auth* corpus till 80 epochs (as opposed to 60 for primary).
- *mixed-from-scratch*: We trained models on *mixed* data from scratch for 40 epochs.[5]
- *synth-from-scratch*: Models were trained only on the *synth* datasets. To adjust for the difference in the size of the respective backtranslated corpora, we trained hsb→de system for 10 epochs and de→hsb system for 30 epochs.
- *synth-auth-finetune*: We took the models trained via the previous system and fine-tuned them on *auth* data for 20 epochs in each translation direction.
- *synth-mixed-finetune*: Same as the last model, except that fine-tuning was done on *mixed* data.

Fine-tuning was carried out by loading pretrained checkpoints and adding extra training flags in `reset-optimizer` and `reset-lr-scheduler`.

5 Results

The systems were evaluated on the blind test set (newstest2020) using automated metrics; no human evaluation was done. Table 2 shows cased BLEU scores for various systems. Our primary systems achieved a BLEU score of 47.6 for Upper Sorbian→German and 45.2 for German→Upper Sorbian translation. We achieved an improvement of 0.3 and 0.4 BLEU points, respectively, by training further till 80 epochs in each direction. We also evaluated a third system, *synth-auth-finetune*, as described in Section 4, which provided a jump of 2.6 points in BLEU score over the primary system for Upper Sorbian→German and 2.5 for German→Upper Sorbian.

In addition to evaluating on blind test sets, we also report BLEU scores on the development test set in the same table. Two outcomes are worth highlighting:

- Model trained only on *synth* data for German→Upper Sorbian translation matched the performance of a similar model trained on the authentic bitext.
- Best results were obtained by fine-tuning a model trained on *synth* data with either *auth* or *mixed*.

[5]We trained further till 60 epochs, but observed no improvement in BLEU scores.

System	Dataset	Epochs	newstest2020	devtest
			hsb→de	
Primary*	*auth*	60	47.6	-
auth-from-scratch	*auth*	80	47.9	45.6
mixed-from-scratch	*mixed*	40	-	45.7
synth-from-scratch	*synth*	10	-	38.0
synth-auth-finetune	*+auth*	20	**50.2**	**49.6**
synth-mixed-finetune	*+mixed*	20	-	48.3
			de→hsb	
Primary	*auth*	60	45.2	-
auth-from-scratch	*auth*	80	45.6	46.4
mixed-from-scratch	*mixed*	40	-	47.4
synth-from-scratch	*synth*	30	-	46.5
synth-auth-finetune	*+auth*	20	**47.7**	49.0
synth-mixed-finetune	*+mixed*	20	-	**49.6**

Table 2: BLEU scores for the blind test set (newstest2020) and the development test set. Bold values in a column indicate the best scores among the evaluated systems. + Additional fine-tuning for models trained with backtranslated corpora. * Only the primary systems were evaluated before deadline.

The second result is notable since the regime of pretraining followed by fine-tuning improves the BLEU scores by up to 4 points on this test set when compared to training only on the original bitext. Moreover, while the model trained on *synth* was not able to match the performance of that trained on *auth* for Upper Sorbian→German, it still provides the same benefits as German→Upper Sorbian model when fine-tuned further. Looking at the small improvements achieved by using only the *mixed* corpus for training, increasing its size by combining upsampled *auth* data with more *synth* data might lead to even further jumps in the BLEU scores.

6 Conclusion

In this paper, we described our Transformer model for supervised machine translation for Upper Sorbian-German language pair. We take note of relatively high BLEU scores achieved by our primary systems (and those of other participants) on this low-resource language pair, which could relate to the high quality of the training corpus. We also report results and takeaways from several experiments with backtranslated data completed post the shared task. A key result is matching the performance of a system trained on the original bitext with one trained on a limited amount of synthetic, backtranslated data. Domain mismatch and a difference in the quality of monolingual corpus might have prevented the system from achieving a similar result in the other direction. We notice big improvements in performance over the primary systems by following a "pretraining then fine-tuning" regime.

An interesting future work would be to measure the applicability of this approach to other low-resource language pairs. Additional systems could be added as well. For instance, models trained on *mixed* data and fine-tuned on *auth* data might provide a meaningful comparison. Prior work (Ding et al., 2019) has shown that the number of BPE merge operations has a significant effect on the performance of NMT systems. This work was pointed out during the review process and should be an avenue for further improvement of the model performance.

Acknowledgments

The author would like to thank his manager for supporting this project, and the anonymous reviewers for their thoughtful comments which helped improve the presentation of this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, California, USA.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder

approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, California, USA.

Tom Kocmi and Ondřej Bojar. 2019. CUNI submission for low-resource languages in WMT news 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 234–240, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

A Poncelas, D Shterionov, A Way, GM de Buy Wenniger, and P Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Martin Popel. 2018. Machine translation using syntactic analysis. *Univerzita Karlova*.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

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, Austin, Texas. Association for Computational Linguistics.