NICT's Supervised Neural Machine Translation Systems for the WMT19 Translation Robustness Task

Raj Dabre and **Eiichiro Sumita**
National Institute of Information and Communications Technology, Kyoto, Japan
{raj.dabre,eiichiro.sumita}@nict.go.jp

Abstract

In this paper we describe our neural machine translation (NMT) systems for Japanese↔English translation which we submitted to the translation robustness task. We focused on leveraging transfer learning via fine tuning to improve translation quality. We used a fairly well established domain adaptation technique called Mixed Fine Tuning (MFT) (Chu et al., 2017) to improve translation quality for Japanese↔English. We also trained bi-directional NMT models instead of uni-directional ones as the former are known to be quite robust, especially in low-resource scenarios. However, given the noisy nature of the in-domain training data, the improvements we obtained are rather modest.

1 Introduction

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) has enabled end-to-end training of a translation system without needing to deal with word alignments, translation rules, and complicated decoding algorithms, which are the characteristics of phrase-based statistical machine translation (PB-SMT) (Koehn et al., 2007). NMT performs well in resource-rich scenarios but badly in resource-poor ones (Zoph et al., 2016).

One such resource-poor scenario is the translation of noisy sentences which are often found on social media like Reddit, Facebook, Twitter etc. There are two main problems: (a) The type of noise (spelling mistakes, code switching, random characters, emojis) in the text is unpredictable (b) Scarcity of training data to capture all noise phenomena. One of the first works on dealing with noisy translation led to the development of the MTNT (Michel and Neubig, 2018) test suite for testing MT models that are robust to noisy text. Fortunately, the problem of noisy text translation can be treated as a domain adaptation problem and there is an abundant amount of Japanese–English text that be leveraged for this purpose. In this paper, we describe the systems for Japanese↔English translation, that we developed and submitted for WMT 2019 under the team name "NICT". In particular our observations can be summarized as follows:

Japanese↔English translation dramatically fails given the limited amount of noisy training data.

Fine-Tuning is simple but has over-fitting risks.

Mixed-Fine-Tuning is a simple but effective way of performing domain adaptation via fine tuning where one does not have to worry about the possibility of quick over-fitting.

Kindly refer to the task overview paper (Li et al., 2019) for additional details about the task, an analysis of the results and comparisons of all submitted systems which we do not include in this paper.

2 Approaches

We used domain adaptation approaches on top of the transformer model.

2.1 The Transformer NMT Model

The Transformer (Vaswani et al., 2017) is the current state-of-the-art model for NMT. It is a sequence-to-sequence neural model that consists of two components: the *encoder* and the *decoder*. The encoder converts the input word sequence into a sequence of vectors. The decoder, on the other hand, produces the target word sequence by predicting the words using a combination of the previously predicted word and relevant parts of the

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 533–536
Florence, Italy, August 1-2, 2019.

input sequence representations. The reader is encouraged to read the original paper (Vaswani et al., 2017) for a deeper understanding.

2.2 Mixed Fine Tuning for Domain Adaptation

The fastest way to adapt an out-of-domain model to an in-domain task is to first train a L1→L2 model on the large out-of-domain data and then fine tune it on the small in-domain data. However, given that NMT models overfit quickly on small data (Zoph et al., 2016), it is important to consider learning rate modification, regularization and sophisticated training schedules. All this can be avoided by performing Mixed-Fine-Tuning (MFT) (Chu et al., 2017) where the out-of-domain model is fine-tuned on a combination of both the out-of-domain data and the oversampled[1] in-domain data. When using this technique there is no risk of overfitting.

2.3 Bi-directional NMT Modeling

Multilingual models (Johnson et al., 2017) enable a model to learn multiple translation directions without increasing the model size. We concatenated the Japanese→English and English→Japanese training corpora after appending the tokens "2en" and "2ja" to the source sentences of the respective corpora. In addition to this, we did not modify the NMT model in any way.

3 Experimental Settings and Results

3.1 Datasets

We used the official Japanese→English and English→Japanese datasets provided by WMT. The out-of-domain (non noisy) datasets are KFTT, JESC and TED Talks, all of which are adequately described in the original MTNT paper (Michel and Neubig, 2018). The total number of out-of-domain sentence pairs is 3,900,772. As for the in-domain corpus, the number of training sentence pairs for Japanese→English translation is 6,506 pairs and for English→Japanese translation there are 5,775 pairs. Upon inspection of the English→Japanese data, we noted that many sentences were actually paragraphs which are almost useless for NMT training as they are trimmed to avoid out-of-memory errors. We tried a naive paragraph splitting method where we split a paragraphs into sentences and keep the splits if there are an equal number of sentences. Upon manual investigation we found out that this splitting leads to correct splits most of the times. As a result, the number of training sentences for English→Japanese translation increases to 10,060 pairs. We pre-processed the Japanese text using KyTea (Neubig et al., 2011). Other than this, we do not perform any pre-processing.

3.2 Model Training Details

We used the tensor2tensor[2] version 1.6 implementation of the Transformer (Vaswani et al., 2017) model. We used the default hyperparameters in tensor2tensor for all our models with the exception of the number of training iterations. Unless mentioned otherwise we use the "base" transformer model hyperparameter settings with a $2^{15} = 32,768$ shared sub-word vocabulary which is learned using tensor2tensor's internal tokenization and sub-word segmentation mechanism. We used a shared sub-word vocabulary because we trained bi-directional models. This allows us to share embeddings between the encoder and the decoder. During training, a model checkpoint is saved every 1000 iterations. All models were trained till convergence on the development set BLEU score. We averaged the last 10 model checkpoints and used it for decoding the test sets. We chose a default beam size of 10 and length penalty of 0.8. We did not ensemble multiple models although it could possibly improve the translation quality even further. When we fine-tuned models, we simply resumed training the last model checkpoint on the noisy in-domain data. We did not change the optimizer nor any other hyperparameters. One might argue that this could lead to overfitting but tensor2tensor uses a learning rate decay by default which prevents this. Furthermore, MFT does not suffer from overfitting.

3.3 Systems

We first trained a (bidirectional) Japanese↔English model using the out-of-domain parallel corpus for 150,000 iterations on 1 GPU with a batch size of 2048 words. We did not train for a larger number of iterations

[1] To balance the highly skewed corpora ratio thereby ensuring that the model sees an equal number of training examples from both domains.

[2] https://github.com/tensorflow/tensor2tensor

Task	BLEU	BLEU cased	IGNORE BLEU (11b)	IGNORE BLEU-cased (11b)	IGNORE BLEU-cased-norm	BEER 2.0
English→Japanese	11.1	11.1	11.1	11.1	11.1	0.354
Japanese→English	8.1	7.4	8.1	7.4	7.8	0.352

Table 1: Results for Japanese↔English translation for the robustness task.

Approach	Ja→En	En→Ja
Bidirectional FT	9.6	10.5
Bidirectional MFT	9.2	**13.4**

Table 2: BLEU scores on the non-blind test set for Japanese–English translation. We show that MFT is either comparable to or significantly better than regular fine-tuning.

because the model had converged sufficiently by 150,000 iterations. We then used this model to perform Mixed-Fine-Tuning (MFT) which uses a combination of the out-of-domain and in-domain corpus. MFT is done for 50,000 iterations on 1 GPU with a batch size of 2048 words.

3.4 Results

Refer to Table 1 for the various automatic evaluation scores. For English→Japanese our submitted system's run achieved a cased BLEU score of 11.1. On the other hand, our Japanese→English system's run achieved a BLEU score of 8.1.

A surface level analysis of our translations showed that the implementation of the Transformer that we used is not well suited to handle noisy text. In most cases it does not handle emojis. We noted that emojis are always missing in the translation. Another problem we observed was that the default KyTea model does not give good morphological segmentations which we believe is one of the reasons for our poor performance in the task. In the future, we will incorporate better pre-processing mechanisms into the tensor2tensor implementation for better translation. Although, we did not mention it in the paper, we tried to use back-translation to translate the monolingual data in the MTNT dataset but were unable to achieve satisfactory results.

3.5 Comparison of Approaches

In Table 2 we give the BLEU scores of our bidirectional models using fine-tuning and mixed-fine tuning. We obtained these BLEU scores on the non-blind test set which was provided along with the training data. We did not use this test set for training or tuning. The BLEU scores are obtained using SacreBLEU (Post, 2018). We can see that while the performance of Japanese to English slightly degrades (not statistically significant), English to Japanese translation improves by approximately 2 BLEU points. As such MFT is either comparable to or significantly better than regular fine-tuning and was the reason why we chose it for the final submission.

4 Conclusion

In this paper we have described our primary Japanese↔English systems whose translations we have submitted to the robustness translation task in WMT2019. In general, we found that bi-directional modeling and Mixed-Fine-Tuning (MFT) work reasonably well for this task although MFT is the main reason behind the improvements. However, these techniques only partially address the problem of training NMT models that are robust to noise. MFT is a robust training approach and does not actually deal with different sources of noise. In the future we will consider applying better pre-processing mechanisms, domain adaptation techniques and data augmentation techniques for even more robust translation systems.

Acknowledgements

We thank the organizers for providing the datasets and the reviewers for their valuable suggestions in improving this paper. This work was conducted under the program "Research and Development of Enhanced Multilingual and Multipurpose Speech Translation Systems" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

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

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

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112, Montréal, Canada.

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 *Proceedings of 30th Advances in Neural Information Processing Systems*, pages 5998–6008.

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, USA.