DFKI-MLT System Description for the WMT18 Automatic Post-editing Task

Daria Pylypenko
DFKI Saarbrücken, Germany
daria.pylypenko@dfki.de

Raphael Rubino
DFKI Saarbrücken, Germany
raphael.rubino@dfki.de

Abstract

This paper presents the Automatic Post-editing (APE) systems submitted by the DFKI-MLT group to the WMT'18 APE shared task. Three monolingual neural sequence-to-sequence APE systems were trained using target-language data only: one using an attentional recurrent neural network architecture and two using the attention-only (*transformer*) architecture. The training data was composed of machine translated (MT) output used as source to the APE model aligned with their manually post-edited version or reference translation as target. We made use of the provided training sets only and trained APE models applicable to phrase-based and neural MT outputs. Results show better performances reached by the attention-only model over the recurrent one, significant improvement over the baseline when post-editing phrase-based MT output but degradation when applied to neural MT output.

1 Introduction

For the 2018 edition of the WMT automatic post-editing (APE) task, two novelties were added compared to the previous editions: post-editing of neural machine translation (NMT) output in addition to phrase-based (PBMT) output, and the availability of larger training sets.

The DFKI-MLT systems developed for this shared task aimed at handling outputs from PBMT and NMT jointly with a single APE model. This was achieved by using artificial tokens indicating which type of MT system was used to produce the source segment and from which corpus the segment pair was extracted (inspired by (Yamagishi et al., 2016; Sennrich et al., 2016a; Johnson et al., 2017)).

Two NMT architectures were used to train our APE models, one using gated recurrent layers with global attention (Bahdanau et al., 2014), and one using attention and feed-forward layers without recurrence (Vaswani et al., 2017). The training data was composed of the official training set released by the shared task organizers plus subsets of the two additional resources filtered with bilingual cross-entropy difference (Axelrod et al., 2011).

The NMT architectures are described in Section 2 and the data preparation process is presented in Section 3. The results obtained by our APE models are compared to the baseline in Section 4. Finally, a conclusion is given in Section 5.

2 APE Architectures

The two neural network architectures used in our experiments were an attentional recurrent neural network with gated units and a multi-head attention-only network.

2.1 Recurrent Neural Network

For the Recurrent Neural Network (RNN) approach, we followed the architecture presented in (Bahdanau et al., 2014) and implemented in OpenNMT (Klein et al., 2017)[1]. Both the encoder and the decoder were 2-layered mono-directional RNNs with LSTM cells. The decoder applies global attention over the source sentence and performs input feeding. The source and target word embeddings, as well as the hidden layers, had 500 dimensions. The dropout probability was set to 0.3. The source and target vocabulary size is limited to 50000 tokens. Standard stochastic gradient descent is used as optimizer with a maximum batch size of 64. These hyper-parameters are the default ones in OpenNMT and were not tuned during the experiments presented in this paper.

[1] We used the Torch version of OpenNMT available at https://github.com/OpenNMT/OpenNMT

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 836–839

https://doi.org/10.18653/v1/W18-64096

2.2 Attention Only

For the attention only approach, we used the architecture described in (Vaswani et al., 2017) and implemented in MARIAN (Junczys-Dowmunt et al., 2018). Two models were trained following this approach with variations in the number of heads (parallel attention layers), using 4 heads and 1024 dimensions for the feed-forward layers for one configuration (noted *Transformer small*) and 8 heads and 2048 dimensions for the second configuration (noted *Transformer large*). For both configurations, 512 dimensions were used for the embedding layers and the positional encodings, the dropout rate was set to 0.1 and the batch size to 32. These hyper-parameters were selected in order to compare the impact of increasing the dimensionality of the encoder and decoder layers, as well as the number of heads, on the post-editing performances.

3 Data Preparation

The training corpora provided for the APE shared task since 2016 were used (Bojar et al., 2016, 2017), as well as the two additional resources made available by the shared task organizers, namely the *artificial training data* presented in (Junczys-Dowmunt and Grundkiewicz, 2016) and the *eSCAPE* corpus (Negri et al., 2018). The target language data (German) was used for both input and output sequences in our APE models, the machine translated text being the source sequences and the corresponding post-edited text the target sequences, without making use of the source language (English). We did not split the machine translated data whether it was produced by a phrase-based (PBMT) or a neural (NMT) system. Instead, we added a specific token at the beginning of every source (machine translated) segment indicating which type of translation system was used to produce it.

The two additional parallel resources (*artificial training data* and *eSCAPE* corpora) were filtered using the bilingual cross-entropy difference approach presented in (Axelrod et al., 2011). We used the APE training data as in-domain corpus and each additional parallel corpus individually as out-of-domain corpus. The top n sentence pairs ranked by their bilingual cross-entropy scores were kept, with n being set by calculating the perplexity obtained on the development set. The resulting corpora used contain approx. 100k, 300k and 360k segment pairs taken from the *eSCAPE* PBMT corpus, the *eSCAPE* NMT corpus and the *artificial training data* respectively. Finally, we added a specific token at the beginning of every source segment indicating from which source it comes from: *eSCAPE*, *artificial* and *wmt*. The latter token was added to the official training data provided for the APE task, and to the development and test sets as well.

All datasets were used together to train our APE models, the artificial tokens inspired by (Yamagishi et al., 2016; Sennrich et al., 2016a; Johnson et al., 2017) allowed for identification of the segment pairs provenance. In order to balance the amount of data coming from different sources, we oversampled the official training data to reach approximately the amount taken from the two additional resources. Similarly, we increased the amount of data produced by a NMT system to balance with the amount produced by a PBMT system. This method was inspired by the work presented in (Chu et al., 2017).

The corpora which were not already tokenized were processed with the tokenizer distributed with the MOSES toolkit (Koehn et al., 2007). Additionally, all corpora were true-cased using a pre-trained true-casing model provided by the WMT organizers[2]. Finally, a byte-pair encoding (Sennrich et al., 2016b) model was trained on the German training data available for the WMT translation task and applied to both source and target sides of all corpora used in our experiments.

4 Evaluation

The three APE models trained for the shared task were used to post-edit the test set released by the organizers. Automatic evaluation with BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) was conducted by the organizers and the obtained scores on the official test set are reported in Table 1. The automatic metrics results are obtained by comparing each system output to the manually post-edited MT output (TER$_{pe}$ and BLEU$_{pe}$), to an independent translation (TER$_{ref}$ and BLEU$_{ref}$) and finally using both post-edited MT output and independent translation simultaneously as a multi-reference evaluation approach (TER$_{pe+ref}$ and BLEU$_{pe+ref}$). The results obtained by the non-post-edited MT output is presented as a baseline.

[2] http://data.statmt.org/wmt18/translation-task/preprocessed/de-en/

System	TER$_{pe}$	BLEU$_{pe}$	TER$_{ref}$	BLEU$_{ref}$	TER$_{pe+ref}$	BLEU$_{pe+ref}$
			PBMT Output			
Baseline	24.24	62.99	48.33	36.42	23.76	66.21
Transformer large	24.19	**63.40**	**47.98**	**36.81**	23.68	**66.66**
Transformer small	24.50	62.78	48.27	36.61	24.04	66.11
RNN	25.30	62.10	48.55	36.19	24.74	65.33
			NMT Output			
Baseline	16.84	74.73	42.24	44.22	16.27	76.83
Transformer large	18.86	70.98	43.74	41.53	18.37	72.93
Transformer small	18.84	70.87	43.79	41.53	18.41	72.95
RNN	19.88	69.35	44.28	40.91	19.43	71.36

Table 1: Automatic metrics results on the test set obtained by our APE models and compared to the baseline using three evaluation methods. Result in bold indicates significant improvement over the baseline.

The automatic evaluation results show that our models significantly degrades the baseline for the NMT output experiments when using the manually post-edited MT output, the independent translation and both simultaneously as gold reference to compute the scores. For the PBMT experiments, the model noted *Transformer large* significantly improves the PBMT output according to the BLEU metric for the three evaluation methods ($+0.4$pt for the post-edited MT output, $+.39$pt of the reference and $+.45$pt for both). However, the TER metric does not indicate significant improvements over the baseline when using the manually post-edited MT output as a gold reference.[3]

The degradation of NMT output in terms of automatic metrics might have at least two explanations. First, the lower amount of available training data produced by this type of MT system and provided by the organizers ($17,753$ unique tokens for NMT and $22,578$ for PBMT after truecasing). We used the over-sampling technique to balance the amount of NMT and PBMT data but this method does not increase the vocabulary coverage. Second, the baseline performances as indicated by the BLEU metric, 74.73 and 44.22 for the post-edited MT output and translation reference used as gold target respectively, are higher than the ones obtained with the PBMT experiments, which might be harder to outperform.

5 Conclusion

This paper presented the DFKI-MLT submissions to the WMT'18 APE shared task, which involved datasets produced by NMT and PBMT systems, as well as larger training data provided by the organizers. We evaluated two different APE architectures based on neural networks and made use of data preprocessing techniques to allow single models to be trained while being able to post-edit both NMT and PBMT outputs and using the target language data only.

The results as indicated by the BLEU metric showed that our approach brings significant improvement over the non post-edited PBMT output when using various gold references to compute the evaluation scores, but fails at improving NMT output. This might be due to the lower amount of training data produced by an NMT system compared to the PBMT produced data, and to the high performance reached by the baseline system on the NMT output as indicated by BLEU.

From the two APE architectures evaluated in our experiments and according to the automatic metrics used, the attention-only model outperformed the gated recurrent one for both types of MT output to post-edit. Both NN architectures could possibly reach better post-editing performances with careful hyper-parameters tuning and we plan to conduct these experiments in the future.

Acknowledgments

The research reported in this paper is partially funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IW17001 (Deeplee) and by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding". Responsibility for the content of this publication is with the authors.

[3] Significance tests were performed by the shared task organizers, more details are available in (Chatterjee et al., 2018).

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 Shared Task on Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*.

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 (Volume 2: Short Papers)*, volume 2, pages 385–391.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 751–758.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. escape: a large-scale synthetic corpus for automatic post-editing. *arXiv preprint arXiv:1803.07274*.

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 on association for computational linguistics*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

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)*, volume 1, pages 1715–1725.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

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.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.