APE through neural and statistical MT with augmented data: ADAPT/DCU submission to the WMT 2019 APE Shared task

Dimitar Shterionov **Joachim Wagner** **Félix do Carmo**
ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
{firstname}.{lastname}@adaptcentre.ie

Abstract

Automatic post-editing (APE) can be reduced to a machine translation (MT) task, where the source is the output of a specific MT system and the target is its post-edited variant. However, this approach does not consider context information that can be found in the original source of the MT system. Thus a better approach is to employ multi-source MT, where two input sequences are considered – the original source and the MT output.

Extra context information can be introduced in the form of extra tokens that identify certain global properties of a group of segments, added as a prefix or a suffix to each segment. Successfully applied in domain adaptation of MT as well as on APE, this technique deserves further attention. In this work we investigate multi-source neural APE (or NPE) systems with training data which has been augmented with two types of extra context tokens. We experiment with authentic and synthetic data provided by WMT 2019 and submit our results to the APE shared task. We also experiment with using statistical machine translation (SMT) methods for APE. While our systems score bellow the baseline, we consider this work a step towards understanding the added value of extra context in the case of APE.

1 Introduction

Automatic post-editing (APE) aims at improving text that was previously translated by Machine Translation (MT). An APE system is typically trained on triplets composed of: a segment in the source language, a translation hypothesis of that segment by an MT system, and the edited version of that hypothesis, created by a human translator.

Currently, neural machine translation (NMT) systems are the state-of-the-art in MT, achieving quality beyond that of phrase-based statistical MT (SMT) (Bentivogli et al., 2016; Shterionov et al., 2018). NMT output is more fluent but may contain issues related to accuracy. However, automatic post-editing of NMT output has proved to be a challenging task (Chatterjee et al., 2018).

In terms of post-editing technology, neural methods as well represent the current state-of-the-art (do Carmo et al., 2019). And while neural post-editing (NPE) has shown substantial improvements when applied on PBSMT output, it has not been as effective in improving output from NMT systems. One of the reasons is that NMT and NPE typically use similar approaches, which can make the latter redundant, as it can be assimilated by the former, e.g., in some cases, by increasing the number of layers of the network. One alternative is to explore features of the data not available while training MT systems. In this paper, we explore the effect of adding tokens that identify partitions in the training data which may be relevant to guide the behaviour of the NPE system. Examples of such tokens are related to basic source and/or target sentence length or to more sophisticated analyses of the text. In this work, we explore two features: *sentence length* and *topic*.

2 Related Work

Adding a token to the input of a sequence model to shape its behaviour is not a new idea. Mikolov and Zweig (2012) aim at improving neural language models and avoid the data fragmentation in multiple datasets by using Latent Dirichlet Allocation (Blei et al., 2003) to construct context vectors and represent topics. Sennrich et al. (2016a) call the added token a 'side constraint', which informs the system about target side features, such as honorific forms of treatment, tense, number, gender, or other grammatical or discourse features, which may not exist or be different in the source side. The authors use an automatic annotator of politeness in the tar-

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 132–138
Florence, Italy, August 1-2, 2019.

get sentences in the training set, which places a token at the end of each sentence to control the politeness level of the output of an NMT model. Yamagishi et al. (2016) also use target side annotations during training to control active versus passive voice in the output. Vanmassenhove et al. (2018) used prefixed tokens identifying the gender of the author to aid the MT system in correctly presenting gender features in discourse.

Special input tokens have also been used to aid training of single models on multilingual translation tasks: Johnson et al. (2017) prefix each source sentence in an NMT system with a token to indicate the target language, training a multilingual model on a scenario with multiple source and target languages. This approach is at the background of the research on zero-shot translation. In the context of low-resource languages, Mattoni et al. (2017) add two tokens, one to specify the source language and another to specify the target language. In their case, the source-language token is used for language specific tokenisation. Similarly, Zhou et al. (2018) found that adding tokens that encode the source and target language family, e. g. `source-family:Germanic` and `target-family:Slavic` for English-Czech translation, may improve the accuracy of the NMT outputs for low-resource languages.

Added tokens in APE were used in a scenario where SMT and NMT outputs were trained jointly in a single model (Pylypenko and Rubino, 2018). An artificial token was added to the data to indicate the system the segments had been produced from. However, this strategy was not very successful, especially when editing NMT output.

Our current work further explores the strategy of adding such tokens about data partitions in NPE. Partitions are derived according to topic models or sentence lengths. Topic models are trained separately on the provided data and aim to identify the topic of each segment of the data.

3 Data and Labels

While the shared task is open to using additional data sources, we only use the data sets linked on the shared task website, aiming at better result reproducibility: i. e. (a) the authentic English-German WMT 2018 APE shared task data (Turchi et al., 2018), (b) the synthetic English-German data of the WMT 2016 AmuNMT system (Junczys-Dowmunt and Grundkiewicz, 2016), (c) the NMT part of the synthetic English-German data of the eSCAPE corpus (Negri et al., 2018), (d) the authentic English-Russian data new in the WMT 2019 APE shared task provided by Microsoft[1] and (e) the synthetic English-Russian data of the eSCAPE corpus.

3.1 Training Data

For the EN-DE experiments, we used the 500k and 4M triplets defined in (Junczys-Dowmunt and Grundkiewicz, 2016). For EN-RU, we used the 8M triplets from the eSCAPE project. Table 1 and Table 2 show statistics about the data used to train our systems.

Size	EN-DE	EN-RU
small	268 840	301 780
medium	795 208	N/A
large	4 660 020	8 037 141

Table 1: Number of SRC-NMT-PE triplets distributed over three data sets used in our experiments.

3.2 Induction of Topic Clusters

We induce ten topic clusters for each language pair using Scikit-Learn's implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We use the English side of the data. The data is the concatenation of the authentic and a sample of the synthetic data. For English-German, we sample 50k segments each of AmuNMT (500k) and eSCAPE data (7M). For English-Russian, we sample 100k of eSCAPE data. The data was cleaned of stop words and words that occur less than five times or in more than 90% of segments.

3.3 Topic Classification

We split the data for training the LDA models, into ten files according to the induced topics and then label each sentence of *all* data according to the most similar topic file. We measure similarity with cosine similarity on character n-gram tf-idf vector representations ($n = 5, 6, 7$). Before n-gram extraction, segments are lowercased and e-mail addresses, URL numbers and characters repeated more than three times are normalised. For tf-idf values, we use plus one smoothing and we avoid zero and negative idf values by adding two to the number of documents. To represent topic clusters, we use the average of its segment vectors.

[1] `http://www.statmt.org/wmt19/ape-task.html` accessed during the task and last 2019-04-30

Size	EN-DE			EN-RU		
	SRC	NMT	PE	SRC	NMT	PE
small	10 771	15 477	18 088	9 125	14 783	15 761
medium	48 227	48 257	48 869	N/A		
large	50 327	50 538	50 790	53 030	50 646	52 970

Table 2: Vocabulary sizes (after applying BPE on the train data set).

3.4 Length Partitions

Another way of partitioning the data is by sentence length. We use the length of the source side of each segment, i. e. the English side to create a partitioning of the data according to the number of tokens. We choose the partition boundaries as thresholds on the number of tokens keeping each partition similar in size within the sample data. Size is measured as $\sum_i s_i^e$ where s_i is the number of tokens in the ith segment and $e = 0.5$. This is a compromise between counting segments ($e = 0$) and counting tokens ($e = 1$).

3.5 Pre-processing for APE Training

We use the available authentic and synthetic data as is. The authentic data, the synthetic AmuNMT data and the synthetic EN-RU eSCAPE data used for training are already tokenised, thus no further tokenisation is conducted. We do not apply lower- nor true-casing, aiming to learn how to correct errors related to the casing. We learn a byte-pair encoding (Sennrich et al., 2016b) of 50 000 operations from our training data which we then apply to split each data set into subword units. After that, the corresponding partition tokens are attached to each segment. In particular, the partition labels are attached to both source and MT segments, i.e., the two sources in our multi-source NPE systems.

4 Experiments

4.1 Objectives

Our experiments aim at two objectives: (i) to investigate the effect of extra information in the form of prefix tokens for NPE; and (ii) to assess whether monolingual SMT[2], can be effective for post-editing of NMT output. The latter is driven by the idea of added benefits from interleaving different MT technologies.

We conduct three types of NPE experiments – (a) baseline experiments, using no extra tokens to build a set of baseline systems; (b) length tokens – prefixed with tokens stating the data partition based on the length and (c) topic tokens – data is prefixed with tokens stating the data partition based on the LDA clustering. For the SMT experiments no additional tokens were attached to the text. We assumed that such augmentation of the source side would increase the difference with respect to word alignment and thus it would have a negative impact on the quality of the system.

4.2 Models

NPE We trained 15 NPE systems: *small*, *medium* and *large* for EN-DE and *small* and *large* for EN-RU, on the data discussed in Section 3.1, for the three different prefix token settings – no token, topic token, length token. For all of them, we employed Marian-NMT[3] to train multi-source sequence-to-sequence models (multi-s2s) with LSTM units.[4] The two sources are the actual source-side data (EN) from the training corpus and its translation (DE or RU). We used cross-entropy as validation metric and the max-length was 150 tokens. The training stops after 5 epochs with no improvement, i.e., early stopping.

SMT We trained 5 SMT models (*small*, *medium* and *large* for EN-DE and *small* and *large* for EN-RU) using Moses (Koehn et al., 2007) release 4.0, Giza++ (Och and Ney, 2003) for word alignment and a 5-gram KenLM language model (Heafield, 2011). Models are tuned with Mert (Och, 2003). We ought to stress that these models are monolingual, i.e., trained only on the original MT output as source and its post-edited variant as target.

[2]In this work, we use the term *monolingual* to define an MT system where the source and the target are in the same language, e.g. the source is a translated sentence in German and the target is its post-edited variant.

[3]https://marian-nmt.github.io/

[4]Options: –mini-batch-fit, –workspace 9000, –layer-normalization, –dropout-rnn 0.2 –dropout-src 0.1 –dropout-trg 0.1, –early-stopping 5, –max-length 150 –max-length-crop, –valid-freq 2000 –save-freq 2000 –disp-freq 1000

4.3 Evaluation and selection for WMT submission

We evaluated our models using BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). For the former, we used the *multi-bleu* implementation provided alongside Moses; and for the latter we used the script provided by the WMT organisation.

We computed BLEU and TER using the human PE side of the data as reference, and the NPE output as hypothesis, e.g. TER(*npe,pe*). We also computed BLEU and TER scores for the original data, i.e. in this case the reference again is the human PE but the hypothesis is the NMT part of the training data: TER(*nmt,pe*). We present our results on the development set in Tables 3 and 4 for EN-DE and EN-RU, respectively. We denote the scores for the original (baseline) MT output with *MT*. Scores are scaled between 0 and 100.

	Model	Prefix	BLEU ↑	TER ↓
MT	Baseline	N/A	76.94	15.08
NPE	small	N/A	63.28	24.09
	medium	N/A	70.57	18.81
	large	N/A	70.29	19.89
	small	topic	60.41	28.59
	medium	topic	73.08	17.81
	large	**topic**	**75.82**	**15.89**
	small	length	62.56	26.91
	medium	length	73.74	17.26
	large	**length**	**75.85**	**15.91**
SMT	small	N/A	76.82	15.33
	medium	**N/A**	**77.04**	**15.17**
	large	N/A	76.82	15.26

Table 3: BLEU and TER scores for the EN-DE NPE and SMT models (dev set). Rows in **bold** indicate submitted system results.

For submission to the shared task, we selected the best models, according to TER, available at the submission deadline. For EN-DE, these are: the NPE-large-topic (primary), the NPE-large-length and the SMT-medium; for EN-RU these are the NPE-large-length (primary) and the SMT-small. In the result tables these are marked in **bold**.

5 Results and Analysis

5.1 Development Observations

Table 3 and Table 4 show the evaluation scores (BLEU and TER) on the development set results. In our experiments, the ranking of the systems' performance scores is always the same, no matter if we use TER or BLEU.

	Model	Prefix	BLEU ↑	TER ↓
MT	Baseline	N/A	80.22	13.13
NPE	small	N/A	50.76	34.45
	large	N/A	59.01	28.01
	small	topic	48.30	41.19
	large	topic	75.39	16.18
	small	length	44.68	44.57
	large	**length**	**73.67**	**19.74**
SMT	**small**	**N/A**	**79.40**	**13.68**

Table 4: BLEU and TER scores for the EN-RU NPE and SMT models (dev set). Rows in **bold** indicate submitted system results.

We can see that all NPE systems in our experiments, whether or not they are augmented with informative tokens, are unable to perform as well as the original NMT translations. So, our NPE systems are not fulfilling their main function. Still, it is worth analysing the evolution of scores from system to system.

As expected, in general, the larger the systems, the better the results. This is most visible in the EN-DE experiments, for which we trained systems in a three-size scale. Systems with small amounts of training data deteriorate the scores very much, which makes them not viable. For augmented systems, in both languages, the addition of more data has a very visible effect, with the largest systems having the best results. The same is not true for the systems with no tokens, in which medium-sized systems achieve better scores than large ones. For the SMT systems, size of the training data was the only factor we tested, but the scores are very close for all systems, with medium-sized systems achieving slightly better results.

The addition of the tokens also has a positive effect in the scores, especially for systems trained with medium-sized and large-sized datasets. For EN-DE, in the systems trained with a small volume of data, the highest scores are for systems with no tokens. But for medium-sized trained systems, the addition of the token *length* achieves the best results. For large systems, the scores are much closer to each other, but augmented systems beat the system with no tokens. In EN-RU, the advantages of adding the tokens is also more visible for the larger datasets, with *topic* as the token that enables the highest scores.

Surprisingly, the APE systems using SMT are the best performing ones, beating all neural ones. In fact, their scores are very close to the original ones, and very consistent, seeming not to be sensitive to the increase in the volumes of training data.

5.2 Final Systems

As noted in Section 4.3 we submitted three systems for EN-DE: the NPE-large-*topic* (as primary), the NPE-large-*length* and the SMT-medium. Only the SMT system scores above the original MT system, and only in terms of BLEU. For EN-RU, we submitted two systems: NPE-large-*length* (as primary) and the SMT-small. None of the system improved on the original MT data, but the SMT system was close. The baseline scores compared to our systems' scores are presented in Table 5 and Table 6.

	System	Model	Prefix	BLEU ↑	TER ↓
MT	Baseline	N/A	N/A	74.73	16.84
NPE	Primary	large	topic	74.29	17.29
	Contrastive I	large	length	74.01	17.41
SMT	Contrastive II	medium	N/A	74.30	17.07

Table 5: BLEU and TER scores for submitted and baseline systems for the EN-DE language pair.

	System	Model	Prefix	BLEU ↑	TER ↓
MT	Baseline	N/A	N/A	76.20	16.16
NPE	Primary	large	length	72.90	18.31
SMT	Contrastive	small	N/A	75.27	16.59

Table 6: BLEU and TER scores for submitted and baseline systems for the EN-RU language pair.

We believe one of the main factors for these results is the initially high quality of the baseline MT systems. The inherent nature of APE systems dictates that they generate a whole new sentence when the inputs are passed through the model. However, in cases when no or barely any changes are required, it will be desirable not generate a new sentence, i.e. the post-edit, but to retain the original one, as any transformation process would be likely to impede the quality. In future work, we will look into combining NPE models with Quality Estimation (QE), to filter NMT output by expected quality and thus control over-correction: the NPE system will then only present alternatives for sentences that require improvements.

6 Conclusions

Although our NPE systems do not fulfill their main aim (improving the output of an NMT system), this paper highlights the potential of two strategies for APE which explore the thin improvement margins allowed by NMT output.

The augmentation strategy is a simple process that requires no system development, but presents its own challenges. The tokens that are used must be informative, so as to guide the NPE system to features in the datasets with a very close relation to the editing patterns the system is supposed to learn. Future work should check the topic model and if necessary switch to a more suitable model. Other types of tokens should also be tested. Furthermore, data augmentation in APE implies pre-analysis of the datasets, since the same tokens are not applicable to different datasets nor use-cases.

The strategy of applying a different MT paradigm, SMT for APE of NMT output, yielded interesting results, albeit still not being able to improve the original NMT output. The margin of development of SMT systems may be limited, but this is also worth experimenting, in view of the challenges APE currently faces with NMT output.

Furthermore, we outlined a hypothesis about the reasons why the post-edited texts score below the baseline system. In particular, we believe this result has to do with the high quality of the baseline MT systems: this implies that some segments should not be post-edited, but our APE system attempted to edit every sentence. We plan to incorporate QE and data selection to mitigate this over-correction issue, offering an APE suggestion only when editing is necessary.

7 Acknowledgements

This research is supported by Science Foundation Ireland through the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Félix do Carmo collaborates in this project in the ambit of a European Unions Horizon 2020 research and innovation programme, under the EDGE COFUND Marie Skłodowska-Curie Grant Agreement no. 713567. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 13/RC/2077.

References

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Félix do Carmo, Dimitar Shterionov, Joachim Wagner, Murhaf Hossari, Eric Paquin, and Joss Moorkens. 2019. A review of the state-of-the-art in automatic post-editing. *Under review for publication: Machine Translation*.

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: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Melvin Johnson, Mike Schuster, Quoc V. 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.

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*, pages 751–758, Berlin, Germany. 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, 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. Association for Computational Linguistics.

Giulia Mattoni, Pat Nagle, Carlos Collantes, and Dimitar Shterionov. 2017. Zero-shot translation for low-resource indian languages. In *Proceedings of MT Summit XVI – Vol.2 Commercial MT Users and Translators Track*, pages 1–10, Nagoya, Aichi, Japan. Asia-Pacific Association for Machine Translation.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *Spoken Language Technologies*. IEEE.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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, Philadelphia, Pennsylvania, USA.

Daria Pylypenko and Raphael Rubino. 2018. DFKI-MLT system description for the WMT18 automatic post-editing task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 836–839, Belgium, Brussels. Association for Computational Linguistics.

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, San Diego, California. 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.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

In *AMTA 2006. Proceedings of the 7th Conference of the. Association for Machine Translation of the Americas. Visions for the Future of Machine Translation*, pages 223–231, Cambridge, Massachusetts, USA.

Marco Turchi, Matteo Negri, and Rajen Chatterjee. 2018. WMT18 APE shared task: En-DE NMT train and dev data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Belgium, Brussels. Association for Computational Linguistics.