Neural Post-Editing Based on Quality Estimation

Yiming Tan, Zhiming Chen, Liu Huang, Lilin Zhang,
Maoxi Li, Mingwen Wang
School of Computer Information Engineering, Jiangxi Normal University
{tt_yymm, qqchenzhiming, ufhuangliu, lilinzhang, mosesli, mwwang}
@jxnu.edu.cn

Abstract

Automatic post-editing (APE) is a challenging task on WMT evaluation campaign. We find that only a small number of edit operations are required for most machine translation outputs, through analysis of the training set of WMT17 APE en-de task. Based on this statistics analysis, two neural post-editing (NPE) models are trained depended on the edit numbers: single edit and minor edits. The improved quality estimation (QE) approach is exploited to rank models, and select the best translation as the post-edited output from the *n*-best list translation hypotheses generated by the best APE model and the raw translation system. Experimental results on the datasets of WMT16 APE test set show that the proposed approach significantly outperformed the baseline. Our approach can bring considerable relief from the overcorrection problem in APE.

1 Introduction

Automatic post-editing (APE) aims to learn how to correct machine translation errors by use of the human post-editing feedback. The traditional statistical post-editing builds monolingual statistical phrased-based machine translation system to translate the wrong raw outputs into good translations (Simard et al., 2007; Bechara et al., 2011; Chatterjee et al., 2015). In recent years, with the great success of deep learning achieved in machine translation, many works have applied neural machine translation (NMT) to the APE task.

Pal et al. proposed to exploit the bidirectional source RNN encoder-decoder model to establish a monolingual machine translation system for APE (Pal et al., 2016). Compared with the traditional statistical post-editing approaches, their approach gained more improvement. In the light of the context information of the translation, Pecina et al. proposed to respectively establish independent encoders for source sentences and raw machine translations (Pecina et al., 2016). Their approach is similar to the multi-source NMT (Zoph et al., 2016); the difference lies in the input information are source sentences and raw machine translation outputs. Grundkiewicz et al. proposed to combine the outputs of monolingual NMT and bilingual NMT to improve the performance of APE task (Grundkiewicz et al., 2016).

This paper presents a new approach for APE which was submitted by the JXNU team to WMT17 APE shared task. In order to effectively reduce the overcorrection problem, we propose to build two specific neural post-editing (NPE) models in term of the edit numbers, and select the best model by machine translation quality estimation (QE). The experiment results indicate that the proposed approach gains great improvement over the baseline officially released by the evaluation campaign.

2 Data analysis

Overcorrection problem refers to edit the machine translation output more times than it really needed, among these edit operations, some are not necessary or even wrong. Overcorrection may cause the resulting outputs of APE have lower translation quality than the raw translation outputs. To estimate the number of edit operations needed on the test set, we count the number of edit operations, including deletion, insert, substitution, and shift of word chunk, for the raw machine translation outputs on the training set of WMT16 and WMT17 APE shared task by the open source

TER script[1]. The combination training set has 23,000 triples that are source sentence, raw machine translation output, and its human reference translation.

The distribution of the number of edit operations needed for raw machine translation outputs on the training set of WMT16 and WMT17 APE shared task are showed as Figure 1. The statistics indicate that the average number of edit operations for the raw machine translation outputs is 4. And the machine translation outputs need more than 1 edit operation account for 20.47%, while 58.03% of machine translation outputs need to be edited 4 times or less.

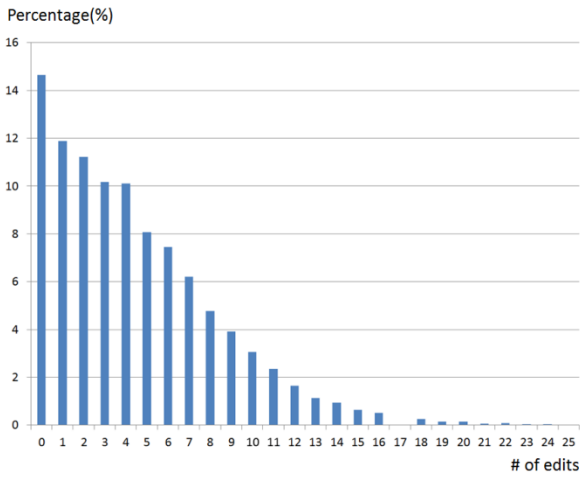

Figure 1: Distribution of the number of edit operations needed for machine translation outputs in the training set of WMT16 and WMT17 APE shared task.

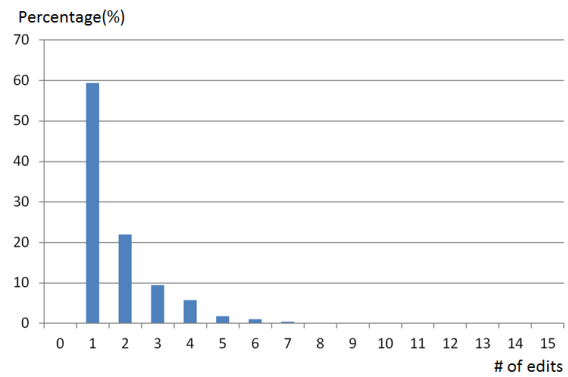

Figure 2: Distribution of the number of only one type edit operations needed for machine translation outputs.

Because the raw machine translation outputs can be converted to good translation by deletion, insert, substitution, and shift of word chunk operations, we also extract the machine translation outputs that only one type of edit operation are needed to convert them into good translation, the distribution of the number of edit operations on the subset is shown as Figure 2, it shows that more than 80% of raw machine translation outputs needed 2 or less one type edit operations.

3 Model

From the distribution of the number of edit operations in the training set, there are a lot of raw machine translation outputs needed a small amount of edit operations, less than 4 times; and there also exist a lot of raw machine translation outputs needed only one type edit operations. Thus, we speculate that this phenomenon is also available for the test set. In order to reduce the overcorrection in the test set, we train two NPE models aiming at these two conditions.

Follow by Grundkiewicz et al. (2016) work, a NPE model is build and trained with the training set officially released by the evaluation campaign, called $\text{NPE}_{\text{BASELINE}}$.

We extract a triplet corpus with raw machine translation outputs needed 4 or less edit operations from the training dataset, and train a NPE system, called $\text{NPE}_{\text{MINOR}}$. In the meantime, in order to strengthen the ability of editing the raw machine translation outputs by one single type edit operations, we use a triplet corpus contained machine translations with 2 or less one single edit operations from the training dataset, and train a NPE system, called $\text{NPE}_{\text{SINGLE}}$.

In order to combine $\text{NPE}_{\text{BASELINE}}$, $\text{NPE}_{\text{MINOR}}$ and $\text{NPE}_{\text{SINGLE}}$, we merge outputs of these three systems which are regarded as an *n*-best list translation hypothesis, and introduce the sentence-level QE approach (Specia et al., 2013) to score and rank the *n*-best list translation hypothesis.

QE approach aim is to estimate the qualities of translation without human references on the basis of features abstracted from the source sentences and machine translation outputs which reflect translation complexity, fluency and adequacy.

Adopted the sentence-level QE approach to score and rank translation outputs in the *n*-best lists, we find that the QE approach can be proved to be very effective when it comes to one source sentence with great difference in qualities of translation, however, it's not very effective when one source sentence with small difference in qualities of translation.

[1]http://www.cs.umd.edu/~snover/tercom/

In order to reduce the impact of misjudgment, a hierarchical classification method is used to select the best translation output among the merged *n*-best list. First, the translation hypotheses are score by the QE method and the scores are converted into to the five-point scale. Thus, if the qualities of translation hypotheses are classified into different level, they can be ranked according to the quality level; if they are in the same level, a statistical language model, SRILM (Stolckeet al., 2002), is introduced to score and rank the translation hypotheses to get the best one.

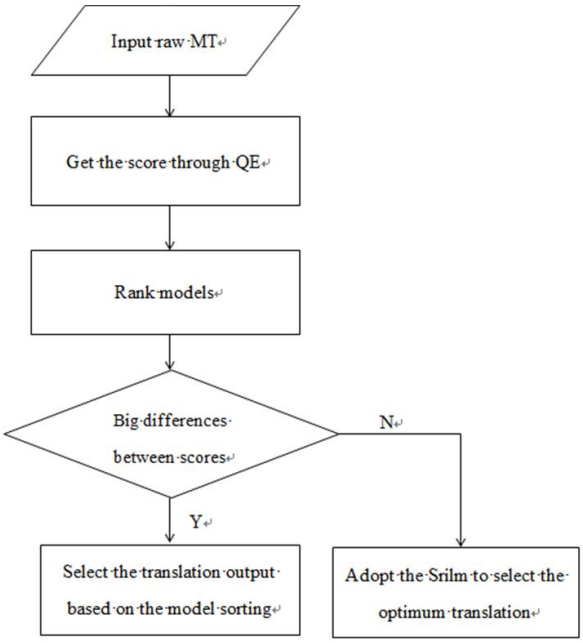

Figure 3: The flow chart of how to select the best translation by the QE approach

4 Experiments

In order to test the performance of the proposed approach, we conduct experiment on the test set of the WMT16 APE Task. The task focuses on the information technology domain, in which English source sentences have been translated into German (en-de) by an unknown MT system. The goal of the APE shared task is to examine automatic methods for correcting errors.

4.1 Experiments setting

Experimental data consist of corpus of WMT16 and WMT17 APE shared task released by the evaluation campaign, and publicly released artificial post-editing data (Grundkiewicz et al., 2016), including source language sentences, raw machine translation outputs and human references. Table 1 shows more details about this corpus.

Due to the provided training triplets for en-de direction is too small to train neural models, Grundkiewicz et al. created artificial training triplets through applying cross-entropy filtering and round-trip translation to extend the provided training triplets and publicly released the extended one (Grundkiewicz et al., 2016). Therefore, we integrate these two corpora into a training set for training NPE systems.

Data set	Sentences	length	TER
WMT16 training set	12,000	17.89	26.22
WMT17 training set	11,000	17.69	24.41
WMT16 development set	1,000	19.75	24.81
WMT16 test set	2,000	17.41	24.76
Artificial data 500K	531,839	20.92	25.28
Artificial data 4M	4,335,715	15.86	36.63

Table 1: Statistics of the provided data sets: number of sentences, average sentence lengths and TER score.

The sentences in the corpus have been tokenized and truecased when preprocessing. To deal with the limited ability of neural translation models to handle out-of-vocabulary words, tokens are splited into subword units (Sennrich et al., 2015b) to improve the systems' performance.

We apply Nematus[2] to train the bidirectional RNN encoder-decoder model with attention mechanism. The size of minibatches is set 80, vocabulary size is set 40000, maximum sentence length is set 50, the dimension of word embeddings is set 500, the size of hidden layers is set 1024, and the optimization algorithm proposed by Adadelta (Zeiler, 2012) is used. Compared with Nematus's approach, AmuNMT[3] based on C++/CUDA (Grundkiewicz et al., 2016) decode at a faster speed on CPU. Thus, we apply AmuNMT's approach to decode to-be-edited machine translations with a beam size of 12 and length normalization when decoding.

4.2 Experiments result

4.2.1 $NPE_{BASELINE}$ system

The APE corpus with size of 4M is used to train the $NPE_{BASELINE}$ system, while the combined corpus of APE corpus with size of 500k and the

[2]http://github.com/rsennrich/nematus
[3]http://github.com/emjotde/amunmt

WMT16 and WMT17 training set are used to optimize the parameters of the system.

4.2.2 NPE_{MINOR} & NPE_{SINGLE} systems

Filtered the above training set by the following rules respectively: machine translations needed 4 or less edits and machine translations needed 2 or less single edit operations, two sub training sets, contained 278.9 K and 160.6 K training triples, are obtained. At the same time, the development set of the WMT16 APE shared task are filtered by the rules, and two sub development sets, contained 1199 and 810 triples, are obtained.

System	TER	BLEU
Raw MT output	12.66	76.13
$NPE_{BASELINE}$	12.20	78.53
NPE_{MINOR}	**10.24**	**81.80**

Table 2: System performance of the NPE_{MINOR} and the $NPE_{BASELINE}$ systems in the sub development set.

System	TER	BLEU
Raw MT output	8.25	82.31
$NPE_{BASELINE}$	8.04	84.48
NPE_{MINOR}	6.20	88.07
NPE_{SINGLE}	**5.58**	**89.02**

Table 3: System performance of NPE systems in the sub development set.

We respectively train and tune the $NPE_{BASELINE}$ model with the sub training set and sub development set, two NPE systems, called NPE_{MINOR} and NPE_{SINGLE}, are gained. The system performance on the two sub development sets are shown in Table 2 and Table 3.

4.2.3 Joint system

To gain better system performance, the outputs of NPE systems and raw machine translations were combined into an *n*-best list of translation hypotheses. The improved machine translation QE was exploited to select the best outputs among the *n*-best list.

As shown in Table 4, the system performance of combining the outputs of $NPE_{BASELINE}$ and NPE_{MINOR} systems and raw machine translations gained 0.7 TER score and 1.76 BLEU score improvement over that of the $NPE_{BASELINE}$ system in the test set of WMT16 APE shared task. The system performance was further improved by 0.75 TER score and 0.61 BLEU score when combined the NPE_{SINGLE} outputs. The result shows the effectiveness of the proposed approach.

System	TER	BLEU
Baseline1(Raw MToutput)	24.76	62.11
Baseline2(Moses PBAPE)	24.64	63.47
$NPE_{BASELINE}$	**23.78**	**64.97**
$NPE_{BASELINE}$ + NPE_{MINOR}	23.08	66.73
$NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE}	**22.33**	**67.34**

Table 4: Results of NPE systems in the WMT16 test set

4.3 Analysis

In order to look into the reasons for system performance improvement, we extract 500 triples from the test set of WMT16 APE shared task, in which the $NPE_{BASELINE}$ system performed worse than the raw machine translations. The machine translations in the 500 triples are all over-corrected by the $NPE_{BASELINE}$ system, however, the total amount of sentences occurring overcorrection reduce to 372 in the outputs of the jointed models. And it was found that 58.8% of machine translation sentences only need 4 or less edits, this illustrates that the jointed model contributes greatly to reducing overcorrection.

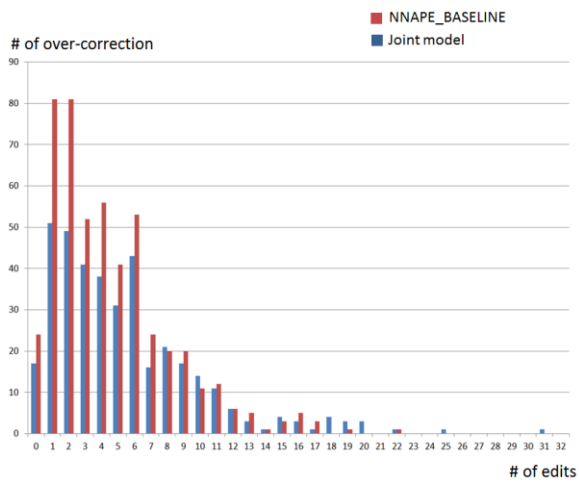

Figure 4: Distribution of the number of edits needed in overcorrection sentences from outputs of $NPE_{BASELINE}$ and jointed systems.

To show their differences on the number of edits more clearly, Figure 4 describes the distribution of the number of edits from outputs of $NPE_{BASELINE}$ system and jointed systems. The Figure 4 reveals that the frequency of

overcorrection of the joint system is lower than the $NPE_{BASELINE}$ system when corrected machine translation needed a small amount of edits (<=4).

5 Conclusion

Our submission to the WMT17 APE shared task en-de translation direction gains significantly improvements over the baselines, scoring 23.30 on TER and 65.66 on BLEU in the official results. This indicates that it is necessary to build a NPE system for machine translations needed a smaller amount of edits. Future work should include the investigation of the proposed approach application to the de-en translation direction of the WMT APE shared task.

Acknowledgements

This research has been funded by the Natural Science Foundation of China under Grant No.6166 2031, 6146 2044, and 61462045. The authors would like to extend their sincere thanks to the anonymous reviewers who provided valuable comments.

References

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

Hanna Béchara, Yanjun Ma and Josef van Genabith. 2011. Statistical post-editing for a statistical MT system. In *MT Summit*, pages 308-315, Xiamen, China.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christ of Monz, MatteoNegri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Rajen Chatterjee, Marco Turchi, and Matteo Negri. 2015a. The FBK participation in the WMT15 automatic post-editing shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 210–215, Lisbon, Portugal.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015b. Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages156–161, Beijing, China.

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, August.

Kevin Knight and Ishwar Chander. 1994. Automated post-editing of documents. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 779-784, Seattle, WA, USA.

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 Association for Computational Linguistics*, pages 177–180.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josefvan Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 281–286, August.

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, ACL'02*, pages 311–318, Stroudsburg, PA, USA.

Kashif Shah, Raymond W. M. Ng, Fethi Bougares and Lucia Specia. 2015. Investigating Continuous Space Language Models for Machine Translation Quality Estimation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1073-1078, Lisbon, Portugal.

Lucia Specia, Kashif Shah, Jose G.C. de Souza and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79-84, Sofia, Bulgaria.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515, Rochester, New York.

Matthew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla and Ralph Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of*

Association for Machine Translation in the Americas, pages 223-231, Cambridge.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv*: 1511.06709.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv*: 1508.07909.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing, volume 2*, pages 901-904