Automatic Grammatical Error Correction for Sequence-to-sequence Text Generation: An Empirical Study: Supplementary Notes

Tao Ge **Xingxing Zhang** **Furu Wei** **Ming Zhou**
Microsoft Research Asia
{tage, xizhang, fuwei, mingzhou}@microsoft.com

1 Model Configuration

We present detailed configuration of the models we implemented in Section 3 of our paper submission.

1.1 Unsupervised SMT/NMT models

Our unsupervised SMT/NMT models are the implementation of Ren et al. (2019) who use 50 million monolingual sentences in NewsCrawl as previous work (Lample et al., 2018) to train MT models in the unsupervised setting. Specifically, word2vec[1] is used to train monolingual word embeddings of each language and vecmap[2] is employed to obtain cross-lingual embeddings.

The NMT model configuration is almost the same with the Transformer model (Vaswani et al., 2017). The vocubulary is a shared 50k BPE codes for both source and target languages. The SMT model is based on the Moses implementation of PBSMT systems with Salm (Johnson et al., 2007) and use default features defined in Moses.

1.2 Style transfer models

The base model for formality style transfer is a 2-layer transformer model with 4 heads. We set the embedding dimension to 256 and the hidden dimension of the feed-forward sub-layer to 1,024. The vocabulary is shared by the source and the target, which is the most frequent 20k BPE codes. We train the model with Adam with learning rate of 0.0005, $\beta_1 = 0.9$, $\beta_2 = 0.997$, learning rate warmup over the first 8,000 steps and inverse square root decay of the learning rate.

Following Xu et al. (2019), we use a Convolutional Neural Network (CNN) model as the style classification model which is used to evaluate style accuracy. The convolutional layer's filter sizes are [3, 4, 5], which is followed by a max-pooling layer. The result is then passed to a fully connected softmax layer to predict the style label (i.e., formal or informal). The CNN model is trained with the 200K sentences with style labels in the GYAFC corpus (Rao and Tetreault, 2018). The accuracy evaluated on the test set in GYAFC is approximately 93%.

Both of the transformer model and CNN model are tuned on the dev set in GYAFC.

1.3 Sentence compression model

We use a 2-layer LSTM seq2seq model, which generates a 0/1 sequence to indicate whether to delete a word, as our sentence compression model based on the idea of Filippova et al. (2015). The vocabulary size is the most frequent 50k words in the training set. The model is optimized by Adam with the learning rate of 0.0002 and tuned on the dev set.

2 Experiments on WMT17 APE task

We conduct experiments on the WMT17 Automatic Post Editing (APE) task. The results are shown in Table 1.

According to Table 1, it seems that our GEC post editing introduces many errors and decreases the translation quality. However, when we manually check and analyze the results, we find it is not

	TER	BLEU
SMT w/o post-editing	15.55	79.54
GEC post editing	16.14	78.55
State-of-the-art	15.29	79.82

Table 1: Results on the WMT17 APE shared task. For TER (short for Translation Error Rate), the lower, the better; For BLEU, the higher, the better. The state-of-the-art approach is the top performing system (Chatterjee et al., 2017) on the WMT17 APE shared task.

[1] https://github.com/tmikolov/word2vec
[2] https://github.com/artetxem/vecmap

BLEU change	Reasons	Examples
54↑	Correction (87.0%)	Base: One gradually reduction in dose or frequency does not appear to infants. (84.2) GEC: One gradual reduction in dose or frequency does not appear to infants. (100) REF: One gradual reduction in dose or frequency does not appear to infants.
	Accidental (13.0%)	Base: The clinical significance of the observed changes in HBV DNA, it is unclear. (67.3) GEC: The clinical significance of the observed changes in HBV DNA, is unclear. (72.7) REF: The clinical significance of the observed changes in HBV DNA are unclear.
240↓	Reference Error (52.9%)	Base: The MAH will continuously will continue to submit yearly PSURs. (100) GEC: The MAH will continuously continue to submit yearly PSURs. (71.1) REF: The MAH will continuously will continue to submit yearly PSURs.
	Correction (25.5%)	Base: Excretion is rapidly and predominantly in the faeces. (75.1) GEC: Excretion is rapid and predominantly in the faeces. (61.0) REF: Excretion occurs rapidly and predominantly in the faeces.
	GEC Error (21.6%)	Base: Patients may not be reconstituted product in use at room temperature for one single period of up to 4 weeks before use. (79.9) GEC: Patients may not be reconstituted product in use at room temperature for a single period of up to 4 weeks before use. (67.3) REF: Patients may store the unreconstituted product in use at room temperature for one single period of up to 4 weeks before use.

Table 2: Reasons for the BLEU changes of the sentences edited by GEC.

true.

Table 2 shows the reasons for the BLEU changes of the sentences edited by GEC. To our surprise, 53% of the cases where BLEU decreases after GEC post editing are due to grammatical errors in the reference sentences. Since the references are edited by humans on the MT outputs, it is very common that human annotators overlooked the grammatical errors in the MT outputs, resulting the existence of grammatical errors in the references. In such cases, GEC corrects the errors yet makes BLEU and TER become worse.

Base: Uncommon: thrombocythaemia, leukocytosis.
GEC: Uncommon: thrombocythaemia and leukocytosis.
REF: Occasionally: thrombocythaemia, leukocytosis.

Table 3: GEC errors in a sentence with a special writing style.

Although it is undeniable that GEC sometimes makes a mistake, as *GEC Error* in Table 2 shows, it usually brings negligible adverse effects to the translation quality. It is notable that among all the GEC errors, approximately 27% are due to the special writing style in some sentences, as shown in Table 3. Therefore, as we conclude in our paper submission, GEC is more beneficial to the seq2seq text generation tasks where target sentences should be in a formal writing style.

References

Rajen Chatterjee, M Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: Fbks participation in the wmt 2017 ape shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638.

Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 129–140.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *arXiv preprint arXiv:1901.04112*.

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.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.