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 vocabulary 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 GYAF corpus (Rao and Tetreault, 2018). The accuracy evaluated on the test set in GYAF is approximately 93%.

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

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...
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.

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


