Improved Data Augmentation for Translation Suggestion

Hongxiao Zhang[1], **Siyu Lai**[1], **Songming Zhang**[1], **Hui Huang**[2], **Yufeng Chen**[1]*
Jinan Xu[1] and **Jian Liu**[1]
[1]Beijing Jiaotong University, Beijing, China
[2]Harbin Institute of Technology, Harbin, China
{hongxiaozhang,siyulai,smzhang22,chenyf,jaxu,jianliu}@bjtu.edu.cn,
huanghui_hit@126.com

Abstract

Translation suggestion (TS) models are used to automatically provide alternative suggestions for incorrect spans in sentences generated by machine translation. This paper introduces the system used in our submission to the WMT'22 Translation Suggestion shared task. Our system is based on the ensemble of different translation architectures, including Transformer, SA-Transformer, and DynamicConv. We use three strategies to construct synthetic data from parallel corpora to compensate for the lack of supervised data. In addition, we introduce a multi-phase pre-training strategy, adding an additional pre-training phase with in-domain data. We rank second and third on the English-German and English-Chinese bidirectional tasks, respectively.

1 Introduction

Translation suggestion (TS) is a scheme to simplify Post-editing (PE) by automatically providing alternative suggestions for incorrect spans in machine translation outputs. Yang et al. (2021) formally define TS and build a high-quality dataset with human annotation, establishing a benchmark for TS. Based on the machine translation framework, the TS system takes the spliced source sentence $\mathbf{x}$ and the translation sentence $\tilde{\mathbf{m}}$ as the input, where the incorrect span of $\tilde{\mathbf{m}}$ is masked, and its output is the correct alternative $\mathbf{y}$ of the incorrect span. The TS task is still in the primary research stage, to spur the research on this task, WMT released the translation suggestion shared task.

This WMT'22 shared task consists of two subtasks: Naive Translation Suggestion and Translation Suggestion with Hints. We participate in the former, which publishes the bidirectional translation suggestion task for two language pairs, English-Chinese and English-German, and we participate in all language pairs.

Our TS systems are built based on several machine translation models, including Transformer (Vaswani et al., 2017), SA-Transformer (Yang et al., 2021), and DynamicConv (Wu et al., 2018). To make up for the lack of training data, we use parallel corpora to construct synthetic data, based on three strategies. Firstly, we randomly sample a sub-segment in each target sentence of the golden parallel data, mask the sampled sub-segment to simulate an incorrect span, and use the sub-segment as an alternative suggestion. Secondly, the same strategy as above is used for pseudo-parallel data with the target side substituted by machine translation results. Finally, we use a quality estimation (QE) model (Zheng et al., 2021) to estimate the translation quality of words in each translation output sentence and select the span with low confidence for masking, and then, we utilize an alignment tool to find the sub-segment corresponding to the span in the reference sentence and use it as the alternative suggestion for the span.

Considering that there is a domain difference between the synthetic corpus and the human-annotated corpus, we add an additional pre-training phase. Specifically, we train a discriminator and use it to filter sentences from the synthetic corpus that are close to the golden corpus, which we deem as in-domain data. After pre-training with large-scale synthetic data, we perform an additional pre-training with in-domain data, thereby reducing the domain gap. We will describe our system in detail in Section 3.

2 Related Work

The translation suggestion (TS) task is an important part of post-editing (PE), which combines machine translation (MT) and human translation (HT), and improves the quality of translation by correcting incorrect spans in machine translation outputs by human translators. To simplify PE, some early scholars have studied translation prediction (Green

*Yufeng Chen is the corresponding author.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1211 - 1216

et al. (2014), Knowles and Koehn (2016)), which provides predictions for the next word (or phrase) when given a prefix. And some scholars have also studied prediction with the hints of translators (Huang et al., 2015).

In recent years, some scholars have devoted themselves to researching methods to provide suggestions to human translators. Santy et al. (2019) present a proof-of-concept interactive translation system that provides human translators with instant hints and suggestions. Lee et al. (2021) utilize two quality estimation models and a translation suggestion model to provide alternatives for specific words or phrases for correction. Yang et al. (2021) propose a transformer model based on segment-aware self-attention, provide strategies for constructing synthetic corpora, and released the human-annotated golden corpus of TS, which became a benchmark for TS tasks.

3 Method

In this section, we describe the translation suggestion system, followed by our strategies for building synthetic corpora, and finally the details of the additional pre-training phase.

3.1 Translation Suggestion System

As defined by Yang et al. (2021), given the source sentence $\mathbf{x}$, its translation sentence $\mathbf{m}$, the incorrect span $\mathbf{w}$ in $\mathbf{m}$, and its corresponding correct translation $\mathbf{y}$, the translation suggestion task first masks the incorrect span $\mathbf{w}$ in $\mathbf{m}$ to get $\mathbf{m}^{-\mathbf{w}}$, and then maximizes the following conditional probabilities:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{m}^{-\mathbf{w}}; \boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta}$ is the parameters of the model.

The construction of the TS system is based on common machine translation models. We introduce the models used in our TS system below:

- **Transformer-base (Vaswani et al., 2017).** The naive transformer model. The encoding and decoding layers are both set to 6, the word embedding size is set to 512, and the attention head is set to 8.

- **Transformer-big (Vaswani et al., 2017).** The widened transformer model. The encoding and decoding layers are both set to 6, the word embedding size is set to 1024, and the attention head is set to 16.

- **SA-Transformer (Yang et al., 2021).** The segment-aware transformer model, which replaces the self-attention of the naive transformer with the segment-aware self-attention, further injects segment information into the self-attention, so that it behaves differently according to the segment information of the token. Its parameter settings are the same as those of Transformer-base.

- **DynamicConv (Wu et al., 2018).** The dynamic convolution model that predicts a different convolution kernel at every time-step. We set both encoding gated linear unit (GLU) and decoding GLU to 1 in the experiment.

3.2 Build Synthetic Corpora

Since there are few golden corpora available for training, it is necessary to build a synthetic corpus to make up for the lack of data. We build synthetic data through the following three strategies and use the mixed data for model pre-training.

3.2.1 Building on Golden Parallel Data

Following the method of Yang et al. (2021), we construct synthetic data on the large-scale golden parallel corpus. Given a sentence pair $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{r} = \{r_1, r_2, \ldots, r_m\}$ from the golden parallel corpus, we randomly sample a sub-segment $\mathbf{w} = \{r_i, r_{i+1}, \ldots, r_j\}$ of $\mathbf{r}$, we mask the sub-segment in sentence $\mathbf{r}$ to get $\mathbf{r}^{-\mathbf{w}} = \{r_1, r_2, \ldots, r_{i-1}, [\text{MASK}], r_{j+1}, \ldots, r_m\}$, and use $\mathbf{w}$ as an alternative suggestion. We perform statistics on the length of golden data to determine the length of masked spans, which is more in line with the golden distribution.

3.2.2 Building on Pseudo Parallel Data

The prediction of alternative suggestions requires the translation context, which cannot be provided by the golden parallel corpus. Therefore, we use the MT model provided by the shared task to infer the source of the large-scale parallel corpus to generate the pseudo-parallel corpus. Then we still follow Yang et al. (2021) and use the same way as described in Section 3.2.1 to construct synthetic data on the pseudo corpora consisting of source sentences and machine translation output sentences.

3.2.3 Building with Quality Estimation

The TS task is to predict the correct alternative proposal given the translation context. However, when sampling on the golden parallel corpus, the context

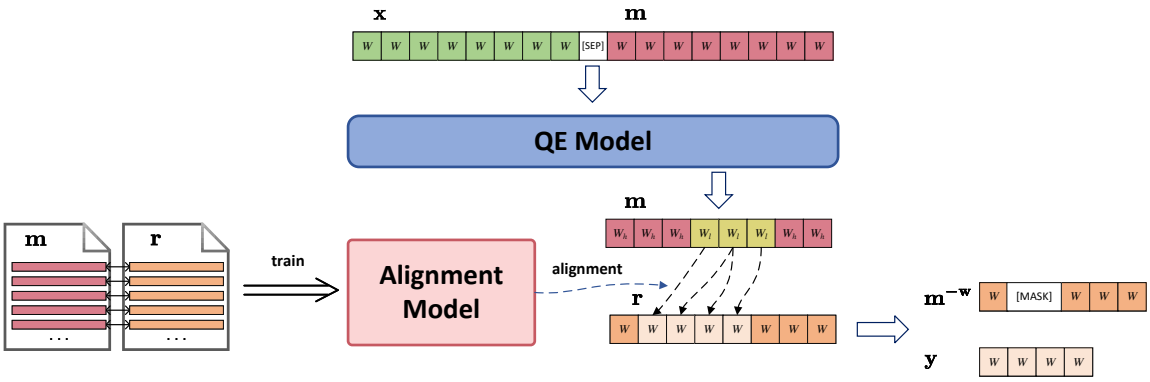

Figure 1: Schematic diagram of building synthetic corpora with quality estimation. $\mathbf{x}$ is the source sentence, $\mathbf{m}$ is the machine translation sentence, $\mathbf{r}$ is the reference sentence, and W_h and W_l represent words with high and low confidence, respectively.

does not match the translation output, and when sampling on the pseudo-parallel corpus, the alternative suggestions may be incorrect. Therefore, the above two construction strategies are not optimal.

We explore a method that is closer to the real scenarios, as shown in Figure 1. First, the word-level translation quality estimation (QE) model is used to estimate the confidence of the words in each translation sentence, and the continuous span with low confidence (that is, poor translation) is selected. Then, the translation sentence is aligned with the reference sentence through the alignment model, and the sub-segment corresponding to the span in the reference is selected as an alternative suggestion.

More specifically, we use a masked language model as our QE model, following the method of Zheng et al. (2021). To train the QE model, we splice the source sentence $\mathbf{x}_i$ and the reference sentence $\mathbf{r}_i$ of the large-scale golden parallel corpus, where some words in $\mathbf{r}_i$ are masked to get $\mathbf{r}_i^{-w}$, and the QE model is optimized to minimize the following loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} \log p(\mathbf{r}_i^w | \mathbf{x}_i, \mathbf{r}_i^{-w}; \boldsymbol{\theta}) \qquad (2)$$

where N is the number of golden parallel sentences, $\mathbf{r}_i^w$ is the masked part of the reference sentence and $\boldsymbol{\theta}$ is the model parameter.

During inference, the source and translation sentences of the pseudo-parallel corpus are spliced and fed into the QE model. The model scores the word of the translation sentence according to the recovery probability of it after being masked, and words with lower scores are considered poor translations.

After that, we train a word alignment model (Lai et al., 2022) using the translated sentences and reference sentences. To ensure high alignment quality, we filter out sentences with lengths less than 5 and greater than 100 and randomly sample 5M sentence pairs for training. We use the trained alignment model to align the machine translation sentence and the reference sentence. The sub-segment in the reference that aligns with the poorly translated span described above is selected as an alternative suggestion.

3.3 Additional Pre-Training Phase with In-Domain Data

The sources of data used to construct large-scale synthetic corpus and human-annotated golden corpus are domain different. To bridge this difference, we introduce an additional pre-training stage. We filter data similar to the golden corpus as in-domain data, which are used as pre-training for the next phase after pre-training model with a large-scale synthetic corpus.

In particular, we use BERT (Devlin et al., 2019) to construct a discriminator to identify in-domain data. The discriminator consists of a binary classifier trained to distinguish between in-domain and out-of-domain sentences. The source sentences from the golden corpus as positive examples and source sentences from the synthetic corpus as negative examples are used to train this discriminator. We upsample the golden corpus by 10 times, and randomly subsample the same amount of sentences from the synthetic corpus. For each input source sentence, the discriminator predicts the probability that the sentence is in-domain. Sentences with probabilities greater than a certain threshold are

Direction	Train	Valid	Test
en⇒de	12387	1890	989
de⇒en	9308	1849	986
en⇒zh	14759	2733	1000
zh⇒en	15207	2767	1000

Table 1: The statistics of golden corpora in four translation directions.

Corpus	golden	pseudo	with QE
LS en⇔de	9.8M	9.8M	4.7M
LS en⇔zh	20M	20M	–
IND en⇒de	0.8M	0.8M	0.4M
IND de⇒en	0.7M	0.7M	0.3M

Table 2: Statistics of constructed synthetic data in our experiments, where LS stands for large-scale data and IND stands for in-domain data.

discriminated as in-domain sentences.

After the above two phases of pre-training, we use the human-annotated golden corpus for fine-tuning and test the final model.

4 Experiments and Results

4.1 Setup

We have submitted English-Chinese (en-zh) and English-German (en-de) bidirectional translation suggestion tasks. We mix en-zh data from WMT'19 and WikiMatrix, and en-de data from WMT'14 and WikiMatrix, respectively, to construct a synthetic dataset. We use the golden Train, Valid and Test set provided by this shared task, and the data statistics are shown in Table 1. We follow Yang et al. (2021) to preprocess the data, and mix the data constructed by the three strategies described in Section 3.2 as our large-scale synthetic data. The statistics of the constructed large-scale (LS) synthetic data and in-domain (IND) synthetic data are shown in Table 2. Note that for the experiments in the en-zh translation direction, we do not apply the construction strategy with QE and the pre-training phase with in-domain data. All our models are implemented based on Fairseq (Ott et al., 2019). We use the same data on each model for two phases of pre-training and fine-tuning.

System	Translation direction			
	zh-en	en-zh	de-en	en-de
Baseline	25.51	**36.28**	31.20	29.48
Ours	**28.56**	33.33	**36.30**	**42.61**

Table 3: BLEU scores on the WMT 2022 TS test set.

System	BLEU
Do nothing	18.24
+ on golden and pseudo corpus	26.91
+ with quality estimation	30.72
+ IND pre-training phase	32.95

Table 4: BLEU scores on the English-German development set for systems based on the SA-Transformer model under different strategies.

Model	BLEU
Transformer-base (A)	32.92
Transformer-big (B)	34.73
SA-Transformer (C)	32.95
DynamicConv (D)	34.03
Ensemble (A + B + C + D)	**35.81**

Table 5: BLEU scores on the development set for systems under different models in the English-German direction.

4.2 Results

We report the results of our method on the development and test set of the translation suggestion task of WMT'22. SacreBLEU[1] is used to compute the BLEU score as quality estimates relative to a human reference. We report the experimental results of our system and the baseline system (Yang et al., 2021) on the test set in Table 3, and for the baseline system, we directly use their experimental results.

As can be seen from Table 3, our system beats the baseline system in three translation directions, especially in the en-de direction, where our system surpasses the baseline by 13.13 BLEU.

We also report the results of the system on the development set of English-German translation directions to analyze the effectiveness of different models and strategies. In Table 4, we show the results of the system based on the SA-Transformer model under different strategies. "Do nothing" means we only train with the provided training set. It can be seen that the strategy of constructing synthetic data with quality estimation (QE) and the additional pre-training with the in-domain (IND) data stage can

[1] https://github.com/mjpost/sacrebleu

bring about a great improvement.

In Table 5, we present the results of systems based on different models and the model ensemble. It can be seen that in the case of the single-model system, the Transformer-big and Dynamic-Conv models achieve better results. Besides, the ensemble model brings obvious improvement and achieves the best results.

5 Conclusion

We describe our contribution to the Translation Suggestion Shared Task of WMT'22. We propose a strategy to construct synthetic data with the quality estimation model to make the constructed data closer to the real scenarios. Furthermore, we introduce an additional phase of pre-training with in-domain data to reduce the gap between synthetic corpus and golden corpus. Experimental results demonstrate the effectiveness of our strategy. Considering the heavy labor of annotating TS data, we think data augmentation is the most important strategy that should be addressed. In the future, we will put more effort into the data generation method, to make the most of openly-accessible parallel data.

Limitations

The strategy of constructing synthetic data based on quality estimation proposed in this paper can automatically sample the incorrectly translated spans in the translations, and find the correct alternative suggestions through the alignment. It is a solution that conforms to real scenarios, and the experimental results have also proved that it is effective. However, our approach to generating synthetic data via QE still has some limitations. First, the quality estimation and alignment phases require a large additional time overhead. And second, the segments from the reference sentences may not fit into the context of the masked translation sentences due to grammar constraints. We hope to explore better solutions in future research.

Acknowledgements

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976016, 61976015, and 61876198). The authors would like to thank the WMT'22 shared task organizers for organizing this competition and for providing open source code and models, as well as the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Spence Green, Sida I Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D Manning. 2014. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1225–1236.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: integrating machine translation effectively and imperceptibly. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Proceedings of the Association for Machine Translation in the Americas*, page 107–120.

Siyu Lai, Zhen Yang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022. Cross-align: Modeling deep cross-lingual interactions for word alignment. *arXiv preprint arXiv:2210.04141*.

Dongjun Lee, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. 2021. Intellicat: Intelligent machine translation post-editing with quality estimation and translation suggestion. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 11–19.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. Inmt: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2018. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Zhen Yang, Yingxue Zhang, Ernan Li, Fandong Meng, and Jie Zhou. 2021. Wets: A benchmark for translation suggestion. *arXiv preprint arXiv:2110.05151*.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334.