# Translation Suggestion based on Pseudo Data generated from Word Alignment

Fuzhu Zhu[1] Xiaotian Wang[1] Takehito Utsuro[1] Masaaki Nagata[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
s2220804_@s.tsukuba.ac.jp

## Abstract

In this paper, we propose a method for generating pseudo data for WMT22 Translation Suggestion task using word alignments. Furthermore, we also adopt the method of Data Augmentation to improve the final performance of the models. Meanwhile, we propose to apply the pre-trained mT5 model to WMT22 Translation Suggestion task. For the results, our proposed approach exceeds the baseline on the Translation Suggestion direction of En-Zh pairs. In addition, the performance of the mT5 model illustrates the possibility of applying pre-trained seq2seq model to WMT22 Translation Suggestion task.

## 1 Introduction

With the development of the machine translation field, an increasing number of specialized tasks and demands are being continually proposed. Post-editing is known as editing machine translation to improve its quality. However, manual post-editing is costly, leading to the proposal of methods which aim to assist post-editing. Translation Suggestion (TS) is one of these methods, which has proven its ability in improving the efficiency of post-editing [1] [2]. To spur the research in TS task, Yang et al. [3] have created a golden corpus dataset, WeTS, for TS task. Besides this, due to the scarcity of WeTS, Yang et al. [3] also propose several methods for generating pseudo data[1]. As for the model, Yang et al. [3] propose the segment-aware self-attention based Transformer for TS task. Through their work, we obtain a benchmark for TS task, including generating the pseudo data, model training and evaluation.

In this study, we propose methods for generating pseudo data for TS task by using word alignments which are generated by OTAlign [6]. Meanwhile, we also use the method of data augmentation for TS task. As for the model, we adopt the mT5-base model [7] which is a large-scale pre-trained model. By applying these methods, the final results exceed the baseline, demonstrating the feasibility of our proposed approaches.

## 2 Related Works

Since Yang et al. [3] provided a benchmark for TS. There have already been some achievements related to this task. Mao et al. [8] used the ΔLM as their backbone model. ΔLM is a pre-trained multilingual encode-decoder model, which outperforms various strong baselines on both natural language generation and translation tasks. For the training data, they construct the pseudo data with two different methods according to its constructing complexity. In their experiments, they find that accuracy indicator of TS can be helpful for efficient PE in practice. On the other hand, the main efforts of Zhang et al. [9] are paid on building the pseudo data. In addition to randomly mask the sub-segment in target reference, they use a quality estimation model to estimate the translation quality of words in translation output sentence and select the span with low confidence for masking. Then, 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.

From these works, we find that there are two main ways to improve the final performance of Translation Suggestion. The first method is to construct a more efficient model or

[1] They use fast-align [4] as one of the methods to generate pseudo data. In their experiments, they have confirmed that fast-align outperforms TERCOM [5]. Meanwhile, Arase et al. [6] have confirmed that OTAlign outperforms fast-align. In this paper, we assume that higher quality word alignment would contribute to generate higher quality pseudo data. Although we do not provide the details, the experimental results show that models pre-trained by pseudo data generated by OTAlign outperform models pre-trained by pseudo data generated by TERCOM.

apply some pre-trained models which is suitable for the Translation Suggestion task. Another method is to find well organized ways to generate the pseudo data as the amount of the golden corpus is limited.

## 3 Training Models for Translation Suggestion Task

The TS task can be described as a sequence input of the source language sentence $s$ concatenates the masked translation $m^{-w}$ ,and a sequence output of the target translation suggestion $t$. The input to the model is formatted as:

$$[s; < sep >; m^{-w}] \tag{1}$$

where $[;]$ means concatenation, and $< sep >$ is a special token used as a delimiter. Following the benchmark [3] method, we apply the pseudo data to pre-train the model. After pre-training, we fine-tune the model by utilizing the golden corpus of WeTS. For more information about TS, please refer to the appendix A.

## 4 Generating Pseudo Data for Translation Suggestion Task

Since the high costs and labor-consuming of creating the golden corpus, as well as the difficulty of training a model for Translation Suggestion task with scare amounts of golden corpus data, it becomes imperative to generate the pseudo corpus automatically. In order to achieve this goal, we firstly collect the parallel English-Chinese data from the United Nations Parallel Corpus v1.0 dataset.[2] For English-Chinese corpora, we remove sentences that are shorter than 50 words or longer than 200 words. After this operation, we finally get a size of around 10M corpora.Then we get the machine translation by feeding the English sentences of the cleaned corpus into a corresponding fully-trained NMT model. Finally, we can generate pseudo corpus data by utilizing the cleaned English-Chinese corpus and the Chinese machine translation.

### 4.1 Generating Pseudo Data by Random Mask

Since this method is described in [3], in this paper, we do not provide a comprehensive explanation. For more information, please refer to the appendix B.

### 4.2 Generating Pseudo Data by Word Alignments

Since the mismatch of distribution between the target sentence and the machine translation sentence is the potential problem of random mask method, we propose the approach of utilizing the monolingual word alignment to generate pseudo data for TS task. Then we can generate the pseudo data for TS task based on the alignment information.[3]

**Table 1** The number of generated pseudo data with three proposed approaches

| Methods | Number of original triples | Number of pseudo data after quality judgment | Number of pseudo data after placeholders concatenation |
|---|---|---|---|
| Randomly mask | 4,040,000 | 4,040,000 | — |
| OTAlign | 4,040,000 | 3,022,368 | 5,378,537 |

#### Prompting Alignment by OTAlign

Optimal Transport (OT) is a theory which aims to find the most effective method for transferring mass from one measure to another.The OT problem generates the OT mapping, which reveals the correspondences between two samples. While OT is commonly used as a distance metric between two measures, the main focus in alignment problems often lies on the OT mapping. For the reason mentioned above, Arase et al. [6] proposed a method for applying the Optimal Transport theory to generate the word alignments. Their experiments' results indicate that OT-base alignment methods are competitive against the state-of-the-arts designed for word alignment [10] [11].

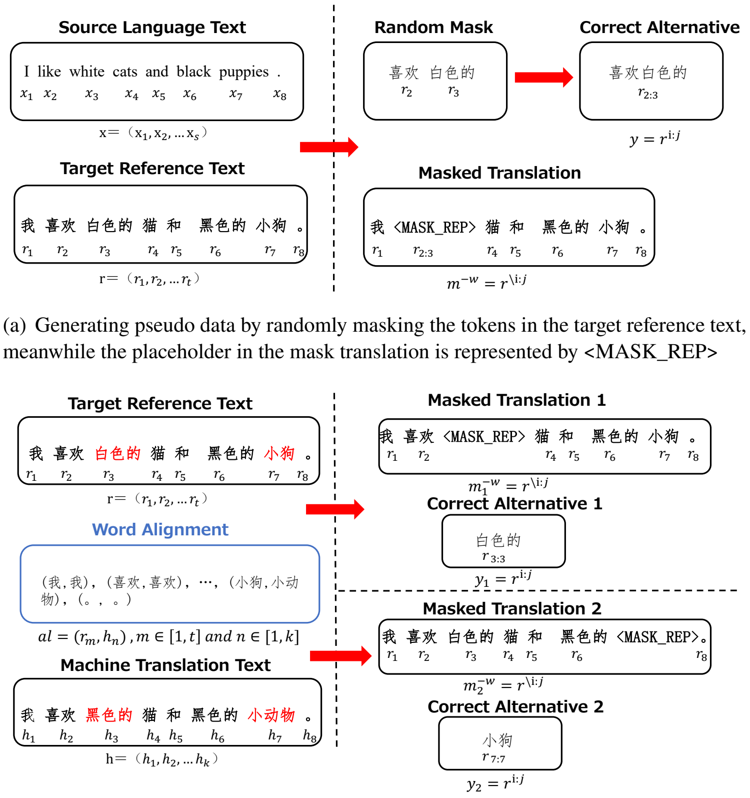
As shown in Figure 1(b), based on the word alignment generated by OTAlign[4], we could create pseudo data for the Translation Suggestion task by following rules:

- **Keep :** If there is an alignment between two tokens (one is from the target reference sentence, another is from the machine translation sentence), and they are completely identical, then this token is remained in the masked translation.
- **Replacement :** If there is an alignment between two tokens (one is from the target reference sentence, another is from the machine translation sentence), but they are not the same, then a placeholder is added to the masked translation, and the corresponding target reference token in the word alignment is added to the correct alternative.

[2] The details of United Nations Parallel Corpus v1.0 dataset can be found at https://conferences.unite.un.org/uncorpus.

[3] We test two different ways for generating word alignments, including TERCOM (Snover et al., 2006) and OTAlign (Arase et al., 2023), and the experimental results show that OTAlign performs better.

[4] Please refer to the appendix C for details of generating word alignments by OTAlign.

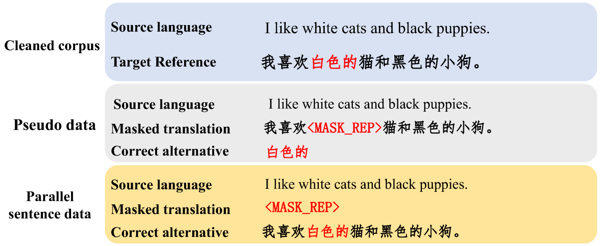

Source Language Text
I like white cats and black puppies .
$x_1$ $x_2$ $x_3$ $x_4$ $x_5$ $x_6$ $x_7$ $x_8$
x= （$x_1, x_2, \ldots x_s$）
Target Reference Text
我 喜欢 白色的 猫 和 黑色的 小狗 。
$r_1$ $r_2$ $r_3$ $r_4$ $r_5$ $r_6$ $r_7$ $r_8$
r= （$r_1, r_2, \ldots r_t$）
Random Mask
喜欢 白色的
$r_2$ $r_3$
Correct Alternative
喜欢白色的
$r_{2:3}$
$y = r^{i:j}$
Masked Translation
我 <MASK_REP> 猫 和 黑色的 小狗 。
$r_1$ $r_{2:3}$ $r_4$ $r_5$ $r_6$ $r_7$ $r_8$
$m^{-w} = r^{\setminus i:j}$

(a) Generating pseudo data by randomly masking the tokens in the target reference text, meanwhile the placeholder in the mask translation is represented by <MASK_REP>

Target Reference Text
我 喜欢 白色的 猫 和 黑色的 小狗 。
$r_1$ $r_2$ $r_3$ $r_4$ $r_5$ $r_6$ $r_7$ $r_8$
r= （$r_1, r_2, \ldots r_t$）
Word Alignment
（我，我），（喜欢，喜欢），…，（小狗，小动物），（。，。）
$al = (r_m, h_n)$ , $m \in [1, t]$ and $n \in [1, k]$
Machine Translation Text
我 喜欢 黑色的 猫 和 黑色的 小动物 。
$h_1$ $h_2$ $h_3$ $h_4$ $h_5$ $h_6$ $h_7$ $h_8$
h= （$h_1, h_2, \ldots h_k$）
Masked Translation 1
我 喜欢 <MASK_REP> 猫 和 黑色的 小狗 。
$r_1$ $r_2$ $r_4$ $r_5$ $r_6$ $r_7$ $r_8$
$m_1^{-w} = r^{\setminus i:j}$
Correct Alternative 1
白色的
$r_{3:3}$
$y_1 = r^{i:j}$
Masked Translation 2
我 喜欢 白色的 猫 和 黑色的 <MASK_REP>。
$r_1$ $r_2$ $r_3$ $r_4$ $r_5$ $r_6$ $r_8$
$m_2^{-w} = r^{\setminus i:j}$
Correct Alternative 2
小狗
$r_{7:7}$
$y_2 = r^{i:j}$

(b) Generating pseudo data by using the word alignment between the target reference text and MT, meanwhile the placeholder in the mask translation is represented by <MASK_REP>

**Figure 1** Two methods of generating pseudo data for Translation Suggestion task



| | | |
|---|---|---|
| Cleaned corpus | Source language | I like white cats and black puppies. |
| | Target Reference | 我喜欢白色的猫和黑色的小狗。 |
| Pseudo data | Source language | I like white cats and black puppies. |
| | Masked translation | 我喜欢<MASK_REP>猫和黑色的小狗。 |
| | Correct alternative | 白色的 |
| Parallel sentence data | Source language | I like white cats and black puppies. |
| | Masked translation | <MASK_REP> |
| | Correct alternative | 我喜欢白色的猫和黑色的小狗。 |

**Figure 2** The examples of the pseudo data and parallel sentence data constructed from the same source language sentence and target reference sentence pair

- **Insertion :** If there is an alignment, and the index of the token from the target reference is 2 or more greater than the index of the previous word alignment,, then we consider that tokens between these two indices need to be inserted. Then placeholders are added in the masked translation, and the corresponding tokens are added to the correct alternative.
- **Quality judgement :** After the four operations mentioned above, we need to assess the quality of the obtained $(m^{-w}, y)$ pairs. In order to do this, we can evaluate the number of placeholders in the mask translation. If the number of placeholders exceeds a certain threshold, then the quality of that masked translation is considered poor and should be filtered out.
- **Placeholders concatenation:** In Equation 1, we can observe that the model input contains only 1 placeholder, but after the operations described above, the current sentence may have multiple placeholders. Therefore, we need to concatenate adjacent placeholders to ensure that the number of placeholders remains consistent with the number of one defined in the TS model input. This may result in generating multiple pseudo data from a single triple of $(r, h, al)$ (Figure 1(b)) .

Based on the rules mentioned above, we ultimately generate the pseudo data by using OTAlign. The numbers of pseudo data after the operation of **quality judgement** and **placeholders concatenation** are shown in Table 1.

## 4.3 Data Augmentation

Xiao et al. [12] propose a method which applies the Data Augmentation (DA) technique into the Bilingual Text-

Infilling task and the final experiment result demonstrates the feasibility of DA. Since the Bilingual Text-Infilling task and the TS task are quite similar (both aim to provide correct alternatives for masked (or missing) segments in the corresponding sentences by the models.), we also propose to apply DA into the TS task. Briefly speaking, for the parallel data $(x, m^{-w})$ and correct alternatives $y$ in Equation 2, we construct the parallel sentence data $(x, \bar{m}^{-w})$ and $\bar{y}$, where $\bar{m}^{-w} = \emptyset$ and $\bar{y} = m^{-w}$. In other words, $\bar{m}^{-w}$ means no tokens are provided in $m^{-w}$. Then we combine the pseudo data and the parallel sentence data to train the TS models in our experiments.

# 5 Experiments

## 5.1 Translation Suggestion Task Settings

For training, we apply the two-state training pipline, where we pre-train the model on the pseudo data in the first stage, and then fine-tune the model on the golden WeTS[5] corpus in the second stage. To compare the performance of the generated pseudo data, we consider the model pre-trained with pseudo data generated by random mask as the baseline model. We divided the existing two types of pseudo data into subsets of 1, 2 and 4 million each. Meanwhile, we add an equal amount of parallel sentence data to each subsets of pseudo data. By comparing the performance of models trained on the same amount of pseudo data and parallel sentence data, we verify the effectiveness of data augmentation for the TS task. Other experimental settings are in appendix D.

## 5.2 Results

As shown in Table 2, as the amount of generated pseudo data increases, the improvement in model's performance is noticeable but not extremely significantly. Compared to models pre-trained using an equal amount of pseudo data generated by randomly mask (Baseline: mT5_base_1M, mT5_base_2M, mT5_base_4M), the performance of models only pre-trained with OTAlign-generated pseudo data is worse than baseline models. However, after fine-tuning, models pre-trained with pseudo data generated by OTAlign's word alignments outperform the baseline. On the other hand, we also want to confirm whether data augmentation methods can enhance the performance of the mT5-base model in the TS task.As shown in Table 2, after adding an equal amount of parallel sentence data, models only pre-trained with OTAlign-generated pseudo data and parallel sentence data don't achieve higher BLEU scores compared to the baseline models with the same amount of data. However, after fine-tuning, models with OTAlign-generated pseudo data and parallel sentence data achieve higher scores than baseline models. Especially, when the mT5-base model is pre-trained with 4M OTAlign-generated pseudo data and 4M parallel sentence data, it achieves a BLEU score of 22.3 after fine-tuning.[6]

**Table 2** The evaluation result of models trained by two types of generated pseudo data. ("*" shows the significant ($p < 0.05$) BLEU scores of models' outputs which are pre-trained and fine-tuned with generated pseudo data in the same amount, compared with mT5_base_xM)

| Models | w/o fine-tuning | w/ fine-tuning |
|---|---|---|
| mT5_base_1M | 10.3 | 14.6 |
| mT5_base_2M | 11.0 | 16.0 |
| mT5_base_4M | 11.3 | 16.9 |
| mT5_otalign_1M | 9.2 | 15.5* |
| mT5_otalign_2M | 10.1 | 16.5* |
| mT5_otalign_4M | 10.8 | 18.1* |
| mT5_base_mix_2M | 11.7 | 16.9* |
| mT5_base_mix_4M | 12.2 | 17.4 |
| mT5_base_mix_8M | 14.1 | 21.8 |
| mT5_otalign_mix_2M | 9.5 | 16.2 |
| mT5_otalign_mix_4M | 11.7 | 18.3* |
| mT5_otalign_mix_8M | 11.9 | 22.3* |

# 6 Conclusion

In order to improve the performance of TS, we have attempted to use OTAlign for generating pseudo data. The experiment results indicate that pseudo data generated using OTAlign is effective for pre-training models, thereby improving the effectiveness of TS. Meanwhile, we show that Data Augmentation technology for TS, the results show that the parallel sentence data can significantly improve the effectiveness of TS.

[5] You can download this dataset from https://openai.weixin.qq.com/translate/description.

[6] Besides these experiments, we used the 4M pseudo-data generated by random mask, 4M OTAlign generated pseudo-data, and about 12M of parallel sentence data (approximately 20M in total) for the pre-training of the mT5-base model. After fine-tuning, the model's BLEU score increased to 26.4. This indicates that continuously increasing the pseudo data volume according to our proposed method can further enhance the model's performance.

## References

[1] D. Lee, J. Ahn, H. Park, and J. Jo. IntelliCAT: Intelligent machine translation post-editing with quality estimation and translation suggestion. In **Proc. 59th ACL and 11th IJCNLP**, pp. 11–19, 2021.

[2] Q. Wang, J. Zhang, L. Liu, G. Huang, and C. Zong. Touch editing: A flexible one-time interaction approach for translation. In **Proc. 1st AACL and 10th IJCNLP**, pp. 1–11, 2020.

[3] Z. Yang, F. Meng, Y. Zhang, E. Li, and J. Zhou. WeTS: A benchmark for translation suggestion. In **Proc. EMNLP**, pp. 5278–5290, 2022.

[4] C. Dyer, V. Chahuneau, and N. Smith. A simple, fast, and effective reparameterization of IBM model 2. In **Proc. NAACL-HLT**, pp. 644–648, 2013.

[5] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In **Proc. AMTA**, Vol. 200, 2006.

[6] Y. Arase, H. Bao, and S. Yokoi. Unbalanced optimal transport for unbalanced word alignment. In **Proc. 61st ACL**, pp. 3966–3986, 2023.

[7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proc. NAACL**, pp. 483–498, 2021.

[8] H. Mao, W. Zhang, J. Cai, and J. Cheng. Transn's submissions to the WMT22 translation suggestion task. In **Proc. 7th WMT**, pp. 1205–1210, 2022.

[9] H. Zhang, S. Lai, S. Zhang, H. Huang, Y. Chen, J. Xu, and J. Liu. Improved data augmentation for translation suggestion. In **Proc. 7th WMT**, pp. 1211–1216, 2022.

[10] W. Lan, C. Jiang, and W. Xu. Neural semi-markov CRF for monolingual word alignment. In **Proc. 59th ACL and 11th IJCNLP**, pp. 6815–6828, 2021.

[11] M. Nagata, K. Chousa, and M. Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In **Proc. EMNLP**, pp. 555–565, 2020.

[12] Y. Xiao, G. Huang L. Liu, Q. Cui, S. Huang, S. Shi, and J. Chen. BiTIIMT: A bilingual text-infilling method for interactive machine translation. In **Proc. 60th ACL**, pp. 1958–1969, 2022.

[13] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. **Computational Linguistics**, Vol. 19, pp. 263–311, 1993.

[14] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proc. EMNLP**, pp. 66–71, 2018.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

# A Translation Suggestion

Translation suggestion is an important tool for post-editing. To further reduce the post-editing time, researchers [1] [2] propose to apply TS into post-editing, where TS provides the sub-segment suggestions for the annotated incorrect word spans in the results of machine translation, and their extensive experiments show that TS can substantially reduce translators’ cognitive loads and the post-editing time.

To enhance the research in the TS area. Yang et al. [3] propose a benchmark for TS. In their proposal, they consider TS task as this: Given the source sentence $x = (x_1, ..., x_s)$, the translation sentence $m = (m_1, ...m_t)$, the incorrect words or spans $w = m_{i:j}$ where $1 \leq i \leq j \leq t$ , and the correct alternative $y$ for $w$, the task of TS is optimized to maximize the conditional probability of $y$ as follows:

$$P(y|x, m^{-w}, \theta) \tag{2}$$

# B Generating Pseudo Data by Random Mask

Random mask on the golden parallel corpus is the most straightforward approach to generate pseudo data for TS. Give the sentence pair $(x, r)$ in the cleaned English-Chinese corpus, where $x$ is the source language sentence and $r$ is the corresponding target reference sentence.We denote $r^{\backslash i:j}$ as a masked version of $r$, which means a portion of $r$ from position $i$ to $j$ is replaced with a placeholder (<MASK_REP>). The $r^{i:j}$ denotes the portion of $r$ from position $i$ to $j$. We consider $r^{i:j}$ and $r^{\backslash i:j}$ as the correct alternative ($y$ in Equation 1) and masked translation ($m^{-w}$ in Equation 1) respectively. In this approach, the mask translation in each example is part of the target reference sentence $r$. (Figure 1(a)).

Specifically, we randomly select one token in target reference sentence with a certain probability[7] as the first token of the incorrect span. Then, in order to generate the correct alternative and the masked translation, we set another certain probability to decide the length of the incorrect span which is started from the first token we selected. Finally, the correct alternative and the masked translation can be generated.

# C Generating word alignments by OTAlign

As for the OT problem, the inputs are a cost function and a pairs of measures. Assuming that the word embeddings of target reference sentence $r$ and the corresponding MT sentence $h$ are at hand. A cost refers to a dissimilarity between $r_t$ and $h_k$ computed by a distance metric such as Euclidean and cosine distances. The cost matrix $\mathbf{C}$ summarises the cost of any word pairs: $C_{t,k} = c(r_t, h_k)$. A measure means a weight each word has. The concept of **measure** corresponds to the notion of **fertility** introduced in IBM Model 3 [13], which defines how many target reference(MT) words a MT word(target reference) word can align. The mass of words in $r$ and $h$ is represented as arbitrary measures $a \in \mathbb{R}_+^t$ and $b \in \mathbb{R}_+^k$. Finally, an alignment matrix $\mathbf{P}$ is computed to minimize the sum of alignment costs under the cost matrix $\mathbf{C}$:

$$L_C(\boldsymbol{a}, \boldsymbol{b}) := \min_{P \in U(a,b)} \boldsymbol{<C, P>}, \tag{3}$$

where $\boldsymbol{<C, P>} := \sum_{t,k} C_{t,k} P_{t,k}$ .With this formulation, we can seek the most reliable word alignment matrix $\mathbf{P}$.

# D Experimental Settings

As for the mT5-base model[8], the pseudo data is jointly tokenized into sub-word units with SentencePiece. [14] During pre-training, the batch size is set as 32, and the learning rate is set to 3e-5. During fine-tuning, the batch size is also set as 32, and the learning rate is set to 1e-5. Especially, we pre-train the mT5-base model using the generated pseudo data and the parallel sentence data for one epoch and fine-tune the model for ten epochs. All the other configurations are remained unchanged as the setting of the benchmark [3]. In order to evaluate the models, we utilize the official evaluation tool scarebleu[9] to evaluate the model’s output(The translation suggestion segment in Chinese) against the reference correct alternatives. In this paper, because we only select the English-Chinese direction, the BLEU [15]score is calculated on the characters with the default tokenizer for Chinese. For experiments related to the mT5-base model we conduct them on an NVIDIA A6000 RTX (48GB) with CUDA 11.3.

[7] We set the blank probability to 0.15 because we observed such a probability in the golden corpus data. In other words, 0.15 is a value that is close to the objective distribution of mask.

[8] We directly utilize the mT5-base model and SentencePiece model provided in https://huggingface.co/google/mt5-base.

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