HW-TSC Translation Systems for the WMT22 Biomedical Translation Task
Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen,
Zongyao Li, Hengchao Shang, Shaofen Li, Ming Zhu, Yuanchang Luo, Yuhao Xie,
Miaomiao Ma, Ting Zhu, Lizhi Lei, Song Peng, Hao Yang, Ying Qin
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,yangjinlong7,raozhiqiang,yuzhengzhe,weidaimeng,chenxiaoyu35,
liangzhongzhao,liangshaofen,zhuming47,luoyuanchang,xieyuhao2,
miaomiao,ma,zhuting20,leilizhi,pengsong2,yanghao30,qinying}@huawei.com

Abstract
This paper describes the translation systems trained by Huawei translation services center (HW-TSC) for the WMT22 biomedical translation task in five language pairs: English ↔ German (en ↔ de), English ↔ French (en ↔ fr), English ↔ Chinese (en ↔ zh), English ↔ Russian (en ↔ ru) and Spanish → English (es → en). Our primary systems are built on deep Transformer with a large filter size. We also utilize R-Drop, data diversification, forward translation, back translation, data selection, finetuning and ensemble to improve the system performance. According to the official evaluation results in OCELoT\textsuperscript{1} or CodaLab\textsuperscript{2}, our unconstrained systems in en → de, de → en, en → fr, fr → en, en → zh and es → en (clinical terminology sub-track) get the highest BLEU scores among all submissions for the WMT22 biomedical translation task.

1 Introduction
Machine translation (MT) refers to the automatic translation of text from one language to another, and the biomedical translation task aims to evaluate the performance of MT systems in the biomedical domain. In this year’s biomedical translation task, our team (HW-TSC) participates in five language pairs, including en ↔ de, en ↔ fr, en ↔ zh, en ↔ ru and es → en (clinical terminology sub-track).

Since the size of in-domain (ID) data is limited, we first use a large amount of out-of-domain (OOD) data to train our baseline neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) system, which is a deep transformer model (Dou et al., 2018; Li et al., 2019) leveraging R-Drop (Wu et al., 2021) training strategy. We then use the collected ID data (except the data from medical database) to further train the NMT model for domain transfer. To better use the limited ID training data, we employ data selection to extract ID data from OOD data, in addition to basic data augmentation strategies including data diversity, forward translation and back translation. Finally, we use finetuning (Dakwale and Monz, 2017) and model ensemble (Wang et al., 2020b) to further improve model performance in the biomedical domain.

This paper is structured as follows: we describe data size and data pre-processing methods in section 2; the model structure and training methods in section 3; final results in section 4; and conclusion in section 5.

2 Dataset
2.1 Data volume
The data size for each language pair for the WMT22 biomedical translation task is shown in Table 1. The OOD bilingual data, used to train our baseline model, comes from the WMT general MT task and our internal corpus; while the ID bilingual and monolingual data, used for transferring the domain (Yang et al., 2021), come from Biomedical Translation, UFAL Medical Corpus and our internal corpus. As there is no ID monolingual data, we use the OOD monolingual instead.

2.2 Data Pre-processing
The data pre-processing process is as follows:

- Remove duplicate sentences (Khayrallah and Koehn, 2018; Ott et al., 2018).
- Remove sentences with mismatched parentheses and quotation marks.
- Filter out sentences of which punctuation percentage exceeds 0.4.
- Filter out sentences with the character-to-word ratio greater than 12 or less than 1.5.

\textsuperscript{1}https://ocelot-wmt22.mteval.org
\textsuperscript{2}https://codalab.lisn.upsaclay.fr/competitions/6696#results
Table 1: The data size for each language pair in the WMT22 Biomedical Translation Task

<table>
<thead>
<tr>
<th></th>
<th>bilingual</th>
<th>monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en↔de</td>
<td>en↔fr</td>
</tr>
<tr>
<td>OOD</td>
<td>200M</td>
<td>600M</td>
</tr>
<tr>
<td>ID</td>
<td>2.75M</td>
<td>6.05M</td>
</tr>
</tbody>
</table>

• Filter out sentences with more than 150 words.
• Apply langid (Joulin et al., 2017, 2016) to filter sentences in other languages.
• Use fast-align (Dyer et al., 2013) to filter out sentence pairs that are poorly aligned.

It should be noted that for en↔de, en↔fr, en↔ru and es→en translation task, we adopt joint SentencePiece model (SPM) (Kudo and Richardson, 2018; Kudo, 2018) for word segmentation, with a vocabulary of 32k. As for en↔zh translation task, we use Jieba tokenizer to pre-segment Chinese sentences, and Moses tokenizer (Koehn et al., 2007) to pre-segment English sentences. Then we use joint Byte Pair Encoding (BPE) (Sennrich et al., 2016) to perform subword segmentation on Chinese and English sentences. The vocabulary size of BPE is also set to 32k.

3 System Overview

3.1 Model

Transformer (Vaswani et al., 2017), as the current mainstream architecture for NMT, adopts a fully self-attention mechanism, which can realize algorithm parallelism, speed up model training, and improve model performance. Deep Transformer, as an improvement of Transformer, increases the number of encoder layers and uses pre-layer-normalization to further improve model performance. Therefore, for all language pairs, we adopt the Deep Transformer (Wei et al., 2021) model architecture: Based on the Transformer-big model architecture, our Deep Transformer model features pre-layer-normalization, 25-layer encoder, 6-layer decoder, 16-head self-attention, 1024-dimension word embedding and 4096-dimension hidden state.

3.2 R-Drop

Dropout (Srivastava et al., 2014) is a powerful and widely used technique for regularizing deep neural networks. Though it can help improve training effectiveness, the randomness introduced by dropouts may lead to inconsistencies between training and inference. R-Drop (Wu et al., 2021) forces the output distributions of different sub-models generated by dropout be consistent with each other. Therefore, we use R-Drop to augment the baseline model for each task and reduce inconsistencies between training and inference.

3.3 Data Diversification

Data diversification (Nguyen et al., 2020) is a simple and effective strategy to improve the performance of NMT. It uses predictions from multiple forward and backward models, and combines the results with the original data to train the final NMT model. The method does not require additional monolingual data and is applicable to all NMT models. It is more efficient than knowledge distillation (Wang et al., 2021) and dual learning (He et al., 2016). In our en↔de, en↔fr, en↔zh and en↔ru translation tasks, we use only a forward model and a backward model to generate synthetic data, and then mix the synthetic data with the bilingual data for NMT model training.

3.4 Forward Translation

Forward translation (Wu et al., 2019), also known as self-training (Imamura and Sumita, 2018), refers to using a forward NMT model to translate source-side monolingual data to generate synthetic bilingual data, which is then used to expand the training data size. Forward translation usually relies on beam search (Freitag and Al-Onaizan, 2017) decoding to generate synthetic data. Therefore, we adopt the forward translation method based on beam search decoding.

3.5 Back Translation

Back translation (Sennrich et al., 2015; Edunov et al., 2018) refers to translating the target monolingual data back to the source language, and then using the synthetic data to increase the training data size. This method has been proven effective in improving the NMT model performance. There are many back translation methods, among which sampling (Graça et al., 2019), noise (Edunov et al.,

\[^3\]https://github.com/fxsjy/jieba
or tagged (Caswell et al.) back-translation methods work better. In the scenario where forward translation and back translation are used in combination (Wu et al., 2019), the improvement effect brought by sampling back translation is more significant. In our translation task, we adopt sampling back translation method.

### 3.6 Data Selection

Data selection (van der Wees et al., 2017) is a data augmentation method that we use to select ID bilingual data from OOD bilingual data. Inspired by the domain feature calculation in curriculum learning (Wang et al., 2020a), we use an ID NMT model and an OOD NMT model to calculate the decoding probability of OOD bilingual data. The bilingual data of which ID decoding probability is higher than OOD decoding probability can be selected as additional ID data. The data selection process is also shown in Algorithm 1:

#### Algorithm 1: Data selection process

**Input**: ID NMT model $\theta_I$, OOD NMT model $\theta_O$ and OOD bilingual data set $D_O$.

**Output**: ID bilingual data set $D_I$.

1. for each sentence pair $(x, y) \in D_O$ do
   1. // $x$ is the source sentence, $y$ is the target sentence.
   2. $\text{score} = \frac{\log P(y|x; \theta_I) - \log P(y|x; \theta_O)}{|y|}$
   3. if $\text{score} > 0$ then
      4. add $(x, y)$ to $D_I$
   5. end
2. end

### 3.7 Finetuning

Finetuning (Dakwale and Monz, 2017) is a way to achieve domain transfer. In our translation task, we adopt a two-stage finetuning strategy. In the first stage, we use ID bilingual data to continue training the OOD NMT model, and then use the data augmentation strategy mentioned above to improve the model performance. In the second stage, we use the development set and synthetic data generated from the source-side text in the test set to finetune the ID model for more fine-grained domain transfer.

### 3.8 Ensemble

Ensemble (Wang et al., 2020b) is a widely used method to integrate different models for better performance. It is worth noting that when using ensemble, increasing the number of models does not always lead to better performance, and sometimes even causes performance deterioration. Therefore, for each track, we train four models on the same data, and go through all combinations of models to choose the one that performs best on the development set. This is also the model selection strategy (Yang et al., 2021) we use in the WMT21 biomedical translation task.

### 4 Experimental Result

During the training phase, we use Pytorch-based Fairseq$^4$ (Ott et al., 2019) open-source framework, and use deep Transformer model architecture as our benchmark system. Each model is trained using 8 GPUs with a batch size of 2048. The update frequency is 4 and the learning rate is $5e^{-4}$, the label smoothing rate ($\text{Szegedy et al., 2016}$) is 0.1, the warm-up steps is 4000, and the dropout is 0.3. Adam optimizer (Kingma and Ba, 2015) with $\beta_1=0.9$ and $\beta_2=0.98$ is also used. Furthermore, we use reg_label_smoothed_cross_entropy as the loss function and set reg-alpha to 5 when applying R-Drop (Wu et al., 2021) training strategy. In the evaluation phase, we use Marian$^5$ (Junczys-Dowmunt et al., 2018) for decoding and then calculate the sacrebleu$^6$ (Post, 2018) on the WMT21 OK-aligned biomedical test set to measure the performance of each model.

#### 4.1 en$\rightarrow$de

For en$\rightarrow$de track, Table 2 shows the results of using the methods mentioned above to improve the model performance. The results show that continuing training with ID bilingual data on the basis of an OOD baseline improves en$\rightarrow$de translation performance by 1.6 BLEU, but has little effect on the de$\rightarrow$en track, with an increase of only 0.1 BLEU. Data selection significantly improves en$\rightarrow$de translation performance by 0.9-1.2 BLEU. In addition, other training strategies also bring small performance improvements.

#### 4.2 en$\rightarrow$fr

Table 3 shows the results of en$\rightarrow$fr model. The results show that data diversity brings the greatest improvement to translation of both directions.

---

4https://github.com/facebookresearch/fairseq
5https://github.com/marian-nmt/marian
6https://github.com/mjpost/sacrebleu
Table 2: BLEU scores of en→de on the WMT21 OK-aligned biomedical test set.

<table>
<thead>
<tr>
<th>System</th>
<th>en→de</th>
<th>de→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD R-Drop baseline</td>
<td>27.3</td>
<td>39.7</td>
</tr>
<tr>
<td>+ ID bilingual data continue training</td>
<td>28.9</td>
<td>39.8</td>
</tr>
<tr>
<td>+ data diversification</td>
<td>29.0</td>
<td>40.1</td>
</tr>
<tr>
<td>+ forward translation &amp; back translation</td>
<td>29.4</td>
<td>41.3</td>
</tr>
<tr>
<td>+ data selection</td>
<td>30.3</td>
<td>42.5</td>
</tr>
<tr>
<td>+ dev set &amp; synthetic test set finetuning</td>
<td>30.8</td>
<td>42.9</td>
</tr>
<tr>
<td>+ ensemble</td>
<td><strong>31.0</strong></td>
<td><strong>43.2</strong></td>
</tr>
</tbody>
</table>

Table 3: BLEU scores of en→fr on the WMT21 OK-aligned biomedical test set.

<table>
<thead>
<tr>
<th>System</th>
<th>en→fr</th>
<th>fr→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD R-Drop baseline</td>
<td>44.8</td>
<td>46.1</td>
</tr>
<tr>
<td>+ ID bilingual data continue training</td>
<td>45.3</td>
<td>46.3</td>
</tr>
<tr>
<td>+ data diversification</td>
<td>46.0</td>
<td>47.6</td>
</tr>
<tr>
<td>+ forward translation &amp; back translation</td>
<td>46.3</td>
<td>47.7</td>
</tr>
<tr>
<td>+ data selection</td>
<td>-</td>
<td>47.8</td>
</tr>
<tr>
<td>+ dev set &amp; synthetic test set finetuning</td>
<td>46.8</td>
<td>48.4</td>
</tr>
<tr>
<td>+ ensemble</td>
<td><strong>46.9</strong></td>
<td><strong>48.6</strong></td>
</tr>
</tbody>
</table>

Table 4: BLEU scores of en→zh on the WMT21 OK-aligned biomedical test set.

<table>
<thead>
<tr>
<th>System</th>
<th>en→zh</th>
<th>zh→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD R-Drop baseline</td>
<td>38.5</td>
<td>32.1</td>
</tr>
<tr>
<td>+ ID bilingual data continue training</td>
<td>41.4</td>
<td>35.0</td>
</tr>
<tr>
<td>+ data diversification</td>
<td>42.5</td>
<td>36.4</td>
</tr>
<tr>
<td>+ forward translation &amp; back translation</td>
<td>42.7</td>
<td>37.3</td>
</tr>
<tr>
<td>+ data selection</td>
<td>42.8</td>
<td>-</td>
</tr>
<tr>
<td>+ dev set &amp; synthetic test set finetuning</td>
<td>43.0</td>
<td>38.7</td>
</tr>
<tr>
<td>+ ensemble</td>
<td><strong>43.1</strong></td>
<td><strong>39.3</strong></td>
</tr>
</tbody>
</table>

Table 5: BLEU scores of en→ru on the WMT21 OK-aligned biomedical test set.

<table>
<thead>
<tr>
<th>System</th>
<th>en→ru</th>
<th>ru→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD R-Drop baseline</td>
<td>35.4</td>
<td>46.8</td>
</tr>
<tr>
<td>+ ID bilingual data continue training</td>
<td>41.0</td>
<td>48.9</td>
</tr>
<tr>
<td>+ data diversification</td>
<td>41.7</td>
<td>50.3</td>
</tr>
<tr>
<td>+ forward translation &amp; back translation</td>
<td>42.3</td>
<td>50.4</td>
</tr>
<tr>
<td>+ data selection</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+ dev set &amp; synthetic test set finetuning</td>
<td>42.4</td>
<td>50.9</td>
</tr>
<tr>
<td>+ ensemble</td>
<td><strong>42.5</strong></td>
<td><strong>51.1</strong></td>
</tr>
</tbody>
</table>

4.3 en→zh

For en→zh track, continuing training with ID bilingual data on the basis of an ODD baseline, as well as data diversity, bring the greatest impact on the model performance, while data selection has the least impact. In addition, the methods such as forward translation & back translation, dev set & synthetic test set finetuning and ensemble have little improvement on en→zh translation, but have a great improvement on zh→en translation. The detailed results of en→zh translation are shown in Table 4.

4.4 en→ru

As shown in Table 5, for the en→ru track, the results are similar to en→zh translation task. Continuing training with ID bilingual data and data diversity have the greatest impact on model performance, while data selection does not lead to performance improvement. In addition, the performance improvements brought by other methods are also relatively limited.

4.5 es→en

We also participate in the es→en clinical terminology sub-track (ClinSpEn-CT) this year. The sample set contains 7,000 terms that are extracted from medical literature and clinical records, with a particular focus on diseases, symptoms, findings, etc. The translations are generated and revised by professional medical translators. We extract 1000 sentences from the sample set as the dev set.

The results are shown in Table 6. All chrF and BLEU scores are calculated on this dev set. Unlike other experiments above, for es→en clinical terminology sub-task, we abandon forward translation method for the sake of maintaining terminology accuracy. Instead, we perform two rounds of back translation using monolingual English ID data. Finally, we finetune the model with 6000 bilingual terms, which results in a significant improvement on the dev set.

4.6 Results In OCELoT Or CodaLab

The BLEU scores of our submissions to the WMT22 Biomedical Translation Task on OCELoT and CodaLab (ClinSpEn-CT) are shown in Table 6.

<table>
<thead>
<tr>
<th>System</th>
<th>chrF</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD R-Drop baseline</td>
<td>0.76</td>
<td>49.5</td>
</tr>
<tr>
<td>+ ID bilingual data continue training</td>
<td>0.77</td>
<td>50.7</td>
</tr>
<tr>
<td>+ back translation</td>
<td>0.79</td>
<td>53.4</td>
</tr>
<tr>
<td>+ 2nd round back translation</td>
<td>0.79</td>
<td>54.1</td>
</tr>
<tr>
<td>+ 6000 bilingual terms finetuning</td>
<td>0.82</td>
<td>56.7</td>
</tr>
<tr>
<td>+ ensemble</td>
<td><strong>0.82</strong></td>
<td><strong>57.2</strong></td>
</tr>
</tbody>
</table>

Table 6: chrF (Popović, 2015) and BLEU scores of es→en on the WMT22 biomedical ClinSpEn-CT 1000 sample set.
7, where our submitted systems achieve the highest BLEU scores in six language directions of the WMT22 biomedical translation task. In conclusion, from the results on the WMT21 OK-aligned biomedical test set, continuing training with ID bilingual data, data diversity, forward translation and back translation have great impacts on the NMT model performance. When the OOD bilingual data contains a certain amount of ID bilingual, the data selection method can also achieve a good boost effect. In addition, dev set & synthetic test set finetuning and ensemble can lead to further performance gains.

5 Conclusion

This paper presents our translation system for the WMT22 en → de, en → fr, en → zh, en → ru and es → en biomedical translation task. During the experiment, we use R-Drop and ID bilingual data finetuning methods to build our ID translation system, and then use data diversity, forward translation, back translation and data selection methods to expand the size of training data for training a better system. We also adopt finetuning and ensemble to further improve the system performance. According to the official evaluation results in OCELOt or CodaLab, our submitted systems achieve the highest BLEU scores in six language directions of the WMT22 biomedical translation task.

References


<table>
<thead>
<tr>
<th>en→de</th>
<th>de→en</th>
<th>en→fr</th>
<th>fr→en</th>
<th>en→zh</th>
<th>zh→en</th>
<th>en→ru</th>
<th>ru→en</th>
<th>es→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>38.7</td>
<td>45.6</td>
<td>38.8</td>
<td>48.6</td>
<td>49.9</td>
<td>43.0</td>
<td>43.3</td>
<td>50.3</td>
<td>41.57</td>
</tr>
</tbody>
</table>

Table 7: BLEU scores of our submission systems on WMT22 Biomedical Translation Task on OCELOt or CodaLab, where the highest BLEU scores among all submissions are bolded.


Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In \textit{EMNLP (Demonstration)}.


