TSMind: Alibaba and Soochow University's Submission to the WMT22 Translation Suggestion Task

Xin Ge[1]*, **Ke Wang**[1]*, **Jiayi Wang**[1], **Nini Xiao**[1,2], **Xiangyu Duan**[2], **Yu Zhao**[1], **Yuqi Zhang**[1]†
[1]Alibaba Group Inc. [2]Soochow University
{shiyi.gx,moyu.wk,joanne.wjy}@alibaba-inc.com
{nnxiaonnxiao,xiangyuduan}@suda.edu.cn, {kongyu, chenwei.zyq}@alibaba-inc.com

Abstract

This paper describes the joint submission of Alibaba and Soochow University, TSMind, to the WMT 2022 Shared Task on Translation Suggestion (TS). We participate in the English ↔ German and English ↔ Chinese tasks. Basically, we utilize the model paradigm fine-tuning on the downstream tasks based on large-scale pre-trained models, which has recently achieved great success. We choose FAIR's WMT19 English ↔ German news translation system and MBART50 for English ↔ Chinese as our pre-trained models. Considering the task's condition of limited use of training data, we follow the data augmentation strategies proposed by Yang et al. (2021) to boost our TS model performance. The difference is that we further involve the dual conditional cross-entropy model and GPT-2 language model to filter augmented data. The leader board finally shows that our submissions are ranked first in three of four language directions in the Naive TS task of the WMT22 Translation Suggestion task.

1 Introduction

Computer-aided translation (CAT) (Barrachina et al., 2009; Green et al., 2014, 2015; Knowles and Koehn, 2016) has become more and more popular to help increase the quality of machine translation (Lopez, 2008; Koehn, 2009) result. It also improves the efficiency of translators by combining the results of machine translation and the content edited by translators in the process of translation or post-editing (Bowker, 2002; Lengyel and Ugray, 2004; Bowker and Fisher, 2010; Bowker, 2014; Chatterjee, 2019).

Post-editing based on machine translation is typical in CAT. Recent works (Domingo et al., 2016; González-Rubio et al., 2016; Peris et al., 2017) propose interactive protocols and algorithms so that humans and machines can collaborate during translation, and machines can automatically provide feedback on humans' edits. One interesting mode is Translation Suggestion (TS) (Yang et al., 2021), which offers alternatives for specific spans of words in the generated machine translation. It will be convenient if the model refines translation results in those specified locations with potential translation errors. Yang et al. (2021) released a benchmark dataset for TS, *WeTS*, which is one of the shared tasks in WMT22. At the same time, they proposed an end-to-end Transformer-like model for TS as the benchmark system.

However, the lack of many labeled TS data limits the training of a large Transformer model to some extent. Though Yang et al. (2021) have tried to utilize XLM-Roberta (Conneau et al., 2019) to initialize the encoder of the Transformer, the decoder has to be trained from scratch, which leads to relatively low BLEU scores for some specific TS spans. We investigate the potential of other encoder-decoder pre-trained models by experiments to see if there is still room for improvement. Finally, we have found that pre-trained Transformer NMT models could be suitable choices to be fine-tuned with the limited size of TS data. In addition, we applied similar data augmentation strategies proposed in Yang et al. (2021), but use the well-trained alignment models between source and target languages from Lu et al. (2020) to filter out high-quality augmented data. Our submissions are ranked first in three of four language directions in the WMT22 Translation Suggestion task.

2 The Model

We train a simple end-to-end Transformer model for each language pair to generate the translation suggestion candidates. The source sentence and the masked translation, in which an incorrect span requiring an alternative has been replaced with a special mask tokens in advance, are concatenated with a special separation token *[SEP]*. Afterward,

[*] indicates equal contribution.
[†] indicates the corresponding author.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1198 - 1204
December 7-8, 2022 ©2022 Association for Computational Linguistics

Symbol	Definition
$\mathbf{x}$	Sentence in source language
$\mathbf{y}$	Machine translation result of $\mathbf{x}$
$\mathbf{r}$	Reference sentence $\mathbf{x}$
$\mathbf{x}^i$	The i-th token of x
$\|\mathbf{x}\|$	Length of $\mathbf{x}$, i.e. the number of tokens in $\mathbf{x}$
$\mathbf{x}^{i:j}$	The fragment of $\mathbf{x}$ from position i to j
$\mathbf{x}^{\neg i:j}$	The masked version of $\mathbf{x}$, in which tokens at the position from i to j of x is replaced with a mask token.
$\hat{\mathbf{p}}$	All aligned-phrase pair between $\mathbf{y}$ and $\mathbf{r}$, pair look likes $(\mathbf{y}^{i:j}, r^{a:b})$
$\hat{\mathbf{y}}$	Replace $\mathbf{y}^{i:j}$ with $\mathbf{r}^{a:b}$ in $\mathbf{y}$, and get another new sentence $\hat{y}$

Table 1: Notations

	WMT22	Filter Length	Filter Quality
en-zh	23.2M	9.78M	6.9M
en-de	30.0M	12.73M	8.18M

Table 2: Number of parallel samples remained after filtering by length and cross-entropy quality score (Lu et al., 2020).

we feed the concatenated sequence as input of the Transformer encoder and the translation suggestion needs to be generated by the Transformer decoder. The model is trained in the same way of a normal translation model.

Considering that the TS task also relies on alignments of hidden representations between the source and the target language, a well-trained translation model can be a good starting point for TS model training. The weights of our model are initialized with a pre-trained Transformer NMT model. Then, a two-phase training pipeline is applied. In the first phase, the model is trained with pseudo corpus derived from data augmentation described in Section 3. In the second phase, we fine-tune the model with the real TS train data released by the organizers.

3 Data Augmentation

We follow the data augmentation methods provided by (Yang et al., 2021) to generate three types of pseudo data for TS model training: data sampled on the golden parallel corpus, data sampled on the pseudo parallel corpus, and data extracted with word alignment. However, the details of the pseudo data augmentation in this paper are slightly different from those of Yang et al. (2021). Full details are exhibited in the following subsections.

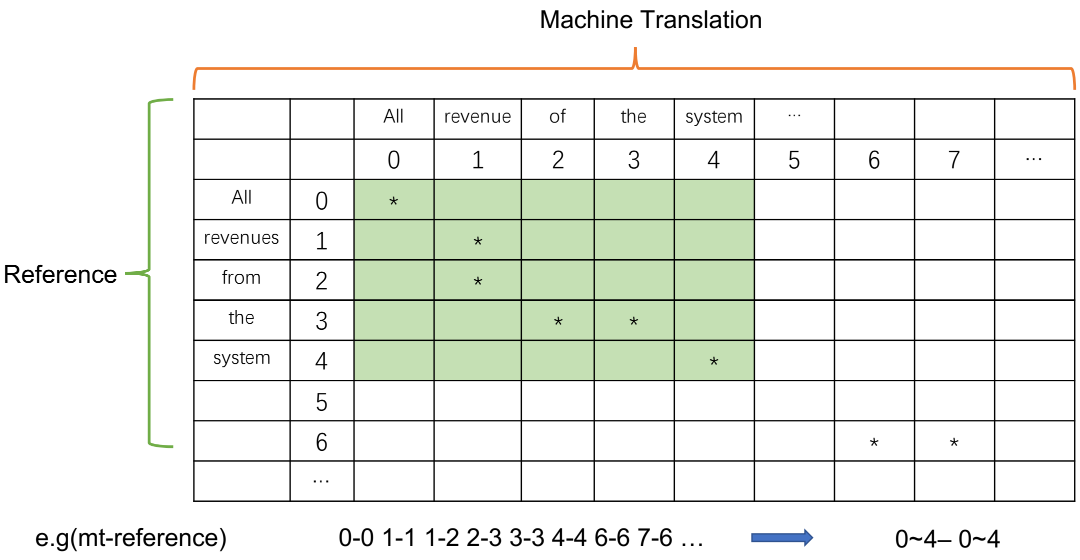
Algorithm 1 Algorithm of Phrase Align

```
Input: $\mathbf{y}$, $\mathbf{r}$, $\mathbf{A}$
Output: $\hat{\mathbf{p}}$
Function GenerateAlign($\mathbf{y}$, $\mathbf{r}$, $\mathbf{A}$):
    $yt = size(\mathbf{y}), rt = size(\mathbf{r})$
    for $i \leftarrow 0$ to $yt$ do
        for $j \leftarrow i$ to $yt$ do
            for $a \leftarrow 0$ to $rt$ do
                for $b \leftarrow a$ to $rt$ do
                    if $IsMatch(\mathbf{y}, \mathbf{r}, i, j, a, b, \mathbf{A})$ then
                        do
                            $i += 1; a += 1$
                        while $\mathbf{y}^i == \mathbf{r}^a$
                        do
                            $j -= 1; b -= 1$
                        while $\mathbf{y}^j == \mathbf{r}^b$
                        $\hat{\mathbf{p}}.add((\mathbf{y}^{i:j}, \mathbf{r}^{a:b}))$
    return $\hat{\mathbf{p}}$
Function IsMatch($\mathbf{y}$, $\mathbf{r}$, $i,j,a,b,\mathbf{A}$):
    for $ii \leftarrow i$ to $j$ do
        let T = $\{t_i | \mathbf{r}^{t_i}$ is aligned with $\mathrm{y}^{ii}$ in $\mathbf{A}\}$
        foreach $t_i \in T$ do
            if $t_i < a$ or $t_i > b$ then
                return False
    for $aa \leftarrow a$ to $b$ do
        let T = $\{t_a | \mathbf{r}^{aa}$ is aligned with $\mathrm{y}^{t_a}$ in $\mathbf{A}\}$
        foreach $t_a \in T$ do
            if $t_a < i$ or $t_a > j$ then
                return False
    return True
```

3.1 Sampling from golden parallel corpus

Raw parallel corpus is firstly filtered by the sentence length. All sentence pairs that have less than 20 words or more than 80 words on any side are removed.

Considering that there might be noise data in the corpus, we apply the dual conditional cross-entropy model (Lu et al., 2020) to obtain a quality score for each sample. Sentence pairs with low quality are



		All	revenue	of	the	system	···			
		0	1	2	3	4	5	6	7	···
All	0	*								
revenues	1		*							
from	2		*							
the	3			*	*					
system	4					*				
	5									
	6							*	*	
	···									

Figure 1: In this example, we have the alignment info between machine translation (MT) and reference sentences: 0-0, 1-1, 1-2, 2-3, 3-3 4-4, 6-6, 6-7, the phrase from $0 \sim 4$ in MT are aligned to $0 \sim 4$ in reference. The rectangle enclosed by the aligned phrases between MT and reference should satisfy that each row and each column has at least one *.

filtered.

Then we generate a pseudo corpus with the remained high-quality parallel corpus. ($\mathbf{x}$, $\mathbf{r}$) is marked as the sentence pair of the parallel corpus, where $\mathbf{x}$ is the source sentence and $\mathbf{r}$ is the golden reference. $\|\mathbf{r}\|$ represents the number of tokens in $\mathbf{r}$.

The first step is to randomly sample the length l to mask for the reference r from a uniform distribution:

$$l \sim U(1, \|\mathbf{r}\|) \tag{1}$$

Then a span with l tokens $\mathbf{r}^{i:j}$ is randomly selected by:

$$i \sim U(0, \|\mathbf{r}\| - l), \;\; j = i + l \tag{2}$$

Finally, we get the TS training data $(\mathbf{x}, \mathbf{r}^{\neg i:j}, r^{i:j})$ from each parallel sentence pair $(\mathbf{x}, \mathbf{r})$, where $\mathbf{r}^{\neg i:j}$ is denoted as the masked version of r, in which $\mathbf{r}^{i:j}$ is replaced with a mask token, e.g <MASK_REP>.

3.2 Sampling on Pseudo Parallel Corpus

In addition, the monolingual corpus is another source for data augmentation. We first filter the monolingual data with a language identification process. Then pseudo parallel corpus is generated with NMT models. Finally, TS training data can be generated as we do in Section 3.1.

3.3 Extracting with Word Alignment

In the task of TS, the labels for the masked span is always correct while the translation contexts of the span, $\mathbf{y}^{\neg i:j}$ are not error-free. Therefore, both of the above two types of pseudo data are biased from the task. In pseudo data sampled from golden parallel corpus, the translation contexts are error-free. And the labels of pseudo data from machine translation results are not always correct. To reduce the bias, another way of data augmentation is proposed in Yang et al. (2021). They utilize the alignment between the machine translation and the golden reference to generate pseudo-training samples for TS. We use the similar idea and the details of our alignment-based data augmentation algorithm are described as follows.

Given the triplet ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{r}$) where $\mathbf{x}$ is the source sentence, $\mathbf{y}$ is the machine translation result generated by NMT models, and $\mathbf{r}$ is the reference, we need to find aligned segment pairs $(\mathbf{y}^{i:j}, \mathbf{r}^{a:b})$ between $\mathbf{y}$ and $\mathbf{r}$.

First, we use the Fast Align toolkit (Dyer et al., 2013) to extract token alignments between $\mathbf{y}$ and $\mathbf{r}$. The align result $\mathbf{A}$ is a list of aligned indexes in the format of i-a, which means token $\mathbf{y}^i$ is aligned to $\mathbf{r}^a$. With the token alignments, the next step is to extract aligned-phrase pairs, denoted as $\hat{\mathbf{p}}$. Figure 1 shows an example of an aligned phrase between MT and reference. The algorithm of the aligned-

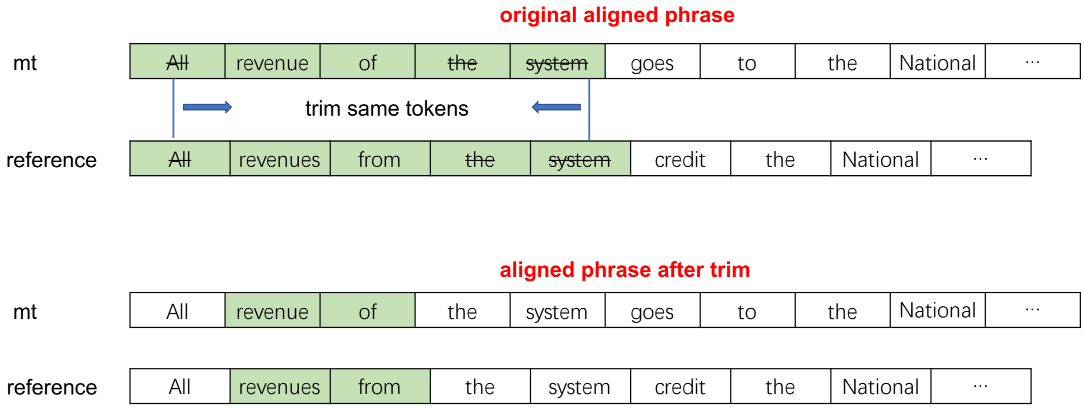


Figure 2: As shown in Figure 1, we get the original aligned phrase between MT and reference which are "All revenue of the system" and "All revenues from the system". We then trim the tokens that appear in both MT and reference to compress the aligned phrase. Finally, we get the trimmed aligned phrase: "revenue of" and "revenues from"

Method	En-De	De-En	En-Zh	Zh-En
TSMind	45.90	43.37	30.21	28.77
-w/o first-phase training	37.14	33.23	21.20	16.44
-w/o second-phase training	37.37	36.83	21.84	19.19

Table 3: Sacre-BLEU on the validation sets of Sub-Task 1 (Naive TS) of the WMT'22 Translation Suggestion Task.

phrase extraction is presented in Algorithm 1 from line 1 to line 13. The aligned phrases are a subset of SMT's phrase extraction (Koehn et al., 2003) with two restricts. 1) Each row and each column of a aligned phrase has at least one token aligned (a * in Figure 1); 2) We take only the longest phrase and the sub-phrases are not taken. After the original aligned phrase is obtained, we remove tokens that appear in both MT and reference to get the trimmed result as shown in Figure 2. We trim these common tokens because we want the model to focus more on the incorrect spans and its alternatives. The pseudo-code of the phrase-alignment is presented in the Algorithm 1. We denote the aligned phrase as $\mathbf{y}^{i:j}$ and $\mathbf{r}^{a:b}$, $\mathbf{y}^{\neg i:j}$ represents the masked version of $\mathbf{y}$ as described in Section 3.2.

Now we need to judge whether $\mathbf{r}^{a:b}$ is better than $\mathbf{y}^{i:j}$ in the context of $\mathbf{y}^{\neg i:j}$. We replace $\mathbf{y}^{i:j}$ with $\mathbf{r}^{a:b}$ in $\mathbf{y}$, and get another new sentence $\hat{\mathbf{y}}$. First, we use the dual conditional cross-entropy model as described in Section 3.1 to calculate the quality score of $(\mathbf{x}, \hat{\mathbf{y}})$. Then, the perplexity of $\hat{\mathbf{y}}$ and $\mathbf{y}$ are given by the language-specific GPT2 models (Schweter, 2020; Radford et al., 2019; Zhao et al., 2019) released on HuggingFace (Wolf et al., 2020) respectively. If the cross-entropy quality score of $(\mathbf{x}, \hat{\mathbf{y}})$ is smaller than the threshold of β_1 and the perplexity loss reduction value of $\mathbf{y} - \hat{\mathbf{y}}$ is at least β_2, then the translation $\hat{\mathbf{y}}$ is most likely better than $\mathbf{y}$. We can treat $\mathbf{y}^{\neg i:j}$ as the masked version of MT and $\mathbf{r}^{a:b}$ as the correct alternative. β_1 and β_2 are the hyper-parameters of the alignment.

Finally, we get the aligned training data ($\mathbf{x}$, $\mathbf{y}^{\neg i:j}$, $\mathbf{r}^{a:b}$) from the triplets $(\mathbf{x}, \mathbf{y}, \mathbf{r})$.

4 Experiment

4.1 Corpus and Setup

Parallel corpora for data augmentation in Section 3.1 and 3.3 and monolingual corpora for Section 3.2 are all downloaded from WMT22 general translation task[1]. For English ↔ German, WikiMatrix (Schwenk et al., 2021), News Commentary v16, Common Crawl Corpora, and Tilde MODEL Corpora (Rozis and Skadiņš, 2017) are used as parallel corpus. For English ↔ Chinese, parallel corpus we used includes UN Parallel Corpus V1.0 (Ziemski et al., 2016) and all parallel corpora from CCMT corpus (Yang et al., 2019) except for the casict2015 corpora. For monolingual corpora, News Commentary and News Crawl are used for all three languages, and Leipzig Corpora (Goldhahn et al., 2012) is also used for Chinese and German.

[1] https://statmt.org/wmt22/translation-task.html

	En-De	De-En	En-Zh	Zh-En	Average
XLM-R	25.12	27.40	32.48	21.25	26.56
Naïve Transformer	28.15	30.08	35.01	24.20	29.36
Dual-source Transformer	28.09	30.23	35.10	24.29	29.43
SA-Transformer	29.48	31.20	**36.28**	25.51	30.62
TSMind	**47.44**	**45.02**	26.41	**31.78**	**37.66**

Table 4: Sacre-BLEU on the test sets of WeTS (Yang et al., 2021)

Then the filtering strategies proposed in Section 3.1 are applied to the raw parallel data. The number of data remained after every filtering step can be found in Table 2.

We download monolingual data from WMT22, and get a total of 45.02 million German, 14.68 million English and 10.01 million Chinese monolingual sentences.

For data augmentation in Sections 3.2 and 3.3, we use the NMT models for English ↔ German and English ↔ Chinese released by Yang et al. (2021)[2] to translate the source sentences. And the hyper-parameter β_1 and β_2 to filter aligned phrases are set to 2.5 and 0.05, respectively.

4.2 Model Training

As mentioned in Section 2, a well-trained NMT model is a good starting point for the TS model. For English ↔ German, we initialize the weights with the NMT models released by Ng et al. (2019) (Winner of WMT'19). For English ↔ Chinese, the one-to-many and many-to-one mBART50 models (Tang et al., 2020) are used.

We use the fairseq toolkit (Ott et al., 2019) to train and evaluate our model. Hyper-parameters are set to the same as examples in the fairseq toolkit except that we reset the learning rate at the beginning of the first phase training and beam size is set as 6 during inference.

4.3 Experimental Results

We evaluate the TSMind by calculating the Sacre-BLEU (Post, 2018) of the top-1 generated translation suggestion candidate on the golden reference. Results of the validation sets of WMT22 are shown in Table 3. Without first-phase training, we get much worse performances. This demonstrates that a large amount of pseudo corpora contributes much to the model. However, without the second-phase training (i.e. without the human-labeled data), we cannot obtain a good translation suggestion model with only pseudo corpora either. Therefore, the design of the two-phase training and the pseudo corpora are essential to set good translation suggestions.

Since the development set of WMT'22 is not the same as the test set used in Yang et al. (2021), to make a fair comparison, we also report the Sacre-BLEU on the test set of WeTS in Table 4. Results of all baseline systems are reported by Yang et al. (2021). TSMind outperforms the strong baseline, SA-Transformer, significantly with a gap of 7.04 BLEU on average for all four language pairs. We notice that TSMind does not perform well on the English to Chinese language pair. The reason might be that the pre-trained model we use is the one-to-many model of mBART50, and the multilingual decoder is not well-trained for Chinese. For example, on the English to Chinese news translation test set of WMT'20 (Barrault et al., 2020), mBART50 only achieves a Sacre-BLEU value of 30.79, while the Sacre-BLEU of state-of-the-art is 49.2.

5 Conclusion

In this paper, we present our translation suggestion systems, TSMind, for the WMT 2022 Translation Suggestion Task. Different from previous work, we use well-trained NMT models as the pre-trained models and applied a two-phase training strategy.

We explore three data augmentation strategies from previous work and utilize the dual conditional cross-entropy model to filter out low-quality augmented data. The leader board finally shows that our submissions are ranked first in three of four language directions in the Naive TS task of WMT22 Translation Suggestion task.

References

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, et al. 2009. Statistical approaches to

[2] https://github.com/ZhenYangIACAS/WeTS

computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Lynne Bowker. 2002. *Computer-aided translation technology: A practical introduction*. University of Ottawa Press.

Lynne Bowker. 2014. Computer-aided translation: Translator training. In *Routledge encyclopedia of translation technology*, pages 126–142. Routledge.

Lynne Bowker and Des Fisher. 2010. Computer-aided translation. *Handbook of translation studies*, 1:60–65.

Rajen Chatterjee. 2019. Automatic post-editing for machine translation. *arXiv preprint arXiv:1910.08592*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Miguel Domingo, Alvaro Peris, and Francisco Casacuberta. 2016. Interactive-predictive translation based on multiple word-segments. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 282–291.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Jesús González-Rubio, Daniel Ortiz-Martínez, Francisco Casacuberta, and José Miguel Benedi Ruiz. 2016. Beyond prefix-based interactive translation prediction. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 198–207, Berlin, Germany. Association for Computational Linguistics.

Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2015. Natural language translation at the intersection of ai and hci. *Communications of the ACM*, 58(9):46–53.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120, Austin, TX, USA. The Association for Machine Translation in the Americas.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Kis Balázs Lengyel, István and Gábor Ugray. 2004. Memoq: A new approach to computer-assisted translation.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

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 NAACL-HLT 2019: Demonstrations*.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Stefan Schweter. 2020. German gpt-2 model.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Muyun Yang, Xixin Hu, Hao Xiong, Jiayi Wang, Yiliyaer Jiaermuhamaiti, Zhongjun He, Weihua Luo, and Shujian Huang. 2019. Ccmt 2019 machine translation evaluation report. In *China Conference on Machine Translation*, pages 105–128. Springer.

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

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).