Findings of the WMT 2022 Shared Task on Translation Suggestion

Zhen Yang, Fandong Meng, Yingxue Zhang, Ernan Li, and Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China
{zieenyang, fandongmeng, yxuezhang, cardli, withtomzhou}@tencent.com

Abstract

We report the result of the first edition of the WMT shared task on Translation Suggestion (TS). The task aims to provide alternatives for specific words or phrases given the entire documents generated by machine translation (MT). It consists two sub-tasks, namely, the naive translation suggestion and translation suggestion with hints. The main difference is that some hints are provided in sub-task two, therefore, it is easier for the model to generate more accurate suggestions. For sub-task one, we provide the corpus for the language pairs English-German and English-Chinese. And only English-Chinese corpus is provided for the sub-task two.

We received 92 submissions from 5 participating teams in sub-task one and 6 submissions for the sub-task 2, most of them covering all of the translation directions. We used the automatic metric BLEU for evaluating the performance of each submission.

1 Introduction

Computer-aided translation (CAT) (Barrachina et al., 2009; Green et al., 2014; Knowles and Koehn, 2016; Santy et al., 2019) has attained more and more attention for its promising ability in combining the high efficiency of machine translation (MT) (Cho et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) and high accuracy of human translation (HT). A typical way for CAT tools to combine MT and HT is PE (Green et al., 2013; Zouhar et al., 2021), where the human translators are asked to provide alternatives for the incorrect word spans in the results generated by MT. To further reduce the post-editing time, researchers propose to apply TS into PE, where TS provides the sub-segment suggestions for the annotated incorrect word spans in the results of MT, and their extensive experiments show that TS can substantially reduce translators' cognitive loads and the post-editing time (Wang et al., 2020; Lee et al., 2021).

As there is no explicit and formal definition for TS, we observe that some previous works similar or related to TS have been proposed (Alabau et al., 2014; Santy et al., 2019; Wang et al., 2020; Lee et al., 2021). However, there are two main pitfalls for these works in this line. First, most conventional works only focus on the overall performance of PE but ignore the exact performance of TS. This is mainly because the golden corpus for TS is relatively hard to collect. As TS is an important sub-module in PE, paying more attention to the exact performance of TS can boost the performance and interpretability of PE. Second, almost all of the previous works conduct experiments on their in-house datasets or the noisy datasets built automatically, which makes their experiments hard to be followed and compared. Additionally, experimental results on the noisy datasets may not truly reflect the model's ability on generating the right predictions, making the research deviate from the correct direction. Therefore, the community is in dire need of a benchmark for TS to enhance the research in this area. To address the limitations mentioned above and spur the research in TS, we make our efforts to construct a high-quality benchmark dataset with human annotation, named *WeTS*,[1] which covers four different translation directions.

The main motivation of this shared task is twofold. The first goal is to analyze the challenges in the area of TS, which can provide some new directions for the further researches and applications in this area. Secondly, we want to make the researchers notice the gaps between the golden and automatically generated synthetic corpus. And we want to see the performance of different techniques on the golden corpus. As the source and translation sentence are both the inputs of TS, it is interesting to see how the interactions between the source and

[1] *WeTS*: We Establish a benchmark for Translation Suggestion

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 821 - 829

translation sentences can improve the final suggestions.

In order to evaluate the quality of the participating systems, we use the automatic metric, BLEU (Papineni et al., 2002). Specifically, we adopt the widely used toolkit, sacrebleu (Post, 2018) to calculate the BLEU score for the top-1 suggestion against the reference sentences.[2] For Chinese, the BLEU score is calculated on teh character with the default tokenizer for Chinese. As for English, the BLEU score is calcualted on the case-sensitive words with the default tokenizer 13a.

Five teams participated in this first campaign of the Translation Suggestion shred task, most of them cover the four translation directions. We will describe each system which submits the technical paper in detail.

2 Task Description

This section describes the task definition in the first edition of TS shared task. We finely divide the task of TS into two sub-tasks, namely *vanilla TS* and *TS with hints*, according to whether the translators' hints are considered.

Vanilla TS. Given the source sentence $\boldsymbol{x} = (x_1, \ldots, x_s)$, the translation sentence $\boldsymbol{m} = (m_1, \ldots, m_t)$, the incorrect words or phrases $\boldsymbol{w} = \boldsymbol{m}_{i:j}$ where $1 \leq i \leq j \leq t$, and the correct alternative $\boldsymbol{y}$ for $\boldsymbol{w}$, the task of *vanilla TS* is optimized to maximize the conditional probability of $\boldsymbol{y}$ as follows:

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

where θ represents the model parameter, and $\boldsymbol{m}^{-\boldsymbol{w}}$ is the masked translation where the incorrect word span $\boldsymbol{w}$ is replaced with a placeholder. [3]

TS with Hints. In the sub-task *TS with hints*, the hints of translators are considered as some soft constraints for the model, and the model is expected to generate suggestions meeting these constraints. The format of the translator's hint is very flexible, which usually requires only a few types on the keyboard by the translator. For English and German, the hints can be the character sequence which includes the initials of words in the correct alternative. As for Chinese, the hints can be the character sequence which includes the initials of the phonetics of words in the correct alternative. In this setting, the model is optimized as:

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

where $\boldsymbol{h}$ indicates the hints provided by translators.

Related tasks. Some similar techniques have been explored in CAT. Green et al. (2014) and Knowles and Koehn (2016) study the task of so-called translation prediction, which provides predictions of the next word (or phrase) given a prefix. Huang et al. (2015) and Santy et al. (2019) further consider the hints of the translator in the task of translation prediction. Compared to TS, the most significant difference is the strict assumption of the translation context, i.e., the prefix context, which severely impedes the use of their methods under the scenarios of PE. Lexically constrained decoding which completes a translation based on some unordered words, relaxes the constraints provided by human translators from prefixes to general forms (Hokamp and Liu, 2017; Post and Vilar, 2018; Kajiwara, 2019; Susanto et al., 2020). Although it does not need to re-train the model, its low efficiency makes it only applicable in scenarios where only a few constraints need to be applied. Recently, Li et al. (2021) study the problem of auto-completion with different context types. However, they only focus on the word-level auto-completion, and their experiments are also conducted on the automatically constructed datasets.

3 Data Description

This section introduces the proposed dataset *WeTS* used in the shred task, which is a golden corpus for four translation directions, including English-to-German, German-to-English, Chinese-to-English and English-to-Chinese.

Translation Direction	Train	Valid	Test
En⇒De	14,957	1000	1000
De⇒En	11,777	1000	1000
Zh⇒En	21,213	1000	1000
En⇒Zh	15,769	1000	1000

Table 1: The sizes for cases in train/valid/test sets. "En⇒De" refers to the direction of English-to-German, and "En⇒Zh" refers to English-to-Chinese.

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

[3] $\boldsymbol{w}$ is null if i equals j, and the model will predict whether some words need to be inserted in position i.

Source Sentence	他们也许并不知道这是一个“假理财”骗局，但也察觉到了诸多可疑之 ta men ye xu bing bu zhi dao zhe shi yi ge jia li cai pian ju, dan ye cha jue dao le zhu duo ke yi zhi chu 处，然而最终还是按照张颖的指使进行了违法违规操作。 ran er zui zhong hai shi an zhao zhang ying de zhi shi jin xing le wei fa wei gui cao zuo
Translation	They may not know this is a "fake financial management" scam, but also aware of many suspicious, and ultimately conduct illegal operations according to Zhang Ying's instructions.
Suggestions	1. suspects 2. doubtful points 3. questionable points

Figure 1: One training example in *WeTS*. For the incorrect word "suspicious" (in red color), there are three correct suggestions. For readability, we also provide the Chinese pinyin format for the Chinese sentence (in blue color).

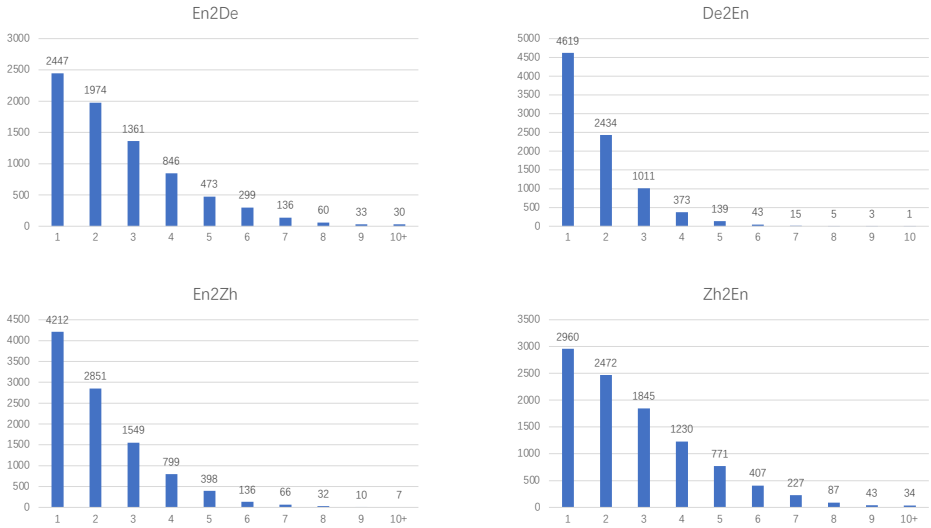

Figure 2: The number of incorrect span in each annotated example.

3.1 Annotation Guidelines

It is non-trivial for annotators to locate the incorrect word spans in the MT sentence. The main difficulty is that, the concept of “translation error" is ambiguous and each translator has his own understanding about translation errors. To easier the annotation workload and reduce the possibility of making errors, we group the translation errors on which we aim to focus into three macro categories:

- Under-translation or over-translation: While the problem of under-translation or over-translation has been alleviated with the popularity of Transformer, it is still one of the main mistakes in NMT and seriously destroys the readability of the translation.
- Semantic errors: For the semantic error, we mean that some source words are incorrectly translated according to the semantic context, such as the incorrect translations for entities, proper nouns, and ambiguous words. Another branch of semantic mistake is that the source words or phrases are only translated superficially and the semantics behind are not translated well.
- Grammatical or syntactic errors: Such errors usually appear in translations of long sentences, including the improper use of tenses, passive voice, syntactic structures, etc.

Another key rule for translators is that annotating the incorrect span as local as possible, as generating correct alternatives for long sequences is much harder than that of shorter sequences.

3.2 Data Construction

As the starting point, we collect the monolingual corpora for English and German from the raw Wikipedia dumps, and extract Chinese monolingual corpus from various online news publications. We first clean the monolingual corpora with a language detector to remove sentences belonging to

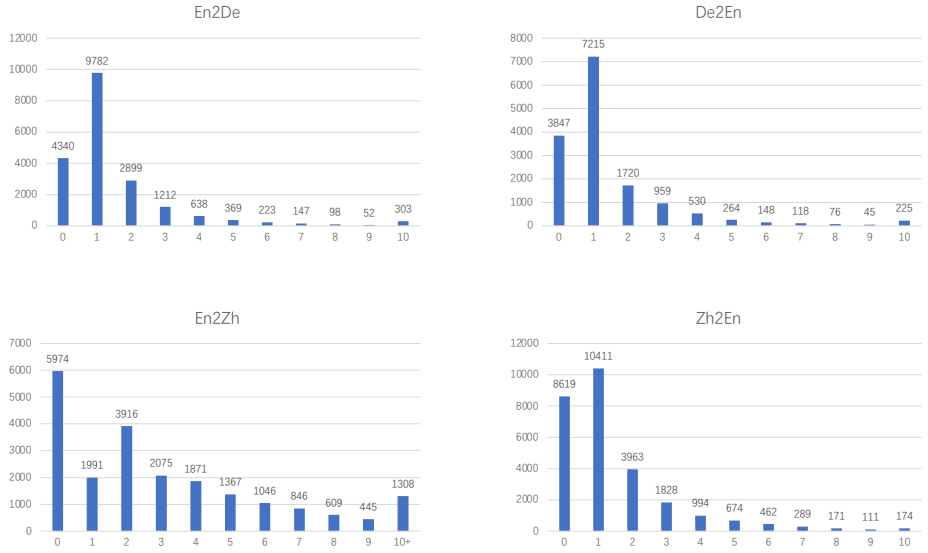

Figure 3: The length of the incorrect span.

other languages.[4] For all monolingual corpora, we remove sentences that are shorter than 20 words or longer than 80 words. In addition, sentences which exist in the available parallel corpora are also removed. Then, we get the translations by feeding the cleaned monolingual corpus into the corresponding fully-trained NMT model. The NMT models for English-German language pairs are trained on the parallel corpus of WMT14 English-German. For Chinese-English directions, the NMT models are trained with the combination between the WMT19 English-Chinese[5] and the same amount of in-house corpus. [6]

Finally, the translators are required to mark the incorrect word spans in the translation sentence and provide at least one alternative for each incorrect span, by using the annotation guidelines. The team is composed by eight annotators with high expertise in translation and each example has been assigned to three experts. There are two phases of agreement computations. In the first phase, an annotation is considered in agreement among the experts if and only if they capture the same incorrect word spans. If one annotation passes the first agreement computation, it will be assigned to other three experts in charge of selecting the right alternatives from the previous annotation. In the second phase of agreement computation, an annotation is considered in agreement among the experts if and only if they select the same right alternatives. With the two-phase agreement checking, we ensure the high quality of the annotated examples. For the annotated examples with multiple incorrect word spans, we can extract multiple examples which have the same source and translation sentences, but different incorrect word span and the corresponding suggestions. Finally the extracted examples are randomly shuffled and then split into the training, validation and test sets.[7] One training example for the translation direction of Chinese-to-English is presented in Figure 1 and the sizes for the train/valid/test sets in *WeTS* are collected in Table 1.

3.3 Detailed Statistics

The number of the incorrect span Each annotated example may contain multiple incorrect spans, we show the number of the incorrect span in each annotated example as Figure 2. We can see that most examples have only a few incorrect spans, and there are more than 70 percent examples containing less than 3 incorrect spans for each translation direction.

The length of the incorrect span Figure 3 represents the length distribution of the incorrect spans. We can find that most of the incorrect spans contain less than 3 words or Chinese characters. This is mainly because of our key rule for annotating the incorrect span as local as possible. Additionally, for all of the four translation directions, the

[4] https://github.com/Mimino666/langdetect

[5] https://www.statmt.org/wmt19/translation-task.html

[6] We have released the models and inference scripts utilized here to make our results easy reproduced.

[7] To keep the fairness of *WeTS*, we ensure the examples among the training, validation and test sets have different source and translation sentences.

Figure 4: The length of the suggestion.

number of the incorrect spans with length 0 ranks top-2 among all the length buckets. This shows that under-translation is still a frequent error of the existing NMT models.

The length of the suggestions Figure 4 shows the length distribution of the suggestions. We can see that in English-to-German, German-to-English and Chinese-to-English, most of the suggestions contain only one word. For English-to-Chinese, most suggestions contain two Chinese characters. Additionally, we can also find that there are quite a few of suggestions with length zero in each translation direction. This shows that over-translation is a non-negligible problem for the existing NMT models.

4 Participants

Five participants submitted their systems to the sub-task one of TS shared task. And two participants submitted their systems to the second sub-task. In sub-task one, 92 runs were submitted in total (each team is only allowed to submit less than 15 runs). Table 2 summarizes the participants and their affiliations.

4.1 Systems

Here we briefly describe each participant's systems as described by the authors and refer the reader to the participant's submission for further details. Since some participants did not submit their papers, we only describe the systems in the submitted papers.

Team	Institution
mind-ts	Soochow University and Alibaba
suda-hlt	Soochow University
Avocados	Beijing Jiaotong University
IOL Research	Transn IOL Technology CO., Ltd.
Slack	Zhejiang University

Table 2: The participating teams and their affiliations.

4.1.1 Baseline

We take the naive Transformer-base (Vaswani et al., 2017) as the baseline and directly apply the implementation of the open-source toolkit, fiarseq.[8] We construct the synthetic corpus based on the WMT parallel corpus, and we refer the readers for details about constructing the synthetic corpus in the paper (Yang et al., 2021). For training, we apply the two-state training pipeline, where we pre-train the model on the synthetic corpus in the first stage, and then fine-tune the model on the golden corpus in the second stage.

4.1.2 IOL Research

The team of IOL Research participates the two sub-tasks and focuses on the En-Zh and Zh-En translation directions. They use the ΔLM as their backbone model. ΔLM is a pre-trained multilingual encoder-decoder model, which outperforms various strong baselines on both natural language generation and translation tasks (Ma et al., 2021). Its encoder and decoder are initialized with the

[8] https://github.com/pytorch/fairseq

pre-trained multilingual encoder InfoXLM (Chi et al., 2020). Their model has 360M parameters, 12-6 encoder-decoder layers, 768 hidden size, 12 attention heads and 3072 FFN dimension. For the training data, they construct the synthetic data with two different methods according to its constructing complexity. During training, they use the two-stage fine-tuning, where they apply the synthetic data to fine-tune the original ΔLM in the first stage and then fine-tune the result of the first stage with the golden corpus. In their experiments, they find that the accuracy indicator of TS can be helpful for efficient PE in practice. Overall, they achieved the best scores on 3 tracks and comparable result on another track.

4.1.3 Avocados

The team of Avocados tries different model structures, such as Transformer-base (Vaswani et al., 2017), Transformer-big (Vaswani et al., 2017), SA-Transformer (Yang et al., 2021) and DynamicConv (Wu et al., 2019). They test different ensemble approaches for better performance. For more details, we refer the readers to their paper (Zhang et al., 2022). Their main efforts are paid on building the synthetic corpus. They apply three different ways to construct the synthetic corpus. Firstly, they 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, 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. To bridge the domain difference between the large-scale synthetic data and human-annotated golden corpus, they apply the pre-trained BERT to 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. Overall, they rank second and third on the English-German and English-Chinese bidirectional tasks respectively.

4.1.4 mind-ts

The team of mind-ts participate in the English-German and English-Chinese translation directions in the sub-task one, and their submissions are ranked first in three of four language directions. For English-German, they initialize the weights with NMT models released by teh winner of WMT19 (Ng et al., 2019). For English-Chinese, the one-to-many and many-to-one mBART50 models are used (Tang et al., 2020). Their main contribution is to construct the synthetic corpus with word alignment. They use the well-trained alignment models between source and target languages to filter out high-quality augment data. Specifically, they first use the Fast Align toolkit to extract the token alignments. Then, they remove tokens that appear in both MT and reference to get the trimmed result. They trim these common tokens because they want the model to focus more on the incorrect span and its alternative. Additionally, they use the dual conditional cross-entropy model to calculate the quality score of the pair between the source and masked translation sentences. If the cross-entropy quality score meets the threshold, they treat the masked translation and the alignment segments as the good examples for TS. Similarly, they also use the two-phase pre-training pipeline to get the final models.

4.2 Submission Summary

The submissions for this year's TS shared task cover different approaches from the pre-trained LMs and the encoder-decode NMT models. From the submissions, we find that the pre-trained models are very useful for the final performance. Additionally, almost all of the submissions have tried different approaches for constructing the synthetic corpus. As the amount of the golden corpus is limited, it is very important to find efficient ways to construct the synthetic corpus. The main problem for constructing synthetic corpus is how to make the synthetic corpus similar to the golden corpus in domain or other aspects. Finally, how to efficiently apply the synthetic corpus also needs much more efforts to investigate. All submissions adopt the two-stage training pipeline to train the models.

4.3 Evaluation Results

We report the BLEU scores of the submissions. The BLEU is calculated automatically with the sacrebleu toolkit. For each run, the participating

team need to submit their top-1 suggestions for each sentence in the test set. Each participating team can submit at most 15 times for each track. We only report the best score for each team. Table 3 and 4 report the results on English-Chinese and English-German respectively in the sub-task one. Table 5 report the results on English-Chinese in the sub-task two.

Team	En-Zh	Zh-En
Baseline	31.02	25.84
mind-ts	33.92(2)	30.07 (1)
Avocados	33.33 (3)	28.56 (3)
IOL Research	39.71 (1)	28.42 (4)

Table 3: Evaluation results on the language pair for English-Chinese in the sub-task one. The number in bracket is the ranked position.

Team	En-De	De-En
Baseline	35.07	37.61
mind-ts	42.91(1)	47.04 (1)
Avocados	42.61 (2)	36.30 (2)

Table 4: Evaluation results on the language pair for English-German in the sub-task one. The number in bracket is the ranked position.

Team	En-Zh	Zh-En
Baseline	41.83	35.02
IOL Research	48.60 (1)	39.95 (1)

Table 5: Evaluation results on the language pair for English-Chinese in the sub-task two. The number in bracket is the ranked position.

5 Discussion and Analysis

Comparing the results of the BLEU scores of all submissions with our baseline systems, there is a significant gap between the submitted and baseline systems. This shows that there is a large space for us to try different techniques to improve the performance of TS. By comparing the results of different submitted systems, we find that different pre-training models have a large difference on the final performance. This is a similar trend with other NLP tasks. Therefore, we believe that this is an interesting and promising direction for us to pay much more efforts.

All submitted systems have investigated different approaches for constructing the synthetic corpus and almost all of them have achieved much improvements with the synthetic corpus. The noise in the synthetic corpus is a major problem which negatively affects the final performance. Therefore, how to filter or decrease the noise is an open question. The team of mind-ts applies the pre-trained LM to filter the synthetic corpus and obtain better performance on 3 out of 4 tracks based on the high-quality synthetic corpus. We can investigate more effective approaches to detect and filter the noise in the synthetic corpus.

However, another interesting direction which are not investigated by the submissions is modeling the interaction between the source and translation sentences efficiently. Compared to MT, the main difference for TS is that the input for TS is dual-source, namely the source and translation sentence. We believe that efficiently modeling the interaction between the source and translation sentences can improve the final performance.

6 Conclusion

We present the results of first edition of the Translation Suggestion shared task. For the goal of this task, we create and release the first golden benchmark dataset, called *WeTS*, which covers the language pairs for English-Chinese and English-German. We wish the released corpus can spur the researches in this area. This year we received 92 submissions from 5 participating teaming in the sub-task one and 6 submissions for the sub-task 2, most of them covering the two translation directions. Results of these submissions show that the pre-trained models and synthetic corpus are two important factors for the final performance.

Acknowledgements

We would like to thank Yaou Li and Ning Zhang for their helps on building the official website of the shared task. The authors would also like to thank the anonymous reviewers of this paper, and the anonymous reviewers of the previous version for their valuable comments and suggestions to improve our work.

References

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448.

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.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

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*.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052.

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

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, Online. Association for Computational Linguistics.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: General word-level Autocompletion for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4792–4802, Online. Association for Computational Linguistics.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Qian Wang, Jiajun Zhang, Lemao Liu, Guoping Huang, and Chengqing Zong. 2020. Touch editing: A flexible one-time interaction approach for translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 1–11.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

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

Hongxiao Zhang, Siyu Lai, Songming Zhang, Hui Huang, Yufeng Chen, Jinan Xu, and Jian Liu. 2022. Improved data augmentation for translation suggestion. *arXiv preprint arXiv:2210.06138*.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.