MiSS: An Assistant for Multi-Style Simultaneous Translation

Zuchao Li[1,2,3,†], **Kevin Parnow**[1,2,3], **Masao Utiyama**[4,*], **Eiichiro Sumita**[4], **Hai Zhao**[1,2,3*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[4]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
charlee@sjtu.edu.cn, {mutiyama, eiichiro.sumita}@nict.go.jp, zhaohai@cs.sjtu.edu.cn

Abstract

In this paper, we present **MiSS**, an assistant for multi-style simultaneous translation. Our proposed translation system has five key features: highly accurate translation, simultaneous translation, translation for multiple text styles, back-translation for translation quality evaluation, and grammatical error correction. With this system, we aim to provide a complete translation experience for machine translation users. Our design goals are high translation accuracy, real-time translation, flexibility, and measurable translation quality. Compared with the free commercial translation systems commonly used, our translation assistance system regards the machine translation application as a more complete and fully-featured tool for users. By incorporating additional features and giving the user better control over their experience, we improve translation efficiency and performance. Additionally, our assistant system combines machine translation, grammatical error correction, and interactive edits, and uses a crowdsourcing mode to collect more data for further training to improve both the machine translation and grammatical error correction models. A short video demonstrating our system is available at https://www.youtube.com/watch?v=ZGCo7KtRKd8.

1 Introduction

With the increasing technological development of the world and the acceleration of globalization, people from different languages and cultural backgrounds communicate more and more, and the needs of translation are becoming more and more important and diverse. Although traditional manual translation works well, with the increasing frequency of international communication, traditional manual translation far from meets demand, and machine translation has correspondingly risen in popularity (Hutchins and Somers, 1992). Recently, Neural Machine Translation (NMT), especially Transformer-based NMT, has emerged as a promising approach with the potential to address many of the shortcomings of traditional rule-based or statistics-based machine translation systems (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). This has significantly improved the performance of machine translation and other related tasks (Huang et al., 2018; Li et al., 2018a,b).

Although neural machine translation has made tremendous improvements and is relatively high-performing, because human language is so complex, machine translation is often still only used as an assistance tool rather than the sole entity responsible for translation. There are several popular and large existing commercial machine translation systems that provide users with effective translation (e.g., Google Translator, Bing Translator, Amazon Translate, and Baidu Translate). As NMT is still very imprecise, however, these web services fall short, as they do not provide sufficient information to users in how good each translation is, which is particularly pertinent to those who have not mastered the target language. VoiceTra[1] included back-translation in the machine translation system to alleviate this deficiency; however, this practice requires users to perform additional manual evaluations, which brings new usage costs.

In mainstream machine translation systems, sentences or paragraphs are used as the units of translation, which means that it takes a relatively long time to provide users with translated content. Simultaneous machine translation, translating sentences in real-time while the user speaks or types, can significantly reduce this translation time, but its performance lags behind that of standard NMT. Although some commercial machine translation systems such

*Corresponding author. †This paper was partially finished when Zuchao Li was a fixed term technical researcher at NICT. This paper was supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

[1] https://voicetra.nict.go.jp/

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 1–10
November 7–11, 2021. ©2021 Association for Computational Linguistics

as Google and Baidu have introduced simultaneous translation feature, due to the integration of simultaneous translation and whole-sentence translation, users cannot easily control whether the system uses simultaneous translation or whole-sentence translation, and the automated control of commercial systems sometimes does not follow the user's wishes.

Since user input errors are unavoidable for any human-computer interaction system, the quality of NMT system also has been shown to significantly degrade when confronted with source-side noise (Heigold et al., 2018; Belinkov and Bisk, 2018; Anastasopoulos, 2019). The previous grammatical error detection and correction work focused on computer-aided writing systems. Some existing computer-aided writing systems (Grammarly[2] and Pigai[3], Write&Improve[4], and LinggleWrite[5]) detect and correct grammatical errors; however, systems such as these have had little attention when considered in the context of input error detection or correction for commercial machine translation systems, as their main focus is generally post-translation editing.

High quality domain specific machine translation systems are in high demand whereas general purpose MT has limited applications because different machine translation users want to generate translations that can be used in the scenario. On the one hand, general purpose translation systems usually perform poorly (Koehn and Knowles, 2017). On the other hand, appropriate translation is also a very important goal to pursue. There are two typical methods to achieve this goal. One is to use the domain adaptation method to obtain a domain-specific model from the existing general machine translation model through transfer learning. The other is to adopt an conditional translation decoder to integrate various domains into the same model and generate translations according to different input conditions (Keskar et al., 2019). At present, the commercial machine translation system mainly adopts the former one, but it also brings the additional deployment cost.

Considering the deficiencies of existing systems, the new needs of users, and the current development of natural language processing, we developed a web-based machine translation demonstration system **MiSS**. In this system, we tried to integrate several new features to provide better services for users. With **MiSS**, users can get real-time translations while writing, flexible control in switching between real-time translation and whole-sentence translation, informative back-translation feedback and scoring, and input error detection and revision suggestions. In addition, the system also supports user interactions that modify the translations or inputs, which provides crowdsourced data for further improving the performance of our machine translation and grammatical error correction. Notably, there were also several interactive translation systems in the past, such as CASMACAT (Alabau et al., 2014), (Knowles and Koehn, 2016), (Peris et al., 2017), and INMT (Santy et al., 2019). The distinctions lie in the abilities of the systems and the features to adapt to the latest user needs.

2 The MiSS System

There are 5 features in our MiSS translation system: simultaneous translation, back-translation for quality evaluation, grammatical error correction, multi-style translation, and crowdsourcing data collection. The system is available at `http://miss.x2brain.com/` until November 12, 2021. We show a screenshot of the system in Figure 1. In the following subsections, we will describe each component of the system.

2.1 Basis: Transformer-based NMT

Transformer (Vaswani et al., 2017) is an attention mechanism-based network. This architecture introduced the innovative self-attention network (SAN) that computes the relationships between all tokens in the source sequence. (Hassan et al., 2018; Läubli et al., 2018; Li et al., 2020a, 2021) observed that Transformer-based NMT has achieved performance similar to human-level performance on some benchmarks, and because of this tremendous performance, this model has been widely used in the field of machine translation. Given the excellent performance of Transformer-based NMT, we use it as the basis for our system. The model includes an encoder and a decoder, which are respectively used for incrementally processing the source and target sentences. Both the encoder and decoder are stacks of L Transformer blocks.

2.2 Feature #1: Simultaneous Translation

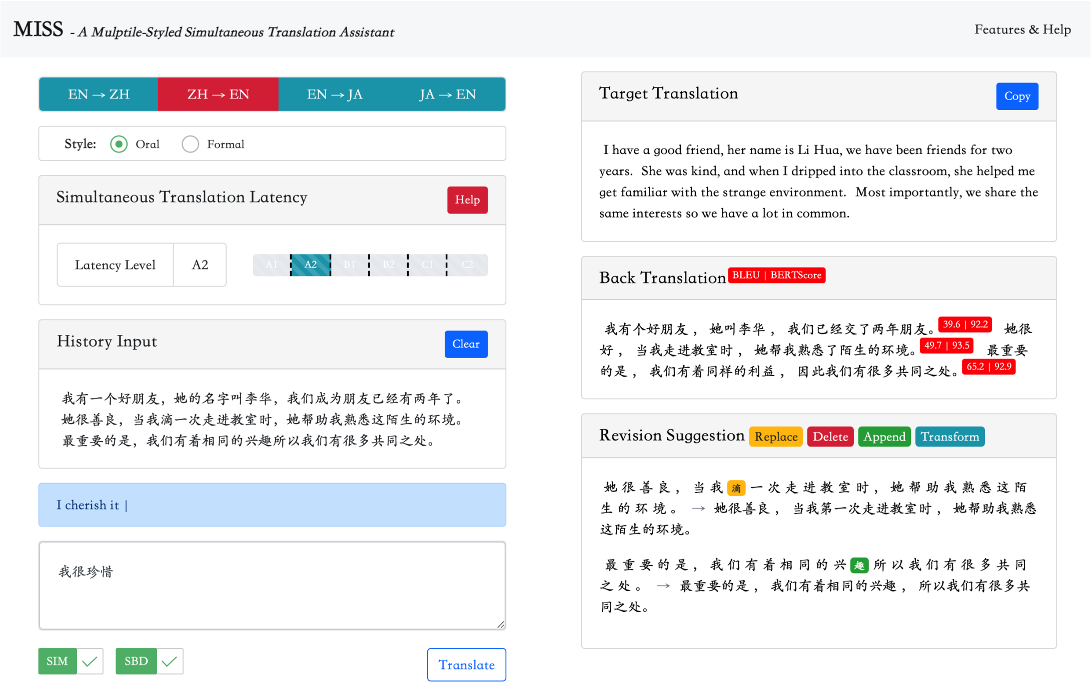
Simultaneous NMT has attracted much attention recently. In contrast to standard NMT, where the

[2] https://www.grammarly.com

[3] https://www.pigai.org

[4] https://writeandimprove.com

[5] https://f.linggle.com



Style: ● Oral ○ Formal

Simultaneous Translation Latency

Latency Level A2

History Input

我有一个好朋友，她的名字叫李华，我们成为朋友已经有两年了。
她很善良，当我滴一次走进教室时，她帮助我熟悉这陌生的环境。
最重要的是，我们有着相同的兴趣所以我们有很多共同之处。

I cherish it |

我很珍惜

Target Translation

I have a good friend, her name is Li Hua, we have been friends for two years. She was kind, and when I dripped into the classroom, she helped me get familiar with the strange environment. Most importantly, we share the same interests so we have a lot in common.

Back Translation

我有个好朋友， 她叫李华， 我们已经交了两年朋友。39.6 | 92.2 她很好， 当我走进教室时， 她帮我熟悉了陌生的环境。49.7 | 93.5 最重要的是， 我们有着同样的利益， 因此我们有很多共同之处。65.2 | 92.9

Revision Suggestion

她很善良， 当我滴一次走进教室时， 她帮助我熟悉这陌生的环境。 → 她很善良， 当我第一次走进教室时， 她帮助我熟悉这陌生的环境。

最重要的是， 我们有着相同的兴趣所以我们有很多共同之处。 → 最重要的是， 我们有着相同的兴趣， 所以我们有很多共同之处。

Figure 1: The screenshot of MISS system.

NMT system can access the full input sentence, simultaneous NMT can only utilize the current state of an input sentence (which may be incomplete). Because of this, the translation task entails more uncertainty and consequently, more difficulty. Current simultaneous NMT systems model the task as a prefix-to-prefix problem. Among them, *wait-k* inference (Ma et al., 2019) is a simple yet effective strategy for simultaneous NMT. In *wait-k*, the decoder is asked to generate the output sequence k words behind the input words. Specifically, the *wait-k* strategy is defined as follows: given an input $x \in X$, the generation of the translation y is always k tokens behind reading x; that is, at the t-th decoding step, we generate token y_t based on $x \leq t - k + 1$. We thus adopt a Transformer-based NMT model with the *wait-k* strategy, aiming for balance between translation performance and efficiency.

2.3 Feature #2: Back-translation for Quality Evaluation

A machine translation model on its own is unable to evaluate the quality of its generated translations, as typical translation quality metrics require reference sentences. This lack of obvious evaluation can cause users to mistrust the translation system and doubt whether it accurately expresses a sentence's true meaning. Back-translation – the 'retranslation' of a translated sequence back into its original language – is a potential method of generating reference sentences for comparison that utilizes the duality of direction in translation (He et al., 2016). Back-translation is currently mainly used as a data-enhancement method for supervised NMT systems (Edunov et al., 2018) and as a crucial training method for unsupervised NMT systems (Conneau and Lample, 2019), though it has been more controversial as a method of assessing translations. According to (Behr, 2017)'s conclusion, while back-translation can give some evaluation of the translation, it often raises issues not noted by human assessors, and more importantly, is less reliable in general, as many problems remain hidden. These shortcomings are mainly are a result of commonly used automatic evaluation methods (like BLEU) using only surface-level similarity; they do not strictly measure , Semantic Equivalence (SE), which is the true goal. Thus, we adopted BERTScore(Zhang et al., 2019), a language generation evaluation metric based on pre-trained BERT contextual embeddings, for semantic equivalence assessment and the evaluation metric BT-BLEU (Li et al., 2020b) (also described in (Nguyen et al., 2021) as reconstruction BLEU) for translation quality evaluation. Furthermore, recent work (Fomicheva et al., 2020) mentions various other unsupervised quality evaluation methodologies, we will include it into the follow-up updates and provide a better reference indicator in our system.

2.4 Feature #3: Grammatical Error Correction

Detecting potential grammatical errors and offering corrective suggestions for them sentence is also a very important feature in MISS. We chose the tag-based modeling approach for this feature based on the fresearch field's latest achievements (Omelianchuk et al., 2020) and our recent work (Parnow et al., 2020, 2021) in the Grammatical Error Correction (GEC).

Specifically, the g-transformations developed by (Omelianchuk et al., 2020) were included in our system in the hopes of providing learners more specific suggestions (i.e., the edit type of an error) to revise the users' input. Predicting edits rather than tokens also increases the generalization of our GEC model. G-transformations are based on several basic transformations: $KEEP (keep the current token unchanged), $DELETE (delete current token), $APPEND_t1 (append new token t1 next to the current token), and $REPLACE_t2 (replace the current token with another token t2). From these basic transformations, further, more task-specific transformations are hand designed (such as $CASE (fix the casing of a word), $MERGE (merge the current token and the next token into a single one) and $SPLIT (split the current token into two new tokens)) and empirically learned (e.g., $REPLACE_cause, which replaces certain words with "cause," and $APPEND_for, which adds "for" when it is needed), resulting in a total tag vocabulary size of 5000.

We train our tag-based GEC model with a multi-stage strategy using the same model architecture and pre-processing script as (Omelianchuk et al., 2020). We use the same synthesis strategy as in (Parnow et al., 2020) to synthesize pseudo data for pre-training in the first stage before fine-tuning on a small, high-quality human-annotated GEC dataset.

2.5 Feature #4: Multi-style Translation

In linguistics, the "style" of a text denotes "the aggregate of contextual probabilities of its linguistic items" (Enkvist, 1964) and can be seen as referring to its deviation from textual norms (Huang, 2015). Machine translation requires generating translated text with different styles, leading to what are known as as domain adaptation tasks (Koehn and Knowles, 2017). In these tasks, there are two main approaches (the data-centric approach and the model-centric approach), but though these approaches produce more powerful in-domain models (i.e., domain-specific models) for their given domains, they bring extra overhead to deployment.

Recently, large-scale Transformer-based language models have shown promising text generation capabilities, as seen with GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), which demonstrated strong generation performance with the Transformer decoder. (Keskar et al., 2019) sought to make a more malleable model and released CTRL, a 1.63 billion-parameter conditional Transformer language model, demonstrating that with enough model capacity, and compute power, language models can adapt to and be successful in multiple domains. Inspired by CTRL's use of control codes, which governed the style and other apsects of its generation, and GPT's use of Transformer decoders, we made a simple modification to the decoder of a Transformer-based NMT model, making this decoder also conditioned on a variety of control codes (Pfaff, 1979; Poplack, 2000). We call our system CTRL-NMT. Formally speaking, the target distribution of CTRL-NMT can be decomposed using the chain rule of probability and trained with a loss that takes the control code into account:

$$p(y|x) = \overbrace{\prod_{i=1}^{n} p(y_i|y_{<i}, x)}^{\text{NMT}} \rightarrow \overbrace{\prod_{i=1}^{n} p(y_i|y_{<i}, x, c)}^{\text{CTRL-NMT}},$$

where x is the source language input, y is the target language translation, and c is the control code.

In CTRL-NMT, the control code uses natural language terms (words) instead of separately defined tokens, so it can share the word embedding and has the ability to continue to expand to more codes. There is little change to the model in comparison to our standard NMT model, so CTRL-NMT can be initialized with the checkpoint of our standard NMT model. Additionally, since we only use a single model, deploying multiple styles will not be more costly.

2.6 Feature #5: Crowdsourcing Data Collection

In machine translation, grammatical error correction, and Semantic Similarity calculation, high-performing models rely on large-scale data, particularly high-quality, manually labeled data. Producing large scale annotated data is an onerous task requiring intensive human effort. This is especially true for machine translation, which requires bilingual speakers. “Crowdsourcing" (Howe, 2006) refers to a data collection method that involves obtaining work, information, or opinions from a large group of people who typically submit their data via internet services. Our MiSS system adopts crowdsourcing data collection as a method of further improving model performance, making MiSS an active learning system.

Specifically, when a user begins to input a sentence, the system responds with translation, back-translation, and revision suggestions. The user’s decisions in response to these suggestions will then constitute the data that we collect.

Operation	NMT	GEC	SE
Acc. Trans.	✓		✓
Edit Trans.	✓	✓	✓
Edit Source		✓	
Acc. Trans.Rv.	✓	✓	✓
Acc. Source.Rv.		✓	

Table 1: User operations used for our crowdsourcing data collection in our MiSS system.

3 Implementation and Training

The full system consists of 4 neural models: (1) a multi-style NMT model, (2) a simultaneous NMT model, (3) a grammatical error correction model, and (4) a BERT model. In our current MiSS release, we translate between three languages (English (EN), Chinse (ZH), and Japanese (JA)) for demonstration.

For the multi-style NMT model, we implement CTRL-NMT using the public fairseq (Ott et al., 2019) toolkit. In our system, we adopt the Transformer (big) setting as in (Vaswani et al., 2017). We did not choose a deeper or wider Transformer (Wang et al., 2019; Sun et al., 2019) model because we wanted to balance performance and efficiency. As in (Li et al., 2019), we used a data-dependent gaussian prior objective (D2GPo) during the NMT model training process for better generalization. Due to resource constraints, our currently deployed model does not perform back-translation of larger sentences. Table 2 lists all our training corpora and their sizes.

For the simultaneous translation model, we implemented the *wait-k* strategy and replaced the bidirectional attention in the encoder side with unidirectional attention. We also used the Transformer model implemented by fairseq as a base for this. Inspired by (Wu et al., 2020), we used beam search for partial tokens during simultaneous translation to obtain better translation sequences. We wanted to emphasize efficient inference, so we adopted a Transformer (Base) setting with fewer parameters. The training data used was the same as that in the multi-style NMT model.

	Provider	Style	Num.
EN-ZH	WMT20 AI Challenger18	Formal Oral	28.3M 12.9M
EN-JA	WMT20 TED+BSD	Formal Oral	17.7M 0.25M

Table 2: All NMT training data

We formulated the GEC task as a sequence labeling problem and thus adopted a neural sequence tagging model to handle the task. We followed (Omelianchuk et al., 2020)’s model architecture, which was an encoder consisting of a pre-trained BERT-like transformer stacked with two linear layers with softmax layers on the top - one for error detection and one for error labeling. As in (Awasthi et al., 2019), the architecture uses an iterative correction strategy in which predicted transformations are applied to the input sequence successively. After errors are detected and predicted, a modified Levenshtein distance guides the generation of a corrected sentence. We limit the maximum number of inference iterations to 4 to speed up the overall correction process while still maintaining good correction accuracy. The training data we used for GEC is shown in Table 3. We trained our English GEC model at the word level and our Chinese and Japanese models at the character level. We used pre-trained language models for initialization; namely, XLNet-large-cased in English, BERT-base-chinese in Chinese, and BERT-base-japanese-char in Japanese.

For translation quality evaluation, we measure the semantic equivalence using BERTScore, an automated evaluation metric that computes token similarity using contextual embeddings. We use RoBERTa-large, BERT-base-chinese, and BERT-base-japanese-char as the respective initial embedding sources for our English, Chinese, and Japanese evaluation models. As (Zhang et al., 2019, 2020) observed that fine-tuning the pre-trained con-

	Provider	Num.
EN	PIE-synthetic	9M
	Lang-8	947K
	NUCLE	56K
	FCE	34K
	W&I+LOCNESS	34K
ZH	NLPCC2018-GEC HSK+Lang8 CGED	1.3M
JA	Lang8	3.1M

Table 3: The GEC training data

Models	EN→ZH	ZH→EN	EN→JA	JA→EN
	separate training			
Transformer-big	37.6	28.0	33.5	18.7
	30.8	28.6	23.2	11.5
	joint training			
Transformer-big	36.9	27.2	33.4	18.5
	28.9	28.0	26.9	15.6
CTRL-NMT	37.5	28.4	33.8	19.2
	31.4	29.1	28.9	16.8
	joint training			
Transformer-base	35.4	25.8	31.7	17.0
	27.5	26.7	25.7	14.6
Sim-NMT (k=3)	31.1	23.3	30.2	16.1
	24.0	23.5	23.2	13.3

Table 4: The performance of our NMT models. Each model presents two lines of results - the top one for formal language and the bottom one for oral language translation.

textualized models on a related task can lead to better evaluation, we fine-tuned the pre-trained contextualized language models using our collected data.

4 Evaluation

We conducted empirical experiments on our models to evaluate the performance of important components in our system. For the NMT component, we chose the WMT2020 test set *newstest2020* as the evaluation set for formal EN-ZH and EN-JA translation and the development set of the AI Challenger 2018 competition as the evaluation set for oral ZH-EN translation. In ZH→EN and JA→EN translation, we used Multi-bleu as our evaluation metric, and we adopted the moses tokenizer for word tokenization, while in EN→ZH and EN→JA, we used character-level Multi-BLEU to remove the influence of different segmenters on BLEU score. For the standard and simultaneous machine translation components, we used the same evaluation sets and metrics.

	PrLM	Dict	P	R	$F_{0.5}$
EN	–	Word	53.46	37.45	54.22
	+XLNet	Word	76.92	41.03	65.47
ZH	–	Word	38.72	15.07	29.47
	–	Char	45.06	19.55	35.73
	+BERT	Char	50.34	33.46	45.72
JA	–	Word	36.83	20.52	31.78
	–	Char	45.68	16.49	33.74
	+BERT	Char	46.56	27.34	40.82

Table 5: The performance of our GEC models

For the GEC component, we followed common practice in the GEC task (Rei and Yannakoudakis, 2016; Omelianchuk et al., 2020) and used precision (P), recall (R), and $F_0.5$ to evaluate our models on all three languages. We evaluated English at the word level and Chinese and Japanese at the character level. We chose the test set of the CoNLL-2014 shared task as our evaluation set for our English GEC model. For Chinese and Japanese, we extracted 5000 sentences from the original training set for the development set and 5000 sentences for the test set and used the rest as the training set. ERRANT[6] was used to convert parallel files to the `m2` format for subsequent scoring with the M^2Scorer (Dahlmeier and Ng, 2012).

The results of our models for standard NMT and simultaneous NMT are shown in Table 4. First, for the evaluation results of standard NMT, we found that the joint training of multiple styles of data does not bring performance improvement compared to separate training, especially when the corpora sizes of the two styles are similar. The translation performance gap between different styles demonstrates that the level difficulty of translation in different styles is different. Since style essentially refers to deviation from standard textual norms, the greater the deviation, the greater the translation complexity is, which explains why different styles will have different levels of difficulty in comparison to standard translation.

In CTRL-NMT, through the incorporation control codes, we found that the translation performance for specific styles using the single model was equivalent to or, in some cases, better than that of training separate models. This shows that the Transformer-based model sufficiently accommodates the generation of multiple styles of language, and leveraging the language commonalities between different styles can bring additional im-

[6] https://github.com/chrisjbryant/errant

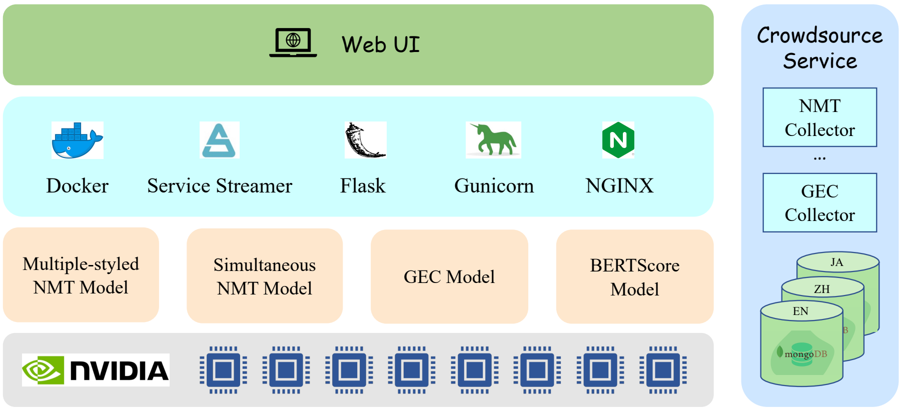


Figure 2: The deployment architecture of MiSS system.

provements.

The results of simultaneous NMT and standard NMT, however, do show that the performance of simultaneous NMT still lags behind that of standard NMT when using the same architecture, as there is less information available to the model during simultaneous translation. Despite this, simultaneous NMT is likely to further approach standard NMT's performance in the future through the use of greater contextual information and input prediction facilitated by a specific input module.

We show the evaluation results[7] of the GEC models in Table 5. The results show that pre-trained language models (PrLMs) can bring large performance improvements. Additionally, comparing Chinese and Japanese models at the word and character levels shows that in tag-based GEC modeling, character-level models outperform their word-level counterparts because of the character-level models' smaller tag sets.

5 Deployment

The architecture diagram of our deployment of the MISS system is shown in Figure 2. Since modern GPUs can bring good inference acceleration for deep neural network models, we choose NVIDIA GPUs as the basis for model deployment. There are four models in the system: the multi-style NMT model, the simultaneous NMT model, the GEC model, and the BERTScore model. We use Docker to install and isolate the environments of each model and use service_streamer to assemble scattered user requests to form a mini-batch to make full use of the GPUs in parallel. Flask and Gunicorn are used to wrap the model into a microservice interface for external calls. NGINX is used to distribute static resources and balance load. We use a basic Web UI to make our service accessible to users. In addition, Mongodb is adopted to store the users' logs, which the system collects.

6 Conclusion and Future Work

In this paper, we presented a translation system, MiSS. This system supports multi-style machine translation, simultaneous machine translation, grammatical error detection and correction, and back-translation-based quality evaluation. Our goal in developing this system is providing users with a more fluid machine translation experience. Using the research of the NLP community, we were able to introduce a variety of translation and translation-related tools to help users. In addition, we leverage the user's operations and feedback in the system as a source of crowdsourced information to potentially use in further improving the performance of the system. Compared with existing commercial translation systems, our system can provide a more comprehensive experience.

With this work, we also lay out steps to take to further improve the machine translation user experience: improve the consistency of translation by integrating document-level context, enhance the performance of models by incorporating back-translation using monolingual data, include more language styles such as academic translation, and explore the data collected through crowdsourcing for further improving overall performance.

[7] In our results, since the evaluation sets of Chinese and Japanese are self-split and character-level, they are not directly comparable to other work.

References

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Germán Sanchis Trilles, and Chara Tsoukala. 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, Gothenburg, Sweden. Association for Computational Linguistics.

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. 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)*, pages 4260–4270, Hong Kong, China. Association for Computational Linguistics.

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*.

Dorothée Behr. 2017. Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, 20(6):573–584.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Nils Erik Enkvist. 1964. *Linguistic and Style: On Defining Style; an Essay in Applied Linguistics; and Approach to the Study of Style*. Oxford University Press.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:539–555.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Jeff Howe. 2006. Crowdsourcing: A definition.

Libo Huang. 2015. Style in translation. In *Style in Translation: A Corpus-Based Perspective*, pages 17–30. Springer.

Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao. 2018. Moon IME: Neural-based Chinese Pinyin aided input method with customizable association. In *Proceedings of ACL 2018, System Demonstrations*, pages 140–145, Melbourne, Australia. Association for Computational Linguistics.

William John Hutchins and Harold L Somers. 1992. *An introduction to machine translation*, volume 362. Academic Press London.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018a. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018b. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411, Brussels, Belgium. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020a. Explicit sentence compression for neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8311–8318. AAAI Press.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2019. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2021. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq R. Joty, Thanh-Tung Nguyen, Kui Wu, and Ai Ti Aw. 2021. Cross-model back-translated distillation for unsupervised machine translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8073–8083. PMLR.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA â†’ Online. Association for Computational Linguistics.

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*.

Kevin Parnow, Zuchao Li, and Hai Zhao. 2020. Grammatical error correction: More data with more context. *The International Conference on Asian Language Processing (IALP)*.

Kevin Parnow, Zuchao Li, and Hai Zhao. 2021. Grammatical error correction as GAN-like sequence labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3284–3290, Online. Association for Computational Linguistics.

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

Carol W Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, pages 291–318.

Shana Poplack. 2000. Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of code-switching. *The bilingualism reader*, 18(2):221–256.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1181–1191, Berlin, Germany. 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, Hong Kong, China. Association for Computational Linguistics.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.

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.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Xueqing Wu, Yingce Xia, Lijun Wu, Shufang Xie, Weiqing Liu, Jiang Bian, Tao Qin, and Tie-Yan Liu. 2020. Learn to use future information in simultaneous translation. *arXiv preprint arXiv:2007.05290*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635. AAAI Press.