Improving Formality-Sensitive Machine Translation using Data-Centric Approaches and Prompt Engineering

Seungjun Lee[1], **Hyeonseok Moon**[1], **Chanjun Park**[1,2], **Heuiseok Lim**[1*]
[1]Korea University, South Korea
[2]Upstage, South Korea
{dzzy6505, glee889, bcj1210, limhseok}@korea.ac.kr
chanjun.park@upstage.ai

Abstract

In this paper, we present the KU x Upstage team's submission for the Special Task on Formality Control on Spoken Language Translation, which involves translating English into four languages with diverse grammatical formality markers. Our methodology comprises two primary components: 1) a language-specific data-driven approach, and 2) the generation of synthetic data through the employment of large-scale language models and empirically-grounded prompt engineering. By adapting methodologies and models to accommodate the unique linguistic properties of each language, we observe a notable enhancement in performance relative to the baseline, substantiating the heightened efficacy of data-driven approaches. Moreover, our devised prompt engineering strategy yields superior synthetic translation instances.

1 Introduction

Neural machine translation (NMT) models have achieved remarkable progress in recent years, as evidenced by their high BLEU scores (Britz et al., 2017; Stahlberg, 2020). Nonetheless, these models generally rely on generic parallel corpora and assume a single target translation for a given source sentence, often overlooking the significance of style and pragmatic aspects in translation, such as formality or politeness (Li et al., 2022). To address this issue, formality-sensitive machine translation (FSMT) has emerged as a research area, aiming to control grammatical formality in translated text across languages (Niu et al., 2017).

The Special Task on Formality Control on Spoken Language Translation introduces a new benchmark with high-quality training datasets for diverse languages, encompassing both supervised and zero-shot language pairs. Despite these new datasets (Nădejde et al., 2022), controlling formality in MT remains a challenging problem due to the absence of gold translations with alternate formality levels and the extensive variation in grammatical formality markers across languages.

Source: **It did**, many people **liked** his show so yeah, **do you like** Chris Pratt?
Korean Formal: **그랬어요**, 많은 사람들이 그의 쇼를 **좋아했죠**. 그래서 **당신** 크리스 프랫 **좋아해요**?
Korean Informal: **그랬어**, 많은 사람들이 그의 쇼를 **좋아했지**. 그래서 **너** 크리스 프랫 **좋아해**?

Table 1: Contrastive translations in formal and informal styles into Korean are presented. Grammatical formality markers, which are bolded, can be aligned through colors.

In the 2023 shared task, an English source segment is paired with two references that are minimally contrastive in grammatical formality, representing both formal and informal levels as shown in Table 1. Training and test samples are provided in the domains of "telephony data" and "topical chat" (Gopalakrishnan et al., 2019) for two supervised language pairs, English-Korean (EN-KO) and English-Vietnamese (EN-VI), and two zero-shot language pairs, English-Portuguese (EN-PT) and English-Russian (EN-RU). Grammatical formality markers differ across these languages. Personal pronouns and verb agreement signal formality in many Indo-European languages (e.g., PT, RU), while in Korean, formality control is notably challenging due to the widespread use of morphological markers to convey polite, respectful, and humble speech, making it an intriguing test case for FSMT.

In this paper, we present our approach to FSMT, focusing on the supervised setting for the English-Korean (EN-KO) and English-Vietnamese (EN-VI) language pairs and evaluating our methods on the zero-shot English-Portuguese (EN-PT) and English-Russian (EN-RU) pairs. Our method consists of two main strategies: 1) a language-specific data-driven approach, and 2) synthetic data gener-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 420–432

ation using large-scale language models and empirical prompt engineering. We apply techniques and models tailored to the linguistic features of each language. For Korean, we utilize a morpheme-centric subword tokenization method, while for Vietnamese, we employ a pre-trained EnViT5 model with high-quality Vietnamese parallel corpora. Additionally, we generate synthetic translation datasets for Portuguese and Russian using prompt engineering and refine these datasets using formality classifiers for fine-tuning our models. Furthermore, we founded significant performance improvements in EN-KO and EN-VI and conducted an ablation study to utilize high-quality synthetic examples.

2 Proposed Method

2.1 Task Definition

In this submission, we focus on the supervised and zero-shot settings on unconstrained formality control machine translation task. Formally, provided with a source segment $\boldsymbol{X} = \{x_1, x_2, \ldots, x_m\}$ and a formality level $l \in \{\text{formal, informal}\}$, the objective is to identify a model defined by parameters Θ that produces the most probable translation $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$ in accordance with the formality level:

$$\boldsymbol{Y} = \arg\max_{Y_l} P(\boldsymbol{X}, l; \boldsymbol{\Theta})$$

In simpler terms, the goal is to find the optimal model parameters Θ that produce the most likely translation $\boldsymbol{Y}$, given the source segment $\boldsymbol{X}$ and the desired formality level l (either formal or informal). This is achieved by maximizing the probability $P(\boldsymbol{X}, l; \boldsymbol{\Theta})$ of obtaining the translation $\boldsymbol{Y}$ at the specified formality level.

2.2 Language Specialized Data-Centric Approach

In this work, we employ a language specialized data-centric approach by integrating transfer learning techniques from Zoph et al. (2016) and language-specific subword methods, such as Unigram (Kudo, 2018) or byte-pair encoding (BPE) (Sennrich et al., 2015b). This combination effectively captures the unique morphological and syntactic structures of the target language, resulting in substantial improvements in translation performance, especially for low-resource languages (Zoph et al., 2016; Bojanowski et al., 2017; Park et al., 2020, 2021). Finally, we fine-tuned the pre-trained model (PLM) on the supervised train set each language pair.

EN-KO We discuss our approach to improve the English-Korean (EN-KO) translation performance by pre-training a Transformer using a high-quality dataset and leveraging morpheme-aware subword tokenization to better capture the linguistic characteristics of the Korean language such as agglutinative nature and structure.

We adopted a data-centric approach by pre-training a Transformer for EN-KO translation. To do so, we used a high-quality dataset from the AI Hub (Park et al., 2022)[1] data platform, which is operated by the Korean government. This comprehensive dataset includes various parallel corpora encompassing diverse domains such as technical and scientific fields, daily life and colloquial expressions, news articles, government and local government websites, publications, administrative regulations, Korean culture, and formal and informal language. By using a dataset specifically tailored for English-Korean translation, we aimed to capture finer nuances in both languages and enhance the translation quality by incorporating domain-specific knowledge and addressing the linguistic variations in different contexts.

Furthermore, we addressed the linguistic characteristics of the Korean language by applying a morpheme-aware subword tokenization method, which combines a segmentation strategy based on linguistic features with subwords. This approach has been shown to be effective in various Korean NLP tasks (Park et al., 2020). We utilized MeCab-ko [2], a widely-used morphological analyzer for the Korean language, for morpheme analysis. After obtaining the morphemes, we applied the Unigram subword tokenization method, which allowed our model to capture linguistic patterns specific to the Korean language, ultimately improving the overall translation performance.

EN-VI For the EN-VI language pair, we employed the EnViT5 (Ngo et al., 2022), a Text-to-Text Transfer Transformer (T5) model proposed by Raffel et al. (2020). We aimed to improve the fine-tuning translation performance of EN-VI in a low-resource setting by applying this data-centric approach to the multi-domain pre-trained EnViT5

[1] https://aihub.or.kr/

[2] https://bitbucket.org/eunjeon/mecab-ko-dic

model, which has been specifically designed for Vietnamese language tasks. Notably, EnViT5 models outperformed existing multilingual models such as mBART and M2M-100 while maintaining a significantly smaller parameter size, making them scalable and promising for both academic and industry applications (Ngo et al., 2022).

EnViT5 was pre-trained with the CC100 Dataset (Wenzek et al., 2020) which comprises monolingual data for over 100 languages. Subsequently, EnViT5 was fine-tuned on the MTet (Ngo et al., 2022) and PhoMT (Doan et al., 2021) datasets. MTet is a multi-domain EN-VI machine translation dataset encompassing a diverse range of domains, including educational videos, software user interfaces, COVID-related news articles, religious texts, subtitles, Wikipedia, and TED Talks (Reimers and Gurevych, 2020). Ultimately, when combined with PhoMT and IWSLT'15 (Cettolo et al., 2015), the final MTet dataset expands the training set size to 6 million examples, covering previously neglected areas such as law and biomedical data, which contains monolingual data for over 100 languages.

2.3 Synthetic Data Generation via Prompt Engineering

Leveraging synthetic examples in machine translation is crucial for improving translation quality, especially in low-resource settings (Edunov et al., 2018; Sennrich et al., 2015a). ChatGPT with GPT-4 engine (OpenAI, 2023), in particular, exhibits translation performance comparable to state-of-the-art WMT system and demonstrate good quality of generation conditioned translation generation in both few-shot and zero-shot settings (Hendy et al., 2023). To generate synthetic data, we employ ChatGPT to condition on formality and translate the IWSLT'22 Formality Track (Salesky et al., 2022) for all language pairs with English as the source language. Furthermore, we use a formality classifier (Rippeth et al., 2022) to filter synthetic examples, ensuring that both formal and informal examples are accurately translated for each language.

Supervised Setting We follow the prompt template depicted in Appendix A, which is based on the approach proposed by Hendy et al. (2023). To provide context for the model, we utilize n randomly selected shots from the English training set of other language pairs in the IWSLT 23 Formality Track (Agarwal et al., 2023). The few-shot examples are sourced from the target language's training set and include both informal and formal levels. ChatGPT is then tasked with translating the input text into either an informal or formal target language, depending on the specified prompt. For the input text, we use English source sentences from the IWSLT 22 Formality Track's other language pairs. After filtering the translated examples using a formality classifier, we fine-tuned the respective PLMs for EN-KO and EN-VI by incorporating synthetic examples into the training sets for each language pair. To verify the effectiveness of data augmentation through prompt engineering, we conduct experiments comparing the results with and without the augmented data.

Language	Size	
	Train	Test
EN-KO	400	600
EN-VI	400	600
EN-PT	0	600
EN-RU	0	600

Table 2: Data statistics in train and test sets of Formality Dataset

Zero-shot Setting In the EN-PT and EN-RU zero-shot settings, we generate synthetic examples for fine-tuning using the IWSLT'22 train set. We translate the source into both formal and informal target language levels, employing suitable prompts and filtering with a formality classifier to ensure conditioned formality. The template, shown in Appendix A, is adapted from the OpenAI Playground's default sentence-level translation task[3]. The model is instructed to translate English input into either informal or formal target language, guided by n random shots from the training set. Generated examples are then filtered using a formality classifier before fine-tuning the pre-trained multilingual translation model.

This zero-shot approach enables effective conditioned task performance with limited exposure to specific language pairs and formality levels. By generating synthetic translation data for fine-tuning, we capitalize on the model's generalization ability across languages and formality levels, enhancing translation performance in zero-shot settings. This highlights the potential of synthetic data in extending pre-trained language models' capabilities, even

[3] https://platform.openai.com/examples/default-translate

with novel language pair and formality combinations.

3 Experiment Settings

3.1 Dataset Details

The IWSLT shared task provides Formality Dataset which contains English source segments, each accompanied by two contrasting reference translations representing informal and formal formality levels. This is available for two language pairs, EN-{KO, VI}, in the supervised setting and two additional language pairs, EN-{PT, RU}, in the zero-shot setting. The statistics for the train and test sets of the dataset are shown in Table 2

For training and testing purposes, we randomly sampled 50 pairs of examples across each domain from the train set of Formality Dataset, and set them aside as validation sets (TASK DEV) for each supervised language. The remaining samples were utilized for training (TASK TRAIN).

Additionally, we utilized external datasets in conjunction with the data provided in the shared task. For EN-KO, we employed a parallel corpus comprising Formal/Informal, Social Science, Technology Science, and News domains from AI Hub for the pretraining of the PLM. For EN-VI, we utilized EnViT5, which was fine-tuned using the MTet (Ngo et al., 2022) and PhoMT (Doan et al., 2021) datasets.

In our research, we leverage ChatGPT for the augmentation of the EN-KO and EN-VI and the generation of synthetic examples for fine-tuning on EN-PT and EN-RU. This was done by using the source data from all available English-other language pairs (EN-XX) in the IWSLT'22 Formality Track (Anastasopoulos et al., 2022). To secure the quality and uniqueness of our training set, we implemented a preprocessing step that excludes duplicate sentences. Furthermore, to determine the optimal hyperparameters, we conducted a case study utilizing TASK DEV (details can be found in Section 4.3). The hyperparameters that led to the highest Matched-Accuracy (M-Acc) were selected for use. For all language pairs, we utilized a temperature of 0.9; specifically, we implemented 4-shot learning for EN-KO and 2-shot learning for EN-VI. For EN-PT and EN-RU, we proceeded with a zero-shot setting. More detailed information regarding the datasets and the preprocessing steps are presented in Table 3.

Language	Size	Source
EN-KO	6M	AI Hub (Formal/Informal + Tech/Sci + Social/Sci + News)
EN-VI	6.2M	MTet (Ngo et al., 2022) + PhoMT (Doan et al., 2021)
EN-{PT, RU}	1.6K	EN source from IWSLT'22 (Anastasopoulos et al., 2022)

Table 3: Additional external datasets used for the formality track in various language pairs.

3.2 Training Details

In the training details for the EN-KO language pair, we applied a morpheme-aware tokenization method to the translation dataset. To achieve this, we followed the training methods proposed by Park et al. (2020) and Gowda and May (2020), using `MeCab-ko` and Unigram to construct a vocabulary of 48K tokens. We then pre-trained the Transformer model (Vaswani et al., 2017). We used the `fairseq` library with 12 encoder and 12 decoder layers, each having 16 attention heads. Both encoder and decoder had an embedding dimension of 1024 and a feed-forward network (FFN) dimension of 4096. During pre-training, we trained for 20 epochs with a learning rate of 5e-4 and 4000 warmup updates. For fine-tuning, we trained for 200 epochs using a learning rate of 4e-5 and 100 warmup updates. We fine-tuned using the TASK TRAIN for all language pairs.

For EN-{VI, PT, RU} pairs, we fine-tuned using the `huggingface` library. For EN-VI, we used the `VietAI/envit5-translation` as the PLM. Fine-tuning was performed for 200 epochs with a learning rate of 4e-5, 200 warmup steps, and a batch size of 64. For EN-{PT,RU} pairs, we used `facebook/mbart-large-50` and trained for 200 epochs with a learning rate of 3e-5, 100 warmup steps, and a batch size of 16. All models were trained using four RTX A6000 GPUs. Detailed hyperparameters and training information can be found in the Appendix B.

3.3 Evaluation Details

In our experimental setting, we used the official test set from Formality Dataset (IWSLT'23) to evaluate our translation model's performance. The evaluation was conducted across two dimensions: overall translation quality and formality control. To assess the overall translation quality, we employed BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) (`eamt22-cometinho-da`) as au-

	Method	EN-KO BLEU	EN-KO COMET	EN-KO %M-Acc	EN-KO %C-F	EN-VI BLEU	EN-VI COMET	EN-VI %M-Acc	EN-VI %C-F
Formal	Official Baseline	4.91	0.211	78.3	98.6	26.71	0.363	96.0	99.7
	ChatGPT	5.65	0.524	83.3	**100.0**	27.07	0.510	**100.0**	98.0
	Ours	**26.60**	**0.727**	**87.0**	**100.0**	**47.00**	**0.669**	99.4	**100.0**
	Ours + Augmentation	17.09	0.667	79.4	99.5	41.57	0.653	99.4	99.7
Informal	Official Baseline	4.85	0.170	97.6	99.5	25.28	0.345	96.0	98.2
	ChatGPT	5.60	0.564	**100.0**	**100.0**	25.83	0.482	**100.0**	**100.0**
	Ours	**27.10**	**0.715**	98.0	95.0	**45.60**	**0.637**	98.8	**100.0**
	Ours + Augmentation	20.35	0.621	98.5	98.8	40.46	0.484	98.7	**100.0**

Table 4: Results on the test set of Formality Dataset for formal and informal supervised settings, obtained via our language specialized data-centric approach.

	Method	EN-PT BLEU	EN-PT COMET	EN-PT %M-Acc	EN-PT %C-F	EN-RU BLEU	EN-RU COMET	EN-RU %M-Acc	EN-RU %C-F
Formal	Official Baseline	27.29	0.448	96.3	97.7	21.96	0.349	96.2	92.0
	ChatGPT	**31.25**	**0.655**	92.0	96.0	**31.25**	0.655	92.0	96.0
	Ours	31.00	0.525	**100.0**	**100.0**	25.80	0.445	**100.0**	**100.0**
Informal	Official Baseline	**30.93**	0.416	**93.2**	**90.8**	21.63	0.348	84.1	85.2
	ChatGPT	27.38	**0.512**	48.4	46.0	**31.25**	**0.655**	92.0	**100.0**
	Ours	19.90	0.249	68.0	90.0	26.30	0.418	**100.0**	**100.0**

Table 5: Results on the test set of Formality Dataset for formal and informal zero-shot settings, achieved through our approach of synthetic data generation via prompt engineering.

tomatic evaluation metrics. We use 13A tokenizer to report SACREBLEU (Post, 2018) scores for all languages.

For formality control, we utilized Matched-Accuracy (M-Acc), a reference-based corpus-level metric that leverages phrase-level formality markers from the references to classify system-generated hypotheses as formal or informal. The corpus-level score is the percentage of system outputs that match the desired formality level.

Additionally, we used a reference-free variant of M-Acc (C-F) [4], which relies on a multilingual formality classifier to label system-generated hypotheses as formal or informal, with the corpus-level score representing the percentage of system outputs matching the desired formality level.

3.4 Prompt Design

We conducted experiments using ChatGPT with GPT-4 engine with `langchain`[5]. For EN-KO and EN-VI language pairs, we used a supervised setting, while for EN-PT and EN-RU pairs, we employed a zero-shot setting. In the supervised setting, we extracted arbitrary n-shot samples using the TASK TRAIN. We designed prompts by leveraging langchain's prompt guide and prompt examples from Hendy et al. (2023). Detailed examples and explanations of the prompts can be found in Appendix A.

4 Result & Findings

4.1 Results for Supervised Setting

Table 4 presents our experimental results in the supervised setting. As demonstrated by our results, our model, trained with the high-quality human-annotated Formality Dataset, exhibited outstanding performance. In particular, with respect to the C-F metric, our model shows almost perfect formality control performance (100% accuracy) for most of the tasks, except for the EN-KO informal task. Additionally, our model shows superior performance for the conventional NMT metrics (*i.e.* BLEU, COMET), outperforming ChatGPT with a 21.50 BLEU score for the EN-KO informal task. The EN-VI pair also exhibits high NMT metric

[4] https://github.com/amazon-science/contrastive-controlled-mt/tree/main/IWSLT2023

[5] https://python.langchain.com/

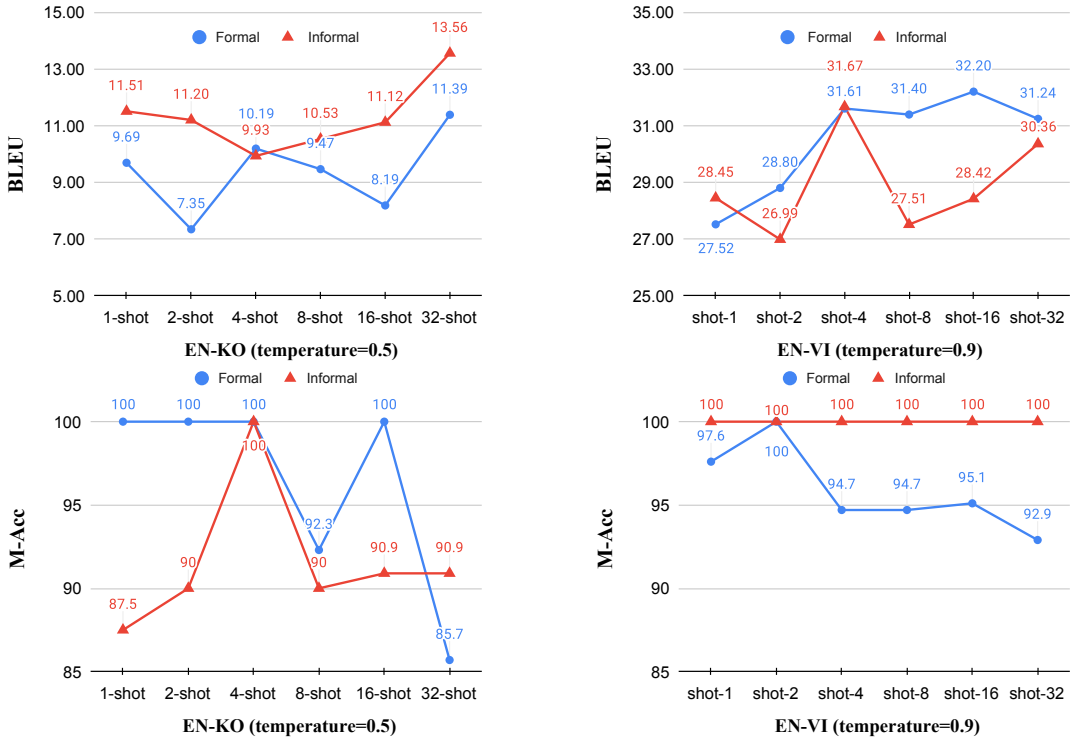

Figure 1: BLEU and M-Acc scores for ChatGPT based on superviesed setting, evaluated on TASK DEV.

scores, M-Acc, and C-F scores compared to the baseline. These results suggest that our language-specific data-centric approach is effective.

Through our experiments, we observed a significant degradation in the quality for supervised settings EN-{KO, VI}. This phenomenon can be attributed to the limitations of synthetic data produced by ChatGPT. While the data generated through ChatGPT exhibits considerable quality, it was not up to par with the sentences derived from our data-centric approach. We found that the integration of ChatGPT-augmented data inadvertently introduced noise into the system, leading to a decrease in overall performance. Despite the exceptional capabilities of ChatGPT, it appears that in this context, the quality of data augmented by conventional NMT methods is still superior. This observation further emphasizes the critical role of data quality over quantity in supervised learning environments, and highlights the potential benefits of more sophisticated prompting techniques that consider formality control, such as stylistic or sentence endings, for improving overall performance.

4.2 Results for Zero-shot Setting

The experimental results for the zero-shot setting are shown in Table 5. As can be seen from the experimental results, our model significantly outperforms the official baseline on all tasks except the EN-PT informal task. Notably, our model demonstrates consistently higher performance in terms of C-F metric compared to ChatGPT, achieving 100% M-ACC and C-F in the majority of tasks.

Exceptionally for EN-PT informal task, the performance of our model is markedly subpar, and ChatGPT even fails to exceed the official baseline. We find this result is highly noteworthy, as it suggest that ChatGPT may generate semantically accurate and plausible data, while the formality can hardly be controlled, especially for the EN-PT language pair. In our experiments, we utilized the same prompt for both EN-PT and EN-RU language pairs, differing only in language specification. The disparity in results between these two language pair suggests that specialized techniques for controlling formality are required for each language pair. This issue can be partially attributed to a data bias in ChatGPT, indicating a potential training data bias concerning formality.

4.3 Case Study

Impact of In-context Shots In this section, we examine the changes in performance based on the number of few-shot samples used for in-context learning, particularly when employing prompt engineering for translation. Previous research sug-

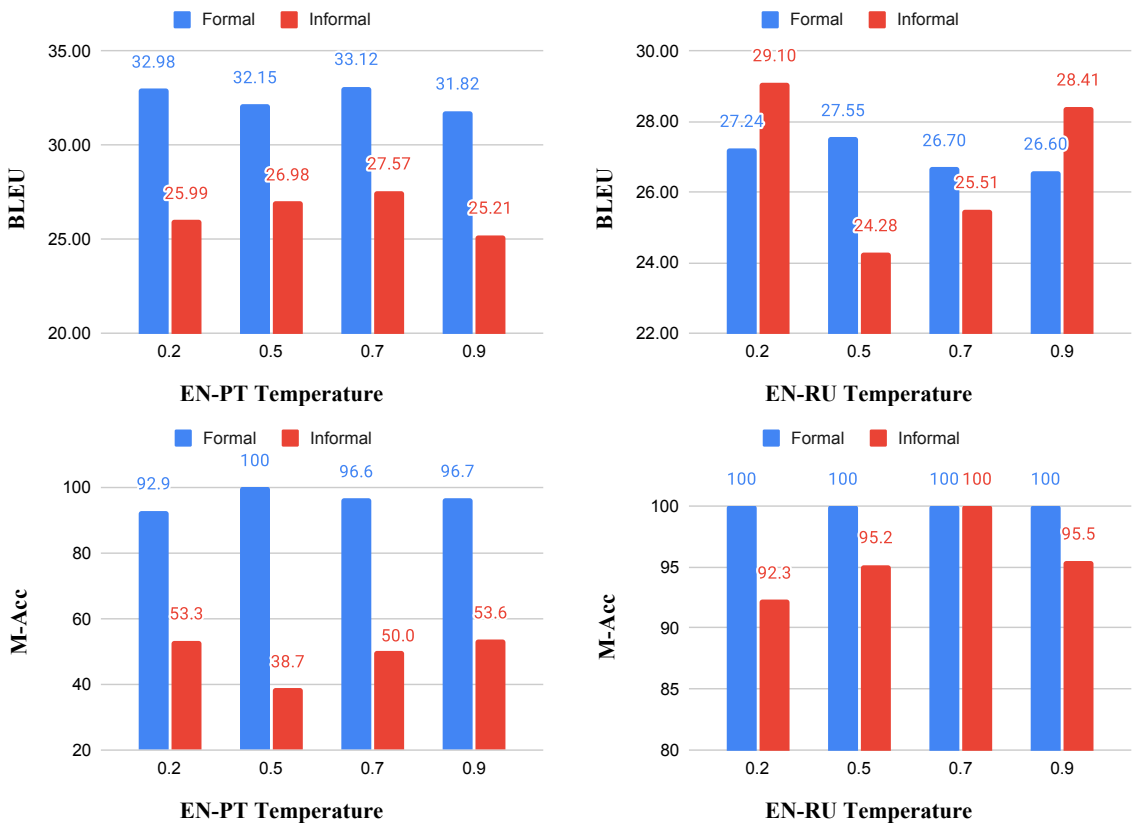

Figure 2: BLEU and M-Acc scores for ChatGPT based on zero-shot setting, evaluated on test set of Formality Dataset.

gests that increasing the number of shots beyond 10 does not significantly impact translation performance when using large language models (Zhang et al., 2023). However, we argue that applying the same perspective to formality control tasks proves challenging. This complexity arises as formality introduces a unique element required for these tasks. Additionally, previous research did not consider unintended consequences arising from this factor.

In pursuit of this, we conducted experiments where the number of shots was incrementally increased from 1 to 32, in powers of 2, using TASK DEV. The aim was to verify the differences in performance resulting from these changes. This process involved translating data via ChatGPT with an increasing number of shots and then evaluating the resulting translation data for its appropriateness. The experimental results are depicted in Figure 1. For this particular experiment, we selected one temperature (from the options of 0.2, 0.5, 0.7, 0.9) that demonstrated the highest performance and evaluated the changes in performance based on the number of shots.

As observed in our experimental results, increasing the number of shots for in-context learning led to an improvement in the general translation performance metric, BLEU. However, the scores of M-Acc and C-F, we found that the best performance was achieved with a smaller number of shots. This suggests that the nature of formality as a feature makes the "formality control" task distinct from conventional NMT, and it may be challenging to directly apply perspectives from conventional NMT to this task. We propose two hypotheses based on these results: (i) there exists a trade-off between translation performance and formality control as the number of shots increases, and (ii) increasing the number of shots while applying random sample selection may have caused confusion in performing formality control. We leave the analysis and validation of these hypotheses for future work.

Impact of Temperature Temperature is an important parameter to make ChatGPT generates varied responses to human queries (Peng et al., 2023). Basically, higher temperatures leads to the higher linguistic variety, while the lower one generates grammatically correct and deterministic text (Ippolito et al., 2019). Previous work suggested that for machine translation, a diverse generation may

impede its translation quality with a high degree of certainty(*i.e.* high temperature) (Peng et al., 2023). In this sense, we experiment with different temperature setting and find the optimal temperature for the formality control data augmentation. In our experiments, we select the most appropriate one among seven shot-candidates (1, 2, 4, 8, 16, 32) for each language pair.

Experimental results reveal that varying temperature can lead to significant performance fluctuations. It is particularly noteworthy that the performance disparity due to temperature changes is exceptionally high for the informal tasks. For formal tasks, the impact of temperature is relatively minor, with the variation in BLEU score is at most 0.95 (EN-RU). However, for informal tasks, the performance shift can reach up to 4.82 points (EN-RU) as temperature changes. Additionally, we find that in informal task, the performance variation depending on the temperature shows distinct trend for each language pair. This is evident from the fact that a moderate temperature(0.7) yielded the highest BLEU performance in the EN-PT informal task, while a similarly moderate temperature(0.5) resulted in the lowest performance. Our findings suggest that handling ChatGPT in informal task necessitates more elaborate control compared to dealing with formal data.

5 Background

In this work, we focus on data-centric approaches to improve Neural Machine Translation (NMT) performance. Several studies have investigated different strategies to address the challenges of low-resource languages and enhance translation quality. Kudo (2018) proposed subword regularization to improve NMT models using multiple subword candidates, effectively increasing data diversity and robustness. Gu et al. (2018) introduced a universal NMT model for extremely low-resource languages, leveraging multilingual knowledge from high-resource languages to assist in translation. Zoph et al. (2016) explored transfer learning for low-resource NMT, utilizing pre-trained models on related high-resource languages to improve the performance on the target low-resource language. Additionally, Sennrich et al. (2015a) proposed a method of improving NMT models by generating synthetic parallel data through back-translation, which has proven successful in various translation tasks. These studies highlight the diverse data-centric approaches in NMT, aiming to improve translation quality and overcome the limitations of low-resource languages.

6 Conclusion

In this paper, we presented the KU x UpStage team's submission for four languages, employing two main strategies: 1) a language-specific data-driven approach, and 2) synthetic data generation using large-scale language models and empirical prompt engineering. While our data-driven approach excelled, particularly in EN-KO and EN-VI, the quality of synthetic data generation was called into question. In light of this feedback, we propose to enhance the quality of synthetic data by integrating Quality Estimation (QE) techniques as an additional filter in the generation process. This step aims to further refine our synthetic examples, potentially improving the overall system performance. We also plan to explore the use of translation models with larger parameters and conduct a thorough analysis through more shot examples and linguistically-grounded data augmentation techniques. Finally, we aim to extend our understanding of factors influencing FSMT performance, such as the impact of formal register versus grammatical formality in training data and a detailed examination of zero-shot transfer.

Acknowledgments

This work was generously supported by multiple grants. The Core Research Institute Basic Science Research Program, funded by the Ministry of Education through the National Research Foundation of Korea (NRF) (Grant NRF-2021R1A6A1A03045425), provided valuable support. Additionally, this research received backing from the Information Technology Research Center (ITRC) support program (IITP-2023-2018-0-01405), which is supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) and funded by the Ministry of Science and ICT (MSIT), Korea. Finally, the Korea government's MSIT also funded a grant through the IITP (No. 2020-0-00368) dedicated to "A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques." We extend our gratitude for this comprehensive support.

Limitations

Due to the random sampling of shots, the results of the experiment may vary between repeated trials. However, we did not conduct repeated experiments under identical conditions, and thus we acknowledge the potential inconsistency of our experimental results.

Ethics Statement

This research study did not involve any human or animal subjects, and no personal data or sensitive information was used in this research. Therefore, no ethical issues were encountered in this study. The authors confirm that the research was conducted in accordance with the relevant ethical guidelines and principles.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Khalid Choukri, Alexandra Chronopoulou, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Benjamin Hsu, John Judge, Tom Ko, Rishu Kumar, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Matteo Negri, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Elijah Rippeth, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Mingxuan Wang, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–14, Da Nang, Vietnam.

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. *arXiv preprint arXiv:2004.02334*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Can Li, Wenbo Wang, Bitty Balducci, Lingshu Hu, Matthew Gordon, Detelina Marinova, and Yi Shang. 2022. Deep formality: Sentence formality prediction with deep learning. In *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 1–5. IEEE.

Maria Nădejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. Cocoa-mt: A dataset and benchmark for contrastive controlled mt with application to formality. *arXiv preprint arXiv:2205.04022*.

Chinh Ngo, Trieu H Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. Mtet: Multi-domain translation for english and vietnamese. *arXiv preprint arXiv:2210.05610*.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chanjun Park, Sugyeong Eo, Hyeonseok Moon, and Heui-Seok Lim. 2021. Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 97–104.

Chanjun Park, Midan Shim, Sugyeong Eo, Seolhwa Lee, Jaehyung Seo, Hyeonseok Moon, and Heuiseok Lim. 2022. Empirical analysis of parallel corpora and in-depth analysis using liwc. *Applied Sciences*, 12(11):5545.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various korean nlp tasks. *arXiv preprint arXiv:2010.02534*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

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, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling translation formality using pre-trained multilingual language models. *arXiv preprint arXiv:2205.06644*.

Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà, editors. 2022. *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, Dublin, Ireland (in-person and online).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A Prompt Template

A.1 Superviesd Setting

```
You are a helpful assistant that translates English to:
1. Informal [target language] or 2. Formal [target language]

####

[shot 1 source]

[shot 2 source]

[shot n source]


1. Informal [target language]: [shot 1 reference]

2. Formal [target language]: [shot 1 reference]


1. Informal [target language]: [shot 2 reference]

2. Formal [target language]: [shot 2 reference]


1. Informal [target language]: [shot n reference]

2. Formal [target language]: [shot n reference]

####

Translate this into only [1. Informal | 2. Formal] [target language]: [input]
```

Figure 3: Prompt template for supervised setting based on Hendy et al. (2023). We utilize n randomly selected shots from the English training set of other language pairs in the IWSLT 23 Formality Track as input for our model, with few-shot examples derived from the target language's training set.

A.2 Zero-shot Setting

```
You are a helpful assistant that translates English to:
1. Informal [target language] or 2. Formal [target language]

[shot n source]

Translate this into only [1. Informal | 2. Formal] [target language]: [input]
```

Figure 4: Prompt template for zero-shot setting, following the recommended instruction and format for the default sentence-level translation task in OpenAI playground[6]. This consistency enables us to maximize the benefits of the instruction finetuning protocol. We use n random shots from the training set.

B Experimental Setup

B.1 EN-KO

In the experimental setup for the EN-KO language pair, we employed a Transformer architecture with shared decoder input-output embeddings. The model's parameters included 1024-dimensional embeddings for both encoder and decoder, 16 attention heads for each, and 12 layers for both encoder and decoder. We used the Adam optimizer with beta values (0.9, 0.98) and a learning rate of 5e-4 scheduled by an inverse square root scheduler with a 4000-step warm-up. To prevent overfitting, we applied a dropout rate of 0.3 and weight decay of 0.0001. Our translation task utilized a label-smoothed cross-entropy criterion with a label smoothing factor of 0.1. The training process was performed with a maximum token limit of 4096 per batch and an update frequency of 4. Model performance was evaluated using BLEU scores with a beam size of 1 and detokenization using the Moses tokenizer. The training process was executed for a maximum of 20 epochs with a log interval of 200 and without epoch checkpoints, while sharing all embeddings.

Parameters for pre-training:

```
fairseq-train \
    --fp16 \
    --fp16-init-scale 4096 \
    --arch transformer --share-decoder-input-output-embed \
    --encoder-embed-dim 1024 --decoder-embed-dim 1024 \
    --encoder-attention-heads 16 --decoder-attention-heads 16 \
    --encoder-ffn-embed-dim 4096 --decoder-ffn-embed-dim 4096 \
    --encoder-normalize-before --decoder-normalize-before \
    --encoder-layers 12 --decoder-layers 12 \
    --optimizer adam --adam-betas '(0.9, 0.98)' --clip-norm 0.0 \
    --lr 5e-4 --lr-scheduler inverse_sqrt --warmup-updates 4000 \
    --dropout 0.3 --weight-decay 0.0001 \
    --task translation \
    --criterion label_smoothed_cross_entropy --label-smoothing 0.1 \
    --max-tokens 4096 \
    --update-freq 4 \
    --eval-bleu \
    --eval-bleu-args '{"beam": 1, "max_len_a": 1.2, "max_len_b": 10}' \
    --eval-bleu-detok moses \
    --eval-bleu-remove-bpe \
    --best-checkpoint-metric bleu --maximize-best-checkpoint-metric \
    --log-interval 200 \
    --max-epoch 20 \
    --skip-invalid-size-inputs-valid-test \
    --no-epoch-checkpoints \
    --share-all-embeddings
```

Parameters for fine-tuning:

```
fairseq-train \
    --batch-size 32 \
    --lr 4e-5  --warmup-updates 200 \
    --max-epoch 200 \
    --restore-file $MODELDIR/checkpoint_best.pt \
    --reset-optimizer --reset-meters --reset-dataloader --reset-lr-scheduler
```

B.2 EN-VI

We fine-tuned our model using the Hugging Face library and the code available at their repository[7]. The fine-tuning was performed with a learning rate of 4e-5, Adam optimizer with beta1 and beta2 values set to 0.9 and 0.98, respectively, and a weight decay of 0.0001. We also used mixed precision training (fp16) to accelerate the process. The learning rate scheduler was set to inverse square root with a warm-up of 200 steps. The training was conducted for 200 epochs with a maximum gradient norm of 0.0, label smoothing factor of 0.1, and a batch size of 64 for both training and evaluation. The model was saved and evaluated at the end of each epoch, and the logging was performed after each training step.

[7] https://github.com/huggingface/transformers/tree/main/examples/pytorch/translation

Parameters for fine-tuning:

```
python train_mt_trainer.py \
    --fp16 \
    --model_name_or_path VietAI/envit5-translation \
    --do_train \
    --do_eval \
    --do_predict \
    --source_lang en \
    --target_lang vi \
    --source_prefix "translate English to Vietnamese: " \
    --learning_rate 4e-5 \
    --adam_beta1 0.9 \
    --adam_beta2 0.98 \
    --max_grad_norm 0.0 \
    --num_train_epochs 200 \
    --lr_scheduler_type inverse_sqrt \
    --warmup_steps 200 \
    --weight_decay 0.0001 \
    --label_smoothing_factor 0.1 \
    --save_strategy epoch \
    --logging_steps 1 \
    --evaluation_strategy epoch \
    --per_device_train_batch_size=64 \
    --per_device_eval_batch_size=64
```

B.3 EN-{PT, RU}

We utilized the same training code as for the EN-VI task and employed the `facebook/mbart-large-50` model.

Parameters for fine-tuning:

```
export langs=ar_AR,cs_CZ,de_DE,en_XX,es_XX,et_EE,fi_FI,fr_XX,gu_IN,hi_IN,
it_IT,ja_XX,kk_KZ,ko_KR,lt_LT,lv_LV,my_MM,ne_NP,nl_XX,ro_RO,ru_RU,si_LK,
tr_TR,vi_VN,zh_CN

python train_mt_trainer.py \
    --fp16 \
    --model_name_or_path facebook/mbart-large-50 \
    --do_train \
    --do_eval \
    --do_predict \
    --source_lang en_XX \
    --target_lang pt_XX \
    --learning_rate 3e-5 \
    --adam_beta1 0.9 \
    --adam_beta2 0.98 \
    --max_grad_norm 0.0 \
    --num_train_epochs 200 \
    --lr_scheduler_type inverse_sqrt \
    --warmup_steps 100 \
    --weight_decay 0.0001 \
    --label_smoothing_factor 0.1 \
    --save_strategy epoch \
    --logging_steps 1 \
    --evaluation_strategy epoch \
    --per_device_train_batch_size=16 \
    --per_device_eval_batch_size=16
```