MAX-ISI System at WMT23 Discourse-Level Literary Translation Task

Li An* **Linghao Jin*** **Xuezhe Ma**
University of Southern California, Information Sciences Institute
{lan72605, linghaoj}@usc.edu
xuezhema@isi.edu

Abstract

This paper describes MaxLab - Information Sciences Institute (MAX-ISI) Translation systems for the WMT23 shared task. We participated in the discourse-level literary translation task - constrained track. In our methodology, we conduct a comparative analysis between the conventional Transformer model and the recently introduced MEGA model, which exhibits enhanced capabilities in modeling long-range sequences compared to the traditional Transformers. To explore whether language models can more effectively harness document-level context using paragraph-level data, we took the approach of aggregating sentences into paragraphs from the original literary dataset provided by the organizers. This paragraph-level data was utilized in both the Transformer and MEGA models. To ensure a fair comparison across all systems, we employed a sentence-alignment strategy to reverse our translation results from the paragraph-level back to the sentence-level alignment. Finally, our evaluation process encompasses sentence-level metrics such as BLEU, as well as two document-level metrics: d-BLEU and BlonDe.

1 Introduction

This paper introduces our submissions to the WMT23 Shared Task: Discourse-Level Literary Translation (Zh-En), Constrained Track. Our submission comprises three translation systems: a primary system employing a paragraph-level transformer, a first contrastive system utilizing a sentence-level transformer, and a paragraph-level Mega model as the second contrastive system.

Until very recently, the predominant focus of context-aware Neural Machine Translation (NMT) research has been on parallel datasets that align at the sentence level, such as IWSLT17 (Cettolo et al., 2017) and OPUS (Tiedemann, 2012). More recent research endeavors have concentrated on literary translation, which is typically more intricate and requires the models to be able to capture long-range context for high-quality translations. For example, Thai et al. (2022) introduced the first multilingual paragraph-aligned dataset PAR3, sourced from public-domain non-English literary works.

We use Transformer as the baseline model. In order to assess whether a more advanced model can excel in modeling long-range sequences using literary data, which contains richer contextual information, we also include the MEGA (Ma et al., 2023) model for comparison. The foundational model architectures we employ are introduced in Section 2.

In Section 3, we provide an extensive explanation of our systems. Within this section, Section 3.1 outlines the data pre-processing step. In this phase, we construct both sentence-level data, which comprises the filtered original data, as well as paragraph-level data. It's worth noting that aligning sentences in literary translation is not always feasible due to the possibility of sentence merging or truncation during the translation process. At the paragraph level, language models can adeptly exploit document-level context, resulting in a reduction of translation errors at the discourse level, as corroborated by human evaluations (Karpinska and Iyyer, 2023). Building on these encouraging findings, we created a dataset aligned at the paragraph level by aggregating multiple sentences from the provided literary dataset. Then, we propose three systems and evaluate those systems with both sentence-level and document-level metrics.

Section 4 presents the results that culminate in our final submissions. Additionally, we discussed challenges we encountered regarding discourse-level translation in Section 5.

*Equal contribution.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 282–286
December 6–7, 2023.

2 Model Architectures

We select the following two model architectures for our systems, taking into account their strong performance in the context of context-aware machine translation.

Transformer The Transformer architecture, as introduced by Vaswani et al. (2017), utilizes an encoder-decoder framework, leveraging a self-attention mechanism. This mechanism enables each position within a given sequence to interact with every other position, facilitating the computation of a comprehensive representation for the entire sequence.

In all our experiments, we employ the Transformer `base` version which consists of 6 encoder layers, 6 decoder layers, a model dimension of 512, and a FFN hidden dimension of 2048.

MEGA The recently unveiled MEGA (Moving Average Equipped Gated Attention) (Ma et al., 2023), addresses two longstanding limitations of the conventional Transformer model, which have impeded its performance on tasks involving long sequences. These limitations pertain to a weak inductive bias and a quadratic computational complexity.

MEGA employs a multi-dimensional, damped exponential moving average (Hunter, 1986) (EMA) in conjunction with a single-head gated attention mechanism to preserve inductive biases. Importantly, MEGA can replace the attention mechanism within the Transformer framework. Additionally, MEGA is of comparable size to the Transformer.

In total, the Transformer architecture is around 75M parameters; the MEGA architecture is around 77M parameters.

3 System Overview

3.1 Data Preprocessing

We first perform the following filtering steps on the training data:

- Remove translators' notes.
- Merge dialogues with tags "#<#" and "#>#" into one instance.
- Combine blank lines with their following line.

We construct sent-level and paragraph-level datasets separately.

Sentence-level dataset is constructed using the sentence alignment information, which is thoughtfully provided.

Paragraph-level dataset Considering the critical role played by context, particularly in literary translation, we further construct a paragraph-aligned corpus. This corpus is established based on the sentence alignment, allowing us to leverage context more effectively in our translations.

Data for each language pair is then encoded and vectorized with byte-pair encoding (Sennrich et al., 2016) using the `SentencePiece` (Kudo and Richardson, 2018) framework. We use separate vocabularies of size 32K for each language Zh and En.

Full corpus statistics are in Table 1.

Subset	Sent-level	Paragraph-level
Train	1742150	290315
Valid1	711	154
Valid2	810	148

Table 1: Instance counts across train and valid subsets.

3.2 System Architectures

Transformer-256 Our primary system employs a `Transformer-base` model at the paragraph level. Prior to tokenization, we structured the data into paragraphs, each with a maximum length of 256 characters on the source side (Zh). The model is subsequently trained and utilized for decoding on the paragraph-aligned corpus mentioned above.

Transformer-Sent In contrast, we conduct training for the transformer-base model using the sentence-level corpus.

MEGA-256 We adopted our proposed paragraph-aligned data as it demonstrates competitive efficacy in comparison to conventional Transformers across established benchmarks, including the LRA dataset, all while maintaining a significantly leaner parameter configuration.

3.3 Training

We train all models on the `fairseq` framework (Ott et al., 2019). All models were trained on 4 NVIDIA A40 GPUs. Following Vaswani et al. (2017); Fernandes et al. (2021), we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a linear decay learning rate scheduler with an initial value of 10^{-4},

System	Subset	BLEU	d-BLEU	BlonDe				
				all	pron.	entity	tense	d.m.
Transformer-Sent	VALID1	**26.40**	**26.40**	37.87	74.40	36.95	69.98	67.22
	VALID2	**16.40**	16.10	29.89	67.34	**49.05**	70.76	52.78
Transformer-256	VALID1	21.90	26.20	**40.92**	**84.17**	**40.47**	**78.72**	**72.71**
	VALID2	13.60	**16.30**	**33.50**	**79.72**	46.44	**81.57**	**68.80**
MEGA-Sent	VALID1	25.00	25.00	37.03	73.55	36.32	68.81	66.55
	VALID2	16.20	15.80	29.54	67.18	47.30	69.95	54.21
MEGA-256	VALID1	22.40	23.90	39.74	81.14	**40.47**	77.29	71.47
	VALID2	13.20	15.80	32.90	77.37	48.17	81.13	66.80

Table 2: Automatic metric results on the valid1 and valid2 sets. All reported BlonDe scores are F1s; pron. stands for pronoun, d.m. stands for discourse marker.

System	Sent-Level				Doc-Level	Human Annotator
	BLEU	chrF	COMET	TER	d-BLEU	Average
Transformer-256	34.1	53.3	78.24	**62.4**	**45.1**	73.59
Transformer-Sent	**34.5**	**54.7**	**79.14**	62.7	44.9	✗
MEGA-256	33.1	52.4	77.84	63.6	44.4	✗

Table 3: Automatic metric results of our submissions on the test set and the average score by different annotators on one sampled document. (Wang et al., 2023).

and increasing to $5e^{-4}$ during a warm-up phase of 4000, and a dropout of 0.2. We run inference on the validation set and save the checkpoint with the best BLEU score.

3.4 Post-processing

Since the final submission requires that each line must be aligned with the corresponding input line in the output files, we add this post-processing step to reverse our paragraph-level translation result to sentence-level alignment. We will discuss this further in the conclusion part.

Sentence-Alignment

1. Use the translated results at the sentence level as a reference

2. Calculated the similarity between each sentence in the translated paragraph and the *M* nearest sentences in the sentence-level translation

3. Align each sentence to the most similar one using Jaccard similarity on N-gram overlap as the similarity metric

3.5 Evaluation

To evaluate the discourse-level translation ability of three systems, we compute three metrics:

BLEU (Papineni et al., 2002) sentence-level BLEU is the most commonly used metric to evaluate the quality of machine-generated translations. We report the standard BLEU score calculated using `sacreBLEU` (Post, 2018)[1] in our systems.

d-BLEU (Liu et al., 2020) document-level sacreBLEU is computed by matching n-grams in the whole document. Note that all evaluations are case-sensitive

BlonDe (Jiang et al., 2022) is introduced as a document-level automatic metric that calculates the similarity-based F1 measure of discourse-related spans across four categories (*pronoun, entity, tense and discourse marker*).

4 Results

The results of our experiments are presented in Table 2. We evaluate our models on the provided two validation sets and list model performances

[1]The sacreBLEU signature is BLEU+case.mixed+lang.src-tgt+numrefs.1+smooth.exp+{test-set}+tok.13a.

on three automatic metrics, i.e., BLEU, d-BLEU, and BlonDe. Given that BLEU scores compare n-grams on a sentence-level basis, we extend our evaluation to encompass d-BLEU and BlonDe metrics, providing a comprehensive assessment of the models' proficiency in discourse-level translation. The results of the test set are presented in Table 3.

Transformer *vs.* MEGA As per the outcomes presented in Table 2, Transformer models slightly surpass MEGA models in both sentence-level and paragraph-level translations. While MEGA demonstrates superior capabilities in long-range sequence modeling, its limited enhancement may be attributed to the fact that current data are not lengthy and doesn't capture sufficient useful context (Jin et al., 2023). Furthermore, the discrepancy in BLEU scores is more pronounced than the variation in BlonDe scores.

Sent-level *vs.* Paragraph-level Based on the results presented in Table 2, there is a discrepancy between BLEU and BlonDe evaluations. Specifically, it is observed that sentence-level translation exhibits a better performance in terms of the BLEU metric, whereas paragraph-level models demonstrate a substantial improvement when assessed using the BlonDe metric.

As delving into the four distinct categories in BlonDe, a consistent trend of enhancement emerges across each category with the adoption of paragraph-level translation. Particularly, marked improvements are observed within the pronoun and tense categories. This can be attributed to the inherent reliance of pronouns and tenses on contextual information. These empirical results demonstrate that paragraph-level data provides more useful contextual signals than sentence-level data.

5 Discussion

Limitation of sentence alignment Literary texts often rely on context that spans beyond individual sentences, making strict sentence alignment impractical. As evidenced in our results, paragraph-level translation excels in preserving contextual information, like pronouns and tenses. However, the insistence on maintaining sentence-level alignment imposes constraints on model selection, hindering flexibility and adaptability.

Limitation of evaluation metrics The current evaluation metrics are not capable enough of measuring document-level machine translation. The most commonly used metric, BLEU, and its variant, d-BLEU, may struggle to fully capture the context awareness and coherence that is crucial at the document level translation.

6 Conclusion

This paper describes the submission to the WMT23 literary translation shared task - constrained track. We compare traditional Transformer models to the newer MEGA model, integrating paragraph-level data into both. Transformer models outperform MEGA in both sentence and paragraph translation on this literary dataset. We observe a discrepancy between BLEU and BlonDe evaluations, with the latter favoring paragraph-level translation. These results emphasize the challenges of document-level translation and the importance of more context-aware evaluation metrics.

References

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine 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 6467–6478, Online. Association for Computational Linguistics.

J. Stuart Hunter. 1986. The exponentially weighted moving average. In *Journal of Quality Technology*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.