Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, Yulin Yuan
vincentwang0229@gmail.com

Abstract

Following last year's WMT, we (Tencent AI Lab and China Literature Ltd.) have continued to host literary translation shared task (Wang et al., 2023) this year, the second edition of the *Discourse-Level Literary Translation*.

First, we (Tencent AI Lab and China Literature Ltd.) release a copyrighted and document-level Chinese-English web novel corpus. Furthermore, we put forth an industry-endorsed criteria to guide human evaluation process. This year, we totally received 10 submissions from 5 academia and industry teams. We employ both automatic and human evaluations to measure the performance of the submitted systems. The official ranking of the systems is based on the overall human judgments. In addition, our extensive analysis reveals a series of interesting findings on literary and discourse-aware MT. We release data, system outputs, and leaderboard at https://www2.statmt.org/wmt24/literary-translation-task.html.

1 Introduction

With the swift progression of MT and the notable advancements in Large Language Models (LLM) (**??**), our curiosity is piqued regarding the efficacy of MT and LLM in the realm of literary translation. We aim to explore the extent to which these technologies can aid in addressing the intricate challenges of translating literary works. Therefore, we hold the first edition of the *Discourse-Level Literary Translation* in WMT 2023. Literary texts encompass a wide range of forms, including novels, short stories, poetry, plays, essays, and more. Among these, *web novels*, also known as online or internet novels, represent a unique and rapidly growing subset of literature. Their popularity, accessibility, and diverse genres set them apart. As they provide not only an extensive volume of text but also exhibit distinctive linguistic features, cultural phenomena, and simulations of societies, web novels can serve as valuable resources and challenging for MT research.

Limitations

We discuss the potential limitations of this edition of the shared task as follows:

- *Language Pair*. This year, we only focus on Chinese→English direction. However, we have a long-term plan to continuously organize this task, and will extend the copyrighted dataset into Chinese-Russian and Chinese-German language pairs next year.
- *Literary Genre*. This year, we mainly used the Web Fiction Corpus which is only one type of literary text. We use Web Fiction for two reasons: (1) its literariness is less complicated than others (e.g. poetry, masterpiece); (2) such bilingual data are numerous and continuously increased. We will consider to extend more literary genres such as poetic translation in the next year.
- *Discourse Benchmark*. We have accumulated some discourse- and context-aware benchmarks (**???**). These benchmarks are pivotal for assessing the proficiency of LLMs in handling complex language structures and contextual nuances. As participation of LLM-based systems in our shared tasks increases, we anticipate integrating these benchmarks more comprehensively into our future evaluations to better measure and understand the evolution of LLM capabilities in linguistic context and discourse comprehension.

References

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: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine*

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 699–700

Type	System	Sent-Level				Doc-Level
		BLEU↑	chrF2↑	COMET↑	TER↓	d-BLEU↑
Baselines	Google⋆	37.4	57.0	80.50	57.4	47.3
	Llama-MT⋆	n/a	n/a	n/a	n/a	43.1
	GPT-4⋆	n/a	n/a	n/a	n/a	43.7
Primary	Cloudsheep⋆	39.5	57.5	81.22	55.5	48.5
	HW-TSC	40.5	58.5	82.61	56.0	50.2
	NLP2CT-UM⋆	**41.6**	**58.7**	**83.56**	**52.7**	**50.9**
	NTU⋆	20.9	41.9	74.53	73.9	34.6
	SJTU-LoveFiction⋆	35.1	54.7	80.79	62.1	47.2
Contrastive	HW-TSC	40.6	58.6	82.59	55.9	50.3
	NLP2CT-UM$^{\star}_{1}$	41.6	58.7	83.54	52.8	50.8
	NLP2CT-UM$^{\star}_{2}$	41.5	58.6	83.38	52.8	50.7
	SJTU-LoveFiction$^{\star}_{1}$	35.7	56.0	82.67	59.7	46.3
	SJTU-LoveFiction$^{\star}_{2}$	38.6	56.5	82.49	57.1	49.6

Table 1: Evaluation results of baseline and participants' systems in terms of **automatic evaluation methods**, including 1) sentence-level metrics BLEU, chrF, COMET, TER; and 2) document-level metrics d-BLEU. Systems marked with ⋆ are unconstrained, while others are constrained. The COMET is calculated with *unbabel-comet* using *Reference 1* while others are calculated with *sacrebleu* using two references. The best primary constrained and unconstrained systems are highlighted.

Translation, pages 55–67, Singapore. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.