Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs

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, Shuming Shi
vinnylywang@tencent.com

Abstract

Translating literary works has perennially stood as an elusive dream in machine translation (MT), a journey steeped in intricate challenges. To foster progress in this domain, we hold a new shared task at WMT 2023, the first 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 14 submissions from 7 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 `http://www2.statmt.org/wmt23/literary-translation-task.html`.

1 Introduction

In past decades, the evolution of machine translation (MT) has undergone significant improvements in accuracy and efficiency, leading to many practical applications in various fields (Bojar et al., 2014; Barrault et al., 2019; Farhad et al., 2021; Kocmi et al., 2022). Despite its success, MT still struggles in certain intricate scenarios to deliver translations that meet high standards (Läubli et al., 2018; Koehn and Knowles, 2017). Translating literary texts is considered to be the greatest challenge for MT due to its complex nature (Toral and Way, 2018; Toral et al., 2018; Ghazvininejad et al., 2018):

- *Rich Linguistic and Cultural Phenomena*: literary texts contain more complex linguistic and cultural knowledge than non-literary ones (Voigt and Jurafsky, 2012; Ghazvininejad et al., 2018). To generate a cohesive and coherent output, MT models require an understanding of the intended meaning and structure of the text at discourse level (Wang et al., 2016, 2018a,b, 2019, 2023b). Furthermore, it demands skillful adaptation of cultural references, idioms, and subtle expressions to capture the essence of the original work in target languages.
- *Limited Data*: existing document-level datasets are news articles and technical documents (Liu and Zhang, 2020; Thai et al., 2022); there is limited availability of copyrighted, discourse-level, parallel data in the literature domain. This makes it difficult to develop models that are able to handle the complexities of literary translation.
- *Long-Range Context*: literature such as novels have much longer contexts than texts in other domains (e.g. news articles). Translation models need to acquire the capacity of modeling long-range context for learning translation consistency and lexical choice (Wang et al., 2017; Wang, 2019; Matusov, 2019; Du et al., 2023).
- *Unreliable Evaluation Methods*: literary evaluation needs to measure the meaning and structure of the text, and the nuances and complexities of the source language. A single automatic evaluation using a single reference is unreliable. Thus, professional translators with well-defined error typologies and targeted automatic evaluation are considered a complement (Matusov, 2019).

With the swift progression of MT and the notable advancements in Large Language Models (LLM) (Ouyang et al., 2022b; OpenAI, 2023), 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

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 55–67
December 6–7, 2023.

Figure 1: The word cloud represents institute and companies from different regions that downloaded the GuoFeng Webnovel Corpus.

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. This year, the shared task mainly focuses on *document-level web novels*, and we introduce a document-level benchmark dataset and establish human evaluation criteria specifically tailored to address the challenges of literary translation:

- **Benchmark Dataset**: We build and release a copyrighted and high-quality Chinese-English training corpus, comprising 2 million sentences sourced from 179 web fictions. This dataset preserves both book-level and chapter-level contexts, and features manually-aligned sentence pairs. We also provide three types of testsets, varying in distribution and document length (in Section 2).
- **Evaluation Methods**: In order to evaluate the translation quality of the participating systems we used both automatic and human evaluation methods. About automatic evaluation, we employ document-level sacreBLEU (d-BLEU) as our metric, which is computed by matching n-grams in the whole document (Liu et al., 2020; Post, 2018). In terms of human evaluation, we propose a well-defined criteria by adapting multidimensional quality metrics (MQM) (Lommel et al., 2014) to fit the context of literary translation. Note that all evaluations are case-sensitive (in Section 3).

Figure 2: Illustration of discourse-level literary translation, which is sampled from our Web Fiction Corpus. Colored words demonstrate rich linguistic phenomena.

We introduce the task overview and submission form in Section 4. This year, 14 submissions were received from 7 different teams, which are detailed in Section 5. We report the evaluation results in Section 6 followed by the conclusion in Section 7.

2 The GuoFeng Webnovel Corpus

We release a copyrighted and high-quality Chinese-English corpus on web novels. Additionally, we provide in-domain pretrained models as supplementary resources. As shown in Figure 1, a total of 45 institutes and companies from various regions have downloaded our dataset, showing that the prposed tasks and data have garnered widespread interest.

2.1 Datasets

Copyright Copyright is a crucial consideration when it comes to releasing literary texts, and it is also one of the primary reasons for limiting the scale of data in this domain. We, Tencent AI Lab and China Literature Ltd., are the copyright owners of the web fictions included in this dataset. In order to promote the advancement of research in this field, we make this data available to the research community, subject to certain terms and conditions.

- After registration, WMT participants can use the corpus for non-commercial research purposes and follow the principle of fair use (CC-BY).

- Modifying or redistributing the dataset is strictly prohibited.
- You should cite the this paper and claim the original download link.

Data Processing The web novels are originally written in Chinese by web novel writers and then translated into English by professional translators. Our data processing involves a combination of automated and manual techniques: 1) we match Chinese books with its English counterparts based on bilingual titles; 2) within each book, Chinese-English chapters are aligned using Chapter ID numbers; 3) within each chapter, we build a MT-based sentence aligner to align sentences in parallel, preserving the sentence order in the chapter; 4) human annotators are engaged to review and correct any discrepancies in sentence-level alignment. To ensure the retention of discourse information, we permit null alignments. We totally spent 6 months addressing copyright issues and around 40,000 euros for human annotation. Figure 2 shows the final format of our corpus.

Training/Validation/Testing Data Table 3 lists data statistics of our dataset. As seen, the *training set* contains 23K continuous chapters from 179 web novels, covering 14 genres such as fantasy science and romance. To enable participants to evaluate model performance by themselves, we provide two *unofficial validation/testing sets* with one reference. For dataset_1, books overlap with the training data, whereas dataset_2 contains unseen books. The participants can regard each chapter as a document to train and test their discourse-aware models. Apart from this, parallel training data in the General MT Task can also be used for data augmentation. In the final testing stage, participants use their systems to translate the *official testing set* (Test_{final}). We select around 20 consecutive chapters from each book. Thus, we participants could treat all chapters within a book as a long document[1]. As seen, the document length of Test_{final} is quite longer than other sets. The final testset contains two references: Reference 1 is translated by human translators and Reference 2 is bult by manually aligning bilingual text in web page. The genres in the valid and test sets are sampled evenly.

[1]The participants can still regard one chapter as a document, which depends on the models' length capability.

2.2 Pretrained Models

Apart from training dataset from web novels, we also provide in-domain pretrained models as supplementary resources. These models can be used to finetune or initialize MT models.

- **RoBERTa (base)**: The original model features a 12-layer encoder and is trained on the Chinese Wikipedia (Liu et al., 2019). It has a hidden size of 768 and a vocabulary size of 21,128 using whole word masking. We continuously train it with Chinese literary texts (84B tokens) (Wang et al., 2023a).
- **mBART (CC25)**: This original model is equipped with a 12-layer encoder and a 12-layer decoder, having been trained on a web corpus spanning 25 languages (Liu et al., 2020). It boasts a hidden size of 1024 and a vocabulary size of 250,000. We continuously train it with English and Chinese literary texts (114B tokens) (Wang et al., 2023a).

Besides, general-domain pretrained models listed in General MT Track are also allowed in this task: mBART, BERT, RoBERTa, sBERT, LaBSE.

3 Evaluation Methods

It is still an open question whether human and automatic evaluation metrics are complementary or mutually exclusive in measuring the document-level and literary translation quality. Thus, we report both automatic and human evaluation methods, and officially rank the systems based on the overall human judgments.

3.1 Automatic Evaluation

We use widely-used sentence- and document-level evaluation metrics: 1) *sentence-level*: we employ sacreBLEU (Post, 2018), chrF (Popović, 2015), TER (Snover et al., 2006) and pretraining-based COMET (Rei et al., 2020); 2) *document-level*: we mainly use document-level sacreBLEU (d-BLEU) (Liu et al., 2020), which is computed by matching n-grams in the whole document. For d-BLEU, We combine all sentences in each document as one line and then conduct sacreBLEU metric. Note that all evaluations are case-sensitive. We employ *sacrebleu*[2] to calculate sacreBLEU, chrF, TER and d-BLEU with *sacrebleu* using two references. The command is: `cat output | python -m`

[2]https://github.com/mjpost/sacrebleu with signature: `nrefs:2|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

Dataset	#Book	#Chap.	#Sent.	#Word	|D|
Train	179	22.6K	1.9M	32.0M	1.4K
Valid_1	22	22	755	18.3K	832
Test_1	26	22	697	19.5K	884
Valid_2	10	10	853	16.0K	1.6K
Test_2	12	12	917	16.7K	1.4K
Test_{final}	12	239	16.7K	337.0K	*28.1K

Figure 3: Data statistics of the GuoFeng Webnovel Corpus on number of book, chapter (#Chap.), sentence (#Sent.), word, and genre distribution in training set. The #Word is based on English texts. For dataset_1, books overlap with the training data, whereas dataset_2 contains unseen books. Thus, each chapter is treated as a separate document. For Test_{final}, around 20 consecutive chapters from each book are selected, treating all chapters within a book as a long document. The document length (|D|) is calculated by dividing #Word divided by the number of documents.

`sacrebleu reference*`. We employ *unbabel-comet*[3] to calculate COMET score using *Reference 1*. The command is: `comet-score -s input -t output -r reference1` (default model).

3.2 Human Evaluation

The human evaluation was performed by professional translators using an adaptation of the multidimensional quality metrics (MQM) framework (Lommel et al., 2014). For example, we consider the preservation of literary style and the overall coherence and cohesiveness of the translated texts. As shown in Table 6, we put forth an industry-endorsed criteria to guide human evaluation process. The main error types are:

- **Accuracy (Acc.)**: The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.
- **Fluency (Flu.)**: Issues related to the form or content of a text, irrespective as to whether it is a translation or not.
- **Style (Sty.)**: The text has stylistic problems.
- **Terminology (Ter.)**: A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified.
- **Locale Convention (Loc.)**: The text does not adhere to locale-specific mechanical conventions and violates requirements for the presentation of content in the target locale.
- **Others (Oth.)**: Other issues such as the signs of MT, gender bias and source errors.

MQM utilizes a scorecard format to quantify the quality assessment results. Evaluators assign numerical values to identified translation errors based on error types, severity, etc., making the assessment results more intuitive. The overall quality score is calculated based on per-word translation accuracy:

$$\text{S} = 1 - \frac{5 \times C_{\text{Min.}} + 10 \times C_{\text{Maj.}} + 25 \times C_{\text{Cri.}}}{\text{Total Word Count}}$$

where where we set four error severity levels: Neutral (Neu.), Minor (Min.), Major (Maj.), Critical (Cri.) with 0/5/10/25 severity penalty. $C_\star$ denotes the number of errors. The "Total Word Count" is calculated based on source input (Chinese word). Considering our task is centered on Zh-to-En translation, we engaged four evaluators who are native English speakers and also fluent in Chinese.

4 Task Description

Overview The shared task will be the translation of literary texts between Chinese→English. Participants will be provided with two types of training datasets: (1) discourse-level GuoFeng Webnovel Corpus; (2) General MT Track Parallel Training Data. Additionally, they are provided two types pretrained models: (1) in-domain pretrained models, including In-domain RoBERTa (base) and In-domain mBART (CC25). (2) other general-domain pretrained models listed in General MT Track. Note that basic linguistic tools are allowed

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

in the constrained condition as well as pretrained language models released before February 2023.

In the final testing stage, participants use their systems to translate an official testing set. The translation quality is measured by a manual evaluation and automatic evaluation metrics. All systems will be ranked by human judgement according to our professional guidelines and translators. Participants can submit either constrained (i.e. only use the training data specified above) or unconstrained (i.e. it allows the participation with a system trained without any limitations) systems with flags, and we will distinguish their submissions.

Goals The main goals of the task are to:

- Encourage research in machine translation for literary texts.
- Provide a platform for researchers to evaluate and compare the performance of different machine translation systems on a common dataset.
- Advance the state of the art in machine translation for literary texts.

Submission and Format Submissions will be done by sending us an email to our official email. Each team can submit at most 3 MT outputs per language pair direction, one primary and up to two contrastive. The requirements of submission format are (1) Keep 12 output files that are identical to the testing input files. (2) In the output files, ensure that each line is aligned with the corresponding input line.

5 Participants' and Baseline Systems

Here we briefly introduce each participant's systems and refer the reader to the participant's reports for further details. Table 1 shows the summary of systems and participant teams.

5.1 MaxLab (constrained)

The team from University of Southern California, Information Sciences Institute introduce three translation systems. The *Primary System* is built on a paragraph-level transformer, trained on a paragraph-aligned corpus (with a source side cap of 256 characters), executing translations at the paragraph level. The *Contrastive System 1* deploys a sentence-level transformer, capitalizing on the sentence alignment data available in the datasets. The *Contrastive System 2* adopts a paragraph-level Mega model (Ma et al., 2022). The Mega model proposed a single-head gated attention mechanism equipped with an exponential moving average, which achieves comparable performance compared to Transformers having with fewer parameters. In pre-processing, the team opted for Byte-Pair Encoding (BPE) for tokenization. And they employed Jaccard similarity for sentence alignment during the post-processing phase.

5.2 MAKE-NMT-VIZ (constrained)

The team from Université Grenoble Alpes introduced three translation systems. The *Primary System* finetune the mBART (CC50) model using Train, Valid_1, Test_1 of the GuoFeng Corpus, adopting settings similar to those described by Lee et al. (2022). Specifically, they finetune models for 3 epochs, utilizing the GELU activation function, a learning rate of 0.05, a dropout rate of 0.1, and a batch size of 16. For decoding, a beam search of size 5 was employed. The *Contrastive System 1* is implemented upon a finetuned concatenation transformer (Lupo et al., 2023) with two training steps: (1) a sentence-level transformer is trained for 10 epochs using General, Valid_1, Test_1 datasets; (2) a document-level transformer is finetuned using pseudo-document data (3-sentence concatenation) from Train, Valid_2, Test_2 data for 4 epochs. They use ReLU as an activation function, along with an inverse square root learning rate, a dropout rate of 0.1, and a batch size of 64. For decoding, a beam search of size 4 was employed. The *Contrastive System 2* is a sentence-level transformer model trained for 10 epochs using General, Valid_1, Test_1 datasets. The training adopted an inverse square root scheduled learning rate, a dropout rate of 0.1, and a batch size of 64. Decoding was done using a beam search of size 4.

5.3 TJUNLP (constrained)

The team from Tianjin University introduced a *Primary System* based on a sentence-level Transformer model. The training consists of two phases: initially, it undergoes 100k steps on a dense model, followed by a 50k step fine-tuning on mixture of experts (MOE). They adopt the Polynomial Decay as their learning rate scheduling strategy, with a learning rate set at 2e-4, a dropout rate of 0.1, and a batch size encompassing 4096 tokens. For decoding, a beam search of size 5 was employed. For pre-processing, the team opted for SentencePiece Model (SPM) for tokenization.

ID	Team	Institution	Flag	#System	Main Methods
1	MaxLab	University of Southern California	⊙	3	para-level Transformer
2	MAKE-NMT-VIZ	Université Grenoble Alpes	⊙	3	mBART
3	TJUNLP	Tianjin University	⊙	1	sent-level Transformer
4	DLUT	Dalian University of Technology	⊗	1	GPT-3.5-turbo
5	NTU	Nantong University	⊗	1	Opus-MT
6	HITer-WMT	Harbin Institute of Technology	⊗	2	Llama-7b
7	HW-TSC	Huawei Translation Services Center	⊗	3	doc2doc Transformer

Table 1: The summary of system submission and their participant teams. We also report the number of systems (#System) and the constrained (⊙) and unconstrained (⊗) flags.

5.4 NTU (unconstrained)

The Nantong University team introduce a *Primary System*. It is based on a pretrained MT model, Opus-MT,[4], which is trained on OPUS dataset (Tiedemann and Thottingal, 2020). The model is finetuned on one NVIDIA Tesla A100 80 GB where the learning rate is 5e-5, batch size is 64, max length is 512 and the epoch number is 10.

5.5 DLUT (unconstrained)

The team form Dalian University of Technology introduce a *Primary System* based on GPT-3.5-turbo (Brown et al., 2020). They mainly propose prompt engineering, data filtering, and document segmentation to activate the capabilities of LLMs for discourse-level translation (Zhao et al., 2023).

5.6 HITer-WMT (unconstrained)

The team form Harbin Institute of Technology (Harbin) introduce two translation systems. The *Primary System* centers on instruction fine-tuning, executed through the Llama-7b model within the Parrot framework (Jiao et al., 2023).[5] Specifically, they build an instruction dataset from two comprehensive chapters of our existing training corpus according to methodologies in Peng et al. (2023). This dataset was fine-tuned using Llama-7b over 3 epochs with a learning rate of 2e-5. The *Contrastive System* utilizes the GuoFeng mBART Model provided by the shared task. This model was trained over 10 epochs at a learning rate of 1e-4, with gradient clipping applied to stabilize training.

5.7 HW-TSC (unconstrained)

The team form Huawei Translation Services Center exploit a variety of techniques. They introduce an unconstrained Document-to-Document Translation system. They first train a sentence-level Transformer-big model with a 25-layer encoder and a 6-layer decoder, and perform domain adaptation with novel data on this model. They obtain a strong baseline using data augmentation methods including Back Translation, Forward Translation, and Data Diversification. They then perform incremental training using the Doc2Doc technique to turn the model into a document-level translation model. They also conduct document-level data augmentation using the Multi-resolutional Document-to-Document approach (Sun et al., 2022), and ensue the consistency of NE translations in a document with TrAining Data Augmentation (TADA). They submit three systems: the *Primary System* uses all strategies. In contrast to the primary system, the *Contrastive System 1* system does not use TADA, and the *Contrastive System 2* sets the beam size to 6 during inference, while 10 for other tasks.

5.8 Baseline Systems (unconstrained)

We select three representative systems as baselines. *Commercial Translation System*: we use Google Translate,[6], which usually performs state-of-the-art in translation performance. *Commercial LLM Systems*: we employ GPT-4 (8K) API[7] to translate documents, which is known for its extensive context modeling capabilities (Ouyang et al., 2022a; Wang et al., 2023c). *Open-sourced LLM Models*: we enhance Llama (2K) (Touvron et al., 2023) on document-level translation by using the 200K general-domain document-level training set (Du et al., 2023). All testing were conducted between August 1st and 30th, 2023. In the future, we will use more diverse model architectures such as non-autoregressive translation model (Gu et al., 2017;

[4] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en.

[5] https://github.com/wxjiao/ParroT.

[6] https://translate.google.com.

[7] https://platform.openai.com.

Type	System	Sent-Level				Doc-Level
		BLEU↑	chrF↑	COMET↑	TER↓	d-BLEU↑
Baselines	Llama-MT⋆	n/a	n/a	n/a	n/a	43.1
	GPT-4⋆	n/a	n/a	n/a	n/a	43.7
	Google⋆	37.4	57.0	80.50	57.4	47.3
Primary (con)	MaxLab	34.1	53.3	78.24	62.4	45.0
	MAKE-NMT-VIZ	37.9	56.6	81.50	58.7	**48.0**
	TJUNLP	32.1	51.9	77.93	64.1	43.3
Primary (uncon)	DLUT⋆	40.5	58.5	82.58	54.6	50.2
	NTU⋆	32.3	52.5	78.07	64.3	43.4
	HITer-WMT⋆	16.1	37.1	69.84	80.1	28.0
	HW-TSC⋆	44.3	61.1	82.69	51.8	**52.2**
Contrastive	MaxLab_1	34.5	54.7	79.14	62.7	44.9
	MaxLab_2	33.1	52.4	77.84	63.6	44.4
	MAKE-NMT-VIZ_1	33.8	51.2	76.91	63.5	45.5
	MAKE-NMT-VIZ_2	35.0	52.7	77.26	61.5	46.2
	$\text{HITer-WMT}^{\star}_1$	30.8	49.2	76.41	67.2	40.6
	$\text{HW-TSC}^{\star}_1$	44.6	61.0	82.67	51.8	52.6
	$\text{HW-TSC}^{\star}_2$	44.4	61.5	82.63	52.1	52.2

Table 2: 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.

Type	System	MQM	Rank
Baselines	GPT-4⋆	54.81	1
	Llama-MT⋆	28.40	2
	Google⋆	22.66	3
Primary (con)	MAKE-NMT-VIZ	**42.36**	**1**
	MaxLab	28.58	2
	TJUNLP	18.34	3
Primary (uncon)	DLUT⋆	**63.35**	**1**
	HW-TSC⋆	53.01	2
	NTU⋆	31.66	3
	HITer-WMT⋆	5.56	4

Table 3: Evaluation results of baseline and primary systems in terms of **human evaluation**. We report MQM score and System Rank.

Ding et al., 2020, 2021; Wang et al., 2023d).

6 Evaluation Results

6.1 Automatic Evaluation

We report the automatic evaluation scores of all submissions in Table 2. The evaluation metrics includes 1) sentence-level BLEU, chrF, COMET, TER; and 2) document-level d-BLEU. To calculate d-BLEU, we first concatenate all continuous sentences in one book as on line, and then employ sacreBLEU to obtain scorers. To compute d-BLEU, we merge all the consecutive sentences from a single book into one continuous line, and then utilize the sacreBLEU to generate the scores.

Among constrained Primary systems, the MAKE-NMT-VIZ system shows impressive performance and achieves the best in terms of all metrics. Similarly, the HW-TSC⋆ Primary system achieves the best in constrained settings. As introduced in Section 5, MAKE-NMT-VIZ mainly finetune the mBART pretrained model while HW-TSC⋆ train a doc2doc Transformer model using a number of data augmentation methods.

In the majority of teams, the primary system exhibits superior performance compared to the corresponding contrastive system. The exceptions to this trend are noted in the cases of HITer-WMT⋆ and HW-TSC⋆, where this pattern does not hold. Among the baseline systems, Google Translate, a commercial translation service, outperforms both

Type	Systems	Annotator				Average
		1	2	3	4	
Baselines	GPT-4⋆	95.84	73.38	76.71	87.52	83.36
	Llama-MT⋆	94.18	65.06	78.37	83.36	80.24
	Google⋆	85.02	42.60	59.23	21.13	52.00
Primary (con)	MAKE-NMT-VIZ	97.50	83.36	92.51	91.68	**91.26**
	MaxLab	86.69	61.73	71.71	74.21	73.59
	TJUNLP	88.02	55.07	20.97	69.22	58.32
Primary (uncon)	HW-TSC⋆	91.68	83.36	83.36	91.68	**87.52**
	DLUT⋆	95.01	69.22	84.19	90.02	84.61
	NTU⋆	85.02	39.27	28.45	62.56	53.83
	HITer-WMT⋆	57.57	21.80	0.00	31.78	27.79

Table 4: Analysis of human scores by different annotators on **one sampled document**. We report four annotators' scores and average score of Baselines, primary constrained and unconstrained (⋆) systems.

Annotator	1	2	3	4
1	-	-	-	-
2	0.858	-	-	-
3	0.824	0.878	-	-
4	0.752	0.875	0.676	-
Average	0.902	0.976	0.927	0.891

Table 5: Pearson correlation coefficient between scores by different annotators in Table 4.

commercial and open-source LLMs (GPT-4 API and Llama-MT) in terms of d-BLEU scores. Interestingly, both the top-1 ranked Primary constrained and the top-2 ranked unconstrained systems surpass the performance of the commercial MT system.

6.2 Human Evaluation

Table 3 presents the results of the human evaluation and system rank for the Primary submissions. We enlisted four human annotators to evaluate 5 documents, comprising a total of 2,194 words sourced from distinct books within the final testset for each translation system.

As seen, the MAKE-NMT-VIZ system outperforms the other three constrained systems, while DLUT⋆ ranks first among the four unconstrained systems. This is not fully consistent with the automatic evaluation results in Table 2. Moreover, the top-2 unconstrained systems outperform the best constrained system, highlighting the benefits of external knowledge. This observation is consistent with that of automatic evaluation.

Among the baseline systems, the LLM system performs the best, whereas the MT system shows the poorest performance, diverging from the observations of automatic evaluation. Interestingly, the literary MT-enhanced models perform comparable with some systems such as MaxLab and Google Translate.

6.3 Analysis

Inter-Annotator Agreement We engaged four annotators to independently review an identical document (i.e. 601 words) selected from the testset. Table 4 outlines the individual scores given by each annotator and the corresponding average scores. The findings illustrate that (1) while there is variance in the exact scores assigned by different annotators, their scoring trends align; (2) the results on this sample may diverge from those obtained from a larger dataset, highlighting the necessity of human evaluation on a larger scale.

In our effort to understand the consistency among the human evaluators, we conducted a Pearson correlation analysis on their scoring patterns. Table 5 illustrates the pairwise Pearson correlation coefficients for the scores given by each annotator. The results indicate a high degree of agreement among the annotators. For example, Annotator 2 demonstrated a very high correlation with Annotator 3 ($r = 0.878$) and Annotator 4 ($r = 0.875$). Besides, the Average Scores also reveal strong evaluator consensus on translation quality. This consistency underscores the reliability of the evaluators' judgments across the assessed translations.

Error Type We further analyze the error distribution in human-annotated results. Figure 4 classifies and counts the errors identified in the evaluated

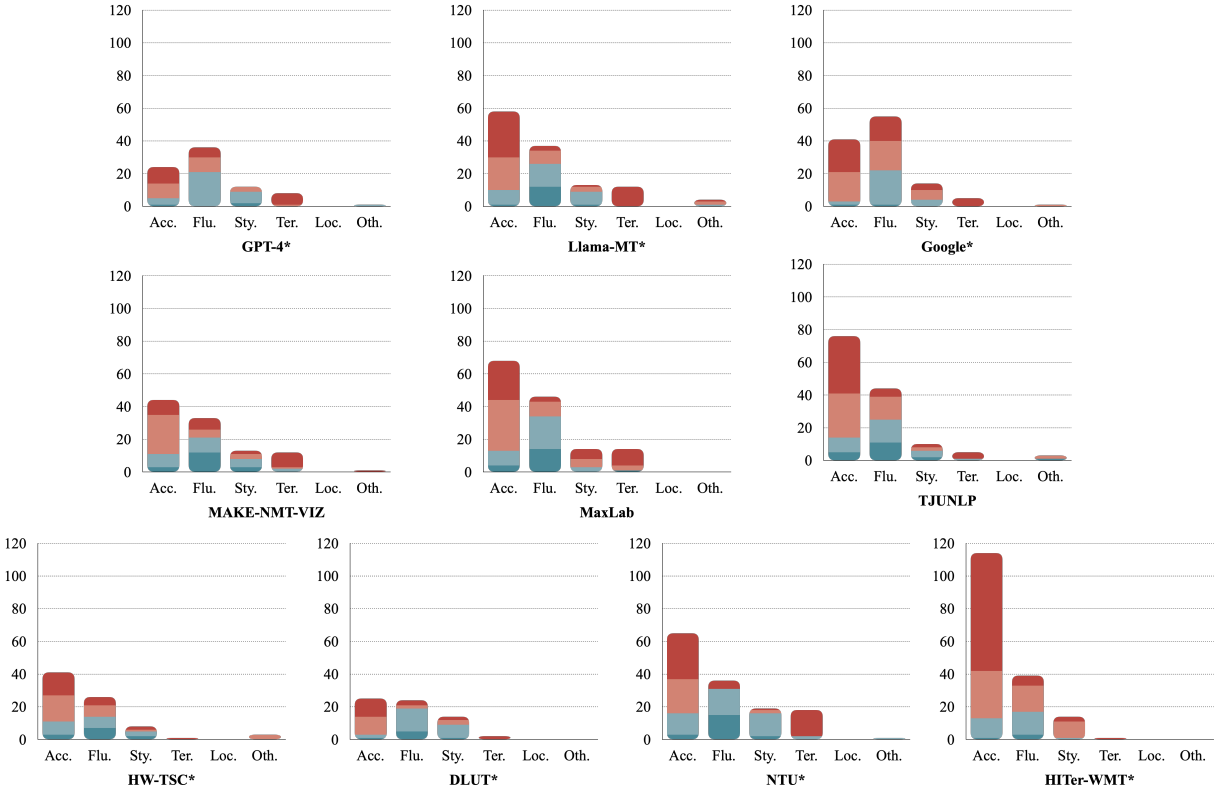

Figure 4: Analysis of error types in human annotations: Accuracy (Acc.), Fluency (Flu.), Style (Sty.), Terminology (Ter.), Localization (Loc.), and Other (Oth.). We report the count of error checkpoints in four evaluated documents. The four error severity levels are presented in different colors: Neutral (blue), Minor (light blue) , Major (light red), Critical (red). Systems marked with ⋆ are unconstrained, while others are constrained.

documents by their severity. This visualization allows for a direct comparison of the error profiles of each system, highlighting their strengths and weaknesses in different aspects of translation quality.

In the baseline systems analysis, GPT-4⋆ registers a higher frequency of Minor errors, particularly in Fluency and Style, indicating areas where refinement could enhance the translation's naturalness and adherence to stylistic norms. Llama-MT⋆, by contrast, has a pronounced incidence of Major and Critical errors in Accuracy and Terminology, raising concerns about the fidelity and technical precision of its translations. Google⋆ stands out with its Fluency errors, suggesting potential issues in maintaining a coherent and natural flow compared to the language models.

Regarding the constrained systems, MAKE-NMT-VIZ displays an even spread of errors, with relatively fewer instances in each category, which points to a well-rounded performance in capturing nuances across various aspects of translation. Both MaxLab and TJUNLP exhibit an increased number of Accuracy and Fluency errors, suggesting challenges in delivering translations that are not only faithful to the source material but also exhibit a seamless and natural flow in the target language.

The unconstrained systems, particularly HW-TSC⋆ and DLUT⋆, show a notable reduction in errors related to Accuracy and Fluency when compared to their constrained counterparts. This trend suggests that the lack of constraints may afford these systems more flexibility, resulting in translations that are more accurate and fluid. However, the overall error distribution across different systems highlights the complex trade-offs and challenges inherent in machine translation, underscoring the need for continued innovation and optimization in the field. In the future, we will also consider hallucination errors (Zhang et al., 2023).

7 Conclusion and Future Work

We believe that the WMT2023 Shared Task on discourse-level literary translation will be a valuable contribution to the field of machine translation and will encourage further research in this area. 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 (Xu et al., 2022, 2023; Wang et al., 2023a). 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.

Machine translation of web novels not only holds research value but also offers practical application prospects (Huang et al., 2021; Lyu et al., 2023). This shared task serves to spur competitive innovation and fosters the advancement of sophisticated machine translation systems capable of navigating the intricate nuances of literary works. Anticipating the future, our objective is to broaden the engagement in the forthcoming shared task, inviting an extensive range of collaborators from industry and academia alike to contribute their unique insights and expertise.

Acknowledgements

We would like to thank the WMT2023 organizers for providing us the opportunity to explore this new task. We also express our gratitude to the experts on the Shared Task Committee for their efforts in organization, evaluation, and advisory roles:

- Longyue Wang, Zhaopeng Tu, Dian Yu, Chenyang Lyu, Shuming Shi (Tencent AI Lab)
- Yan Gu, Yufeng Ma, Weiyu Chen (China Literature Ltd.)
- Bonnie Webber (University of Edinburgh)
- Siyou Liu, Yulin Yuan (University of Macau)
- Philipp Koehn (Johns Hopkins University)
- Liting Zhou, Andy Way (Dublin City University)
- Yvette Graham (Trinity College Dublin)
- Chao-Hong Liu (Potamu Research Ltd.)
- Qingsong Ma (Tencent AI Evaluation Lab)

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. *arXiv preprint arXiv:2012.14583*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Progressive multi-granularity training for non-autoregressive translation. *arXiv preprint arXiv:2106.05546*.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2023. On extrapolation of long-text translation with large language models. *arXiv preprint*.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *NAACL*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of EMNLP*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *LREC*.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation*, pages 10–19.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas (AMTA)*.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of ACL*.

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. *arXiv preprint arXiv:2210.14250*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*.

Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Ph. D. thesis, Dublin City University, Dublin, Ireland.

Longyue Wang, Zefeng Du, Donghuai Liu, Cai Deng, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023a. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *arXiv preprint arXiv:2307.08074*.

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023b. A survey on zero pronoun translation. *arXiv preprint arXiv:2305.10196*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023c. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *EMNLP-IJCNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *NAACL*.

Zhihao Wang, Longyue Wang, Jinsong Su, Junfeng Yao, and Zhaopeng Tu. 2023d. Revisiting non-autoregressive translation at scale. *arXiv preprint arXiv:2305.16155*.

Mingzhou Xu, Longyue Wang, Siyou Liu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2023. A benchmark dataset and evaluation methodology for chinese zero pronoun translation. *Language Resources and Evaluation*.

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. GuoFeng: A benchmark for zero pronoun recovery and translation. In *EMNLP*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. 2023. Dutnlp system for wmt23 discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation*.

Type	Granular	Definition	Examples
Accuracy	Addition	The target text includes text not present in the raw.	A translation includes portions of another translation that were inadvertently pasted into the document or the translator has added too many details of his own.
	Omission	Content is missing from the translation that is present in the source.	A paragraph present in the source is missing in the translation.
	Mistranslation	The target content does not accurately match the raw.	A source text states that a medicine *should not be* administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).
	Misnomer	The target text is more/less specific than the raw.	1. The source text refers to a boy but is translated with a word that applies only to young boys rather than the more general term. 2. The source text uses words that refer to a specific type of military officer but the target text refers to military officers in general.
	Untranslated	Content that should have been translated has been left untranslated.	A sentence to be translated into English was left in Chinese.
Fluency	Punctuation	Punctuation marks missing or used in a wrong way.	An English text uses a semicolon where a comma should be used.
	Spelling	Issues related to spelling of words. (Including those of capitalization, hyphenated words, and use of asterisk for censored swear words.)	The English word "Translation" is spelled "Transaltion".
	Grammar	Issues related to the grammar or syntax of the text, other than spelling and orthography. (especially inconsistency of the tenses and conditionals)	An English text reads "The man was seeing the his wife."
	Inconsistency	The text shows internal inconsistency.	A text uses both "app." and "approx." for "approximately".
Style	Awkwardness	A text is written with an awkward style.	A text is written with many embedded clauses and an excessively wordy style. While the meaning can be understood, the text is very awkward and difficult to follow.
	Inconsistent	Style is inconsistent within a text.	One part of a text is written in a light and terse style while other sections are written in a more wordy style.
	Unidiomatic	The content is grammatical, but not idiomatic.	The following text appears in an English translation of "我们衷心感谢他": "We thanked him with heart" where "with heart" is an understandable, but non-idiomatic rendering, better stated as "heartily".
Terminology	Mistranslation	A genre-specific or cultural-specific terminology is wrongly translated.	A Chinese word "修士" is translated into "practitioner" rather than the expected "cultivator".
	Inconsistent	Terminology is used in an inconsistent manner within the text.	"斗罗大陆" is translated into "Douluo Land" in the first few chapters and then into "Soul Land".
Locale Convention	Location Format	Using the wrong format for address, name etc.	A Chinese address "北京市朝阳区花园路22号" is translated into "Beijing, Chaoyang district, Huayuan Road N.22" instead of the expected "N.22, Huayuan Road, Chaoyang District, Beijing".
	Number Format	The translated date, time, currency, telephone use formats inappropriate for its locale.	An English text has 2012-06-07 instead of the expected 06/07/2012.
Others		Other issues that haven't been included in this list.	E.g. signs of MT, mimetic word, gender bias, source errors etc.

Table 6: The MQM-based evaluation criteria for literary translation.