Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels

Yuchen Eleanor Jiang[ζ] **Tianyu Liu**[ζ] **Shuming Ma**[γ]
Dongdong Zhang[γ] **Mrinmaya Sachan**[ζ] **Ryan Cotterell**[ζ]
[ζ]ETH Zürich [γ]Microsoft Research Asia
{yuchen.jiang,tianyu.liu,ryan.cotterell,mrinmaya.sachan}@inf.ethz.ch
{shuming.ma,dongdong.zhang}@microsoft.com

Abstract

Several recent papers claim to have achieved human parity at sentence-level machine translation (MT)—especially between high-resource language pairs. In response, the MT community has, in part, shifted its focus to document-level translation. Translating documents requires a deeper understanding of the structure and meaning of text, which is often captured by various kinds of discourse phenomena such as consistency, coherence, and cohesion. However, this renders conventional sentence-level MT evaluation benchmarks inadequate for evaluating the performance of context-aware MT systems. This paper presents a new dataset with rich discourse annotations, built upon the large-scale parallel corpus BWB introduced in Jiang et al. (2022a). The new BWB annotation introduces four extra evaluation aspects, i.e., entity, terminology, coreference, and quotation, covering 15,095 entity mentions in both languages. Using these annotations, we systematically investigate the similarities and differences between the discourse structures of source and target languages, and the challenges they pose to MT. We discover that MT outputs differ fundamentally from human translations in terms of their latent discourse structures. This gives us a new perspective on the challenges and opportunities in document-level MT. We make our resource publicly available to spur future research in document-level MT and its generalization to other language translation tasks.

https://github.com/EleanorJiang/BlonDe/tree/main/BWB

1 Introduction

The field of machine translation (MT) has made tremendous strides in recent years thanks to neural machine translation (NMT) models that can utilize massive quantities of parallel training data (Vaswani et al., 2017; Junczys-Dowmunt et al., 2018, *inter alia*). Unfortunately, most parallel corpora are aligned only at the sentence level, and document-level translation has remained limited to small-scale studies (Ansari et al., 2020; Lison et al., 2018; Koehn, 2005; Tiedemann, 2012).

There are, however, inherent differences between sentence-level translation and document-level translation, as documents consist of complex discourse structures that go beyond the mere concatenation of individual sentences. Three key discourse features are particularly important in document-level translation. First, the translations of *named entities* have to be consistent. For example, in Fig. 1, the same terminologies are not consistently referred to with the same translations (i.e., **Weibo** vs. **micro-blog**, **Qiao Lian** vs. **Joe** vs. **Joe love**), and as a result, the sentence-level MT system fails to capture discourse dependencies across sentences. Second, the *coreference relationship* in the source language needs to be preserved. In particular, the relations between entities and their pronominal anaphora should be preserved, as well as the transition of discourse centers[1]. In Fig. 1 and Fig. 8, coreference chains are color-coded, visually demonstrating the preservation of anaphoric referential relations and the transition chains of centers. Third, the *conversational structure*, such as transitions between speakers, must be maintained.

Inferring latent discourse structures from documents is essential for translation coherence because of these discourse features. As a result, conventional sentence-level MT pipelines that do not consider context are unable to generate natural and coherent translations (Lapshinova-Koltunski et al., 2018; Werlen and Sadiht, 2021). While efforts have been made to develop context-aware NMT models over the past few years (Tiedemann and

[1] In this context, discourse centers refer to the entities that are in the readers' focus at a certain point in the discourse. It can be realized by either pronouns or nominal mentions (Grosz et al., 1995).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7853–7872
July 9-14, 2023 ©2023 Association for Computational Linguistics

	SOURCE	REFERENCE	MT
(1)	乔恋攥紧了拳头，垂下了头。	**Qiao Lian** clenched **her** fists and lowered **her** head.	**Joe** clenched **his** fist and bowed **his** head.
(2)	其实他说得对。	Actually, **he** ***was*** right.	In fact, **he*’s*** right.
(3)	⟦⟧自己就是一个蠢货，竟然会[...]。	⟦**She**⟧ was indeed an idiot, as only an idiot would [...]	⟦**I**⟧ ***am*** a fool, even ***will*** [...]
(5)	她点进去，发现是凉粉群，所有人都在@她[...]	**She** logged into ⟦**her**⟧ account and saw that a large number of fans in the **Liang fan group** had tagged **her**.[...]	**She** nodded in and found it was a **cold powder group**, and everyone was on **her**.[...]
(7)	【川流不息：乔恋，快看微博头条！微博头条！】	[**Chuan Forever**: **Qiao Lian**, look at the headlines on **Weibo**, quickly!]	**Chuan-flowing**: **Joe love**, quickly look at the **micro-blogging** headlines! **Weibo** headlines?
(8)	她微微一愣，拿起手机，登陆微博，在看到头条的时候，整个人一下子愣住了！	**She** froze momentarily, then picked up ⟦**her**⟧ cell phone and logged into **Weibo**. When ⟦**she**⟧ saw the headlines, ⟦**her entire body**⟧ immediately froze over again!	**She** took a slight look, picked up the phone, landed on the **micro-blog**, when ⟦**she**⟧ saw the headlines, ⟦**the whole person**⟧ suddenly choked!

Figure 1: Selected examples of the annotations in BWB-test. The mentions in the same coreference chain are marked with the same color. Pronoun omissions are marked with ⟦⟧. The mistranslated verbs are marked with ***teal***, and the mistranslated named entities are **underlined**. (7) is a quotation with the speaker annotated as **Chuan Forever**. The full chapter is given in Fig. 8. MT is the output of a Transformer-based sentence-level machine translation system.

Scherrer, 2017; Agrawal et al., 2018; Voita et al., 2018; Bawden et al., 2018; Zhang et al., 2018), a reliable evaluation method for document-level MT has not yet been developed.

Traditional MT evaluation metrics, such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005), are limited to evaluating translation quality at the sentence level and do not consider contextual information. Even with the use of more complex neural evaluation methods, e.g., COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020) and BERTSCORE (Zhang et al., 2020), it is not possible to accurately assess the coherence of translations within the context of a document. BLONDE (Jiang et al., 2022a) has attempted to evaluate document-level translation quality and obtained a higher correlation with human evaluation. Yet, it is a coarse-grained automatic evaluation metric based on surface syntactic matching, thus being unable to evaluate deeper discourse features, such as coreferences and quotations. Vernikos et al. (2022) propose to concatenate context to extend pretrained metrics to the document level. However, its lack of interpretability leaves it unclear as to why it assigns a particular score to a given translation, which is a major reason why some MT researchers are reluctant to employ learned metrics in order to evaluate their MT systems (Karpinska et al., 2022).

To address this issue, we have annotated a new benchmark based on the large-scale bilingual parallel corpus BWB proposed in previous work (Jiang et al., 2022a). This benchmark includes four evaluation aspects annotated on both source and target sides: entity, terminology, coreference, and quotation. It consists of 80 documents across multiple fiction genres, with 15,095 mentions covered in 150,287 words in total.

Using these annotations, we systematically evaluate the latent discourse structures of MT models. We show that, while context-aware MT systems perform better than phrase- and sentence-level ones in terms of entity translation, they still lag behind human translation in terms of proper noun and personal name translation. In addition, we demonstrate that humans are far more adept at preserving coreference chains than MT models.

The main contributions of our paper are:

- We propose a new benchmark that includes four types of discourse annotations that are closely related to document-level translation.
- We systematically investigate the similarities and differences between the discourse structures of source and target languages, and the challenges they pose to machine translation.
- We demonstrate that machine translations differ fundamentally from human translations in terms of their latent discourse structures. We believe these new insights can lead to potential improvements in future MT systems.

2 Corpus

We draw our source material for annotation from the texts in the test set of BWB (Jiang et al., 2022a), which consists of 65,107 words, 2,633 sentences in 80 different documents drawn from 6 web novels across different genres. The distribution of genres in BWB are shown in Fig. 2. The statistics of BWB are shown in Tab. 1. We refer the readers to App. D

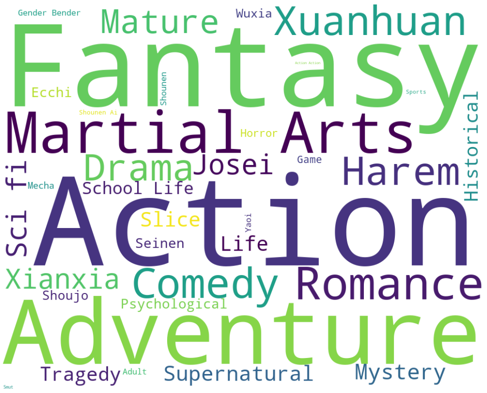

Figure 2: Genre distribution of novels in the BWB corpus. Action, fantasy and adventure are the most common genres.

for the dataset creation details.

	Size				Averaged Length	
Split	#word	#sent	#chap	#book	#word/sent	#sent/chap
Train	325.4M	9.57M	196K	378	34.0	48.8
Valid	68.0K	2.63K	80	6	25.8	32.9
Test	67.4K	2.63K	80	6	25.7	33.1

Table 1: Statistics of the BWB corpus.

3 Annotation

In this section, we describe the annotation scheme and the process of annotating the corpus, and provide examples of how the annotations can be used to study discourse phenomena in the two languages. The annotation was conducted by eight professional translators. A total of 8,585 mentions on the English side and 6,070 mentions on the Chinese side have been annotated.

3.1 Named Entities

The mistranslation of named entities (NEs) can significantly harm the quality of translation, although often under-reported by automatic evaluation metrics (e.g., BLEU, COMET). Therefore, we annotate named entities according to the ACE 2005 guidelines (Walker et al., 2005). These annotations identify six categories of entities: people (PER), facilities (FAC), geo-political entities (GPE), locations (LOC), vehicles (VEH), and organizations (ORG). In addition, each unique entity is assigned an entity_id for the purpose of coreference resolution. For example, the following text includes annotations for two entities:

(1) · · · [$_{\text{PER,6}}$ the scion of [$_{\text{ORG,5}}$ the Ri family]]

SOURCE	REFERENCE	MT
[$_{\text{PER,T,1}}$ 木子]	[$_{\text{PER,T,1}}$ Mu Zi]	Wood
[$_{\text{VEH,N,2}}$ 马车]	[$_{\text{VEH,N,2}}$ the carriage]	the carriage
[$_{\text{FAC,N,3}}$ 屋子]	[$_{\text{FAC,N,3}}$ the house]	the house
[$_{\text{GPE,N,4}}$ 欧洲]	[$_{\text{GPE,N,4}}$ Europe]	Europe
[$_{\text{LOC,N,7}}$ 林]	[$_{\text{LOC,N,7}}$ the forest]	the forest
(omitted)	[$_{\text{ORG,T,12}}$ the Ri family]	not translated
[$_{\text{ORG,T,14}}$ 三泉宗]	[$_{\text{ORG,T,14}}$ the Three Spring Sect]	Sanquanzong

Table 2: Examples of entity annotations and terminology annotations. (1) Terminology is more prone to mistranslation and situations where translation is inconsistent with context. (2) MT often produces translations of poor quality when the Chinese subject is omitted. This is likely due to the difficulty in understanding the semantics of the sentence with ellipsis, as the subject is an integral part of the meaning of the sentence.

Nested entities are also annotated. The annotated dataset contains a total of 5,984 English entities and 4,853 Chinese entities. There are fewer entities annotated in Chinese compared to English because the subject is often dropped in Chinese. This phenomenon also serves as a significant contributor to translation errors in context-agnostic MT systems when translating from Chinese to English, as shown in Fig. 1 sentence (3).

3.2 Terminology

Terminology refers to specialized words or phrases conventionally associated with a particular subject matter or whose use is agreed upon by a community of speakers. In the context of a novel, terminology refers to the specific words and phrases that are used to describe the concepts, characters, settings, and events in the story. These words and phrases may be specific to the genre or style of the novel, or they may be unique to the world or setting created by the author. For example, in a fantasy novel that uses made-up magic terms, incorrect terminology translation could lead to misunderstandings of the powers and abilities of characters, the rules of the magical world, or the significance of certain events. Similarly, incorrect translations of named entities could lead to confusion or misunderstandings about the identities and roles of characters in the story. Inconsistent terminology translation can compromise the integrity and cohesion of a work and make it difficult for readers to fully grasp the intended meaning. Therefore, we have included a layer for terminology identification. This layer is a binary classification of whether a certain span counts as terminology (T) or not (N). Tab. 2 shows some term and non-term annotations in the dataset. There are 2,156 English terms and 2,290 Chinese terms that have been annotated, accounting for approximately 52% of all entities.

3.3 Coreference

Coreference is the phenomenon in which a word or phrase in a text refers to another word or phrase that has been previously mentioned in the text. Establishing coreference is important for determining gender and number marking on pronouns, determiners, and adjectives as well as for making lexical decisions. Errors can compromise coherence and accuracy. For example, consider the following sentence in English:

SRC In fairness, Miller did not attack the statue itself.
... But he did attack its meaning ...
REF Um fair zu bleiben, Miller griff nicht die Statue selbst an.
... Aber er griff deren Bedeutung an ...
MT Fairerweise hat Miller die Statue nicht selbst angegriffen.
... Aber er griff seine Bedeutung an ...

In this example, both *itself* and *its* refer back to *the statue*. However, in MT, *its* is incorrectly translated as seine, which is the masculine possessive pronoun in German, as a result of the incorrect coreference resolution.

We follow the OntoNotes Coreference Annotation Guidelines for Chinese and English (Weischedel et al., 2012), and consider proper noun phrases, common noun phrases, and personal pronouns as coreference candidates. In particular, we follow the three important principles of the OntoNotes guidelines.

Maximal Spans. We include all modifiers in an annotated span. I.e., we annotate [the surrounding passersby, who were actually reporters in disguise] rather than simply [passersby].

Rather Lack Than Abuse. When the annotators were in doubt, they were told to not mark a difficult coreference decision.

Ellipsis. Omitted pronouns are marked with O and other pronouns are marked with P. For instance, consider:

SRC [$_{\text{PER,T,1}}$ 乔恋] 攥紧了拳头,[$_{\text{P,1}}$ 她] 垂下了头。
REF [$_{\text{PER,T,1}}$Qiao Lian] clenched [$_{\text{O,1}}$ her] fists and lowered [$_{\text{P,1}}$ her] head.

In this example, the first *her* is omitted in Chinese and is therefore marked as <O,1>, where 1 is the entity id of *Qiao Lian*.

3.4 Quotations

The final annotation layer is quotation. In this stage of the process, we identify instances of direct speech and attribute the speech to its speaker. The inclusion of direct speech is common in literature, and its proper translation is essential. For instance, the same person can be addressed by different names by different people. Furthermore, MT systems that lack contextual awareness may have difficulty correctly identifying the speaker in instances of direct speech, leading to inconsistencies in overall contextual translation. For example,

(2) [$_{\text{Q,2}}$ “Oh dear! Oh dear! I shall be late!”]

where 2 is the entity id of the speaker and *“Oh dear!”* is an exclamation. In this example, discriminating exclamations from vocatives is vital for the cohesiveness of the story. In addition, there are cases where knowing the speaker is important for coreference, e.g., John (Mary, respectively) said to Mary (John, respectively), “Oh, this is your dog. Her (His, respectively) dog barked.” In BWB test set, there are 840 (31.9%) sentences that contain quotations, and there are 25 distinct speakers in total.

	Mention	Entity	Terminology	Coreference	Quotation
English	80.7	79.3	89.2	74.1	90.1
Chinese	86.5	74.1	84.8	65.9	89.7

Table 3: Inter-annotator agreement. For coreference, the average CoNLL F1 scores are reported.

Inter-Annotator Agreement. To ensure annotation quality, we randomly select 10 documents and have them independently annotated by another expert, following previous work (Bamman et al., 2019). We then calculate the mention overlap and F1 scores for entity, terminology, coreference, and quotation between the two annotations, using the regular annotator's result as the hypothesis and the second annotator's as the reference. We achieve comparable inter-annotator agreements with previous work (Cohen et al., 2017; Bamman et al., 2019). The results are reported in Tab. 3.

4 Bilingual Analysis

We conduct a thorough bilingual analysis of the novel evaluation aspects using the new annotation. The following two research questions are being investigated:

- How different (or similar) are the discourse structures in the source language (Chinese) and the target language (English)?
- How do the differences in discourse structures affect the translation quality of MT systems?

	EN			ZH		
Type	Count	Freq.	Repet.	Count	Freq.	Repet.
PER	3,387	80.5%	16	3,552	81.6%	26
FAC	360	8.6%	2	325	7.5%	3
ORG	232	5.5%	5	241	5.5%	6
LOC	108	2.6%	3	115	2.6%	4
VEH	79	1.9%	3	77	1.8%	3
GPE	44	1.0%	1	41	0.9%	1
NOM	2,054	48.8%	8	2,061	47.4%	11
TERM	2,156	51.2%	18	2,290	52.6%	30

Table 4: The distributions of different types of entities in both English and Chinese in BWB-test. *Freq.* and *Repet.* stand for the frequency and the average number of repetitions, respectively.

Lang	MASCULINE	FEMININE	NEUTER	EPICENE	**Omitted**
EN	1,633	2,521	608	391	**64.9%**
ZH	654	967	14	118	9.4 %

Table 5: The distributions of different types of pronouns in both English and Chinese in the BWB test set.

Entity types and terminology. Tab. 4 demonstrates the distribution of entity annotations and term annotations in both Chinese and English in the BWB test set. First, the BWB test set is dominated by person and facility entities, with a much lower proportion of geo-political entities. In addition, terminology entities and person entities are repeated more frequently, requiring better translation consistency. An important sanity check from our analysis is that the number and distribution of entities in Chinese and English are similar. This suggests that at the discourse level, the information being conveyed in these two languages is largely the same.

Pronouns. Pronoun translation has been the focus of discourse-level MT evaluation (Hardmeier, 2012; Miculicich Werlen and Popescu-Belis, 2017). We compare the numbers of different types of pronouns in Chinese and English in Tab. 5. As can be seen, Chinese has significantly fewer pronouns due to its pronoun-dropping property. This poses additional challenges for NMT systems, as they must be able to resolve anaphoric references. In addition, Tab. 5 reveals a notable difference in the frequency of neuter pronouns. English exhibits a higher prevalence of neuter pronouns, possibly due to the presence of a larger number of expletive subjects in the language.

Coreference. We further conduct analyses to investigate the differences and similarities between coreference behaviors in Chinese and English. Fig. 3a examines the distribution in distances to the *nearest* antecedent for both Chinese and English. The average distance between antecedents and anaphora in English is shorter than in Chinese. This may be connected to the more common use of pronoun ellipsis in Chinese—when referring to closer antecedents, pronouns are often omitted. Fig. 3b illustrates the distance between the first and last mention of an entity within each coreference chain. It can be observed that, although English coreference chains tend to be longer in general compared to Chinese coreference chains, the distribution of these lengths is consistent. This represents another language-independent discourse feature in addition to entity distribution. The spread distributions of the two languages are further depicted in Fig. 8 and Fig. 9. Finally, we present an analysis of the size of coreference chains in Fig. 3c. Our results indicate that the number of mentions in English coreference chains tends to be larger than in Chinese.

Error Type	#	Description	An
ENTITY	43.3%	error(s) due to the mistranslation of named entities.	✓
TERM	19.2%	error(s) caused by the mistranslation of terminologies.	✓
COREF	34.0%	error(s) caused by coreference resolution failure(s).	✓
ZEROPRO	17.3%	error(s) caused by the omission of pronoun(s).	✓
QUOTE	1.1%	error(s) caused by the misinterpretation of quotation(s).	✓

Table 6: The types of NMT errors and their description. # represents the percentage of the error in the BWB test set. ✓ indicates "with annotation".

Challenges for MT. So far our analyses have revealed that the Chinese source language and English target language exhibit several distinct linguistic characteristics, leading to various discourse phenomena, including the omission of pronouns in Chinese and the use of expletive subjects in English, as well as shorter English coreference chains, among others. These discourse phenomena present challenges for MT. In order to systematically analyze translation errors caused by these phenomena, we conducted a study in which four professional translation experts compared reference translations to those generated by a commercial MT system

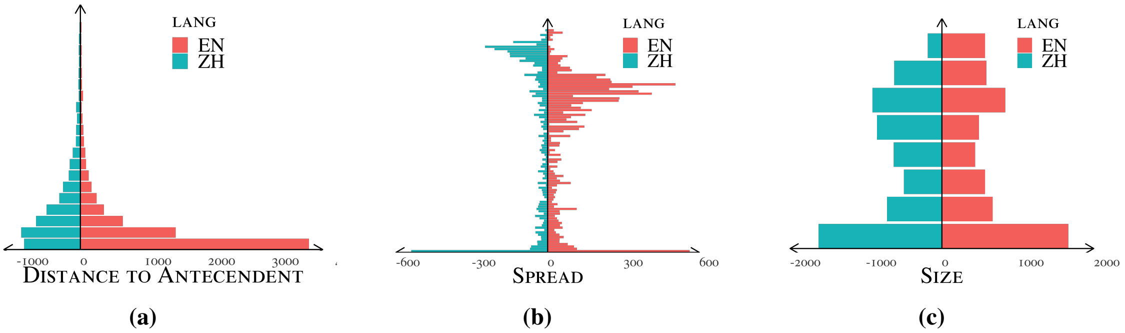

Figure 3: The histograms of coreference chains. The blue and red bars are the count bins of the distance to the nearest antecedent for each mention, the spreads between the first and last mention of entity, and the number of mentions in each coreference chain, respectively.

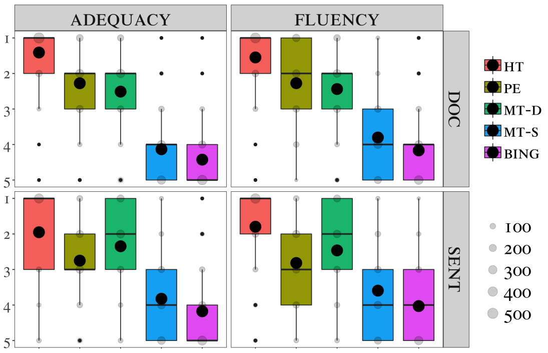

Figure 4: Human evaluation results on BWB. Each ● represents the average of the system's rankings.

Google Translate. The annotation guidelines are as follows:

1. Identify translation errors and identify whether the translation error is at the DOCUMENT level, e.g., the translation is inconsistent with the context or does not comply with the global criterion of coherence.
2. Categorize the DOCUMENT errors in accordance with the discourse phenomena and mark the corresponding spans in the reference (English) that cause the MT output to be incorrect.

The results, summarized in Tab. 6, indicate that named entity translation is the most significant issue, with terminology translation, coreference resolution failures, and pronoun omission also being notable problems. This analysis emphasizes the importance of accurately inferring latent discourse structures in the translation of long texts.

5 Exploring Discourse Features of MT Systems and Human Translation

The existence of a bilingual corpus with discourse annotations allows us to test the performance of both human translations and MT systems—including the popularly used commercial systems and those trained on the in-domain dataset BWB. The following 6 MT systems are adapted:[2]

- **SMT**: phrase-based baseline (Chiang, 2007).
- **BING**, **GooGLe**, **BaiDu**: commercial systems.
- **SENT**: the sentence-level transformer-based (Vaswani et al., 2017) baseline model.
- **DOC**: the document-level NMT model that adopts two-stage training (Zhang et al., 2018).
- **PE**: the post-edited BING outputs produced by professional translators. They were instructed to correct only discourse-level errors with minimal modification.

Overall Quality. The quality of the translations produced by each system was first evaluated using a range of automatic translation metrics, including BLEU, METEOR, BERTSCORE, COMET, and BLONDE. The results of this evaluation are presented in Tab. 7. In addition, we conducted a human evaluation of the translations, the results of which are shown in Fig. 4.[3] The large gap between the performances of HT and MT indicates that the genre of BWB, i.e., literary translation, is challenging for MT, and NMT systems are far beneath human parity. DOC performs significantly better than SENT, suggesting that BWB contains rich discourse phenomena that can only be translated accurately when the context is taken into account. It is also worth noting that even though PE is the post-edit of the relatively poor-performing system BING, it still achieves surprisingly better performance than DOC at the document level. This observation confirms

[2] SENT and DOC are trained on BWB by fairseq (Ott et al., 2019), and the training details are in App.C.

[3] App. E describes how human assessment is carried out. The inter-rater agreement is reported in Tab. 11.

	Automatic Metrics						Discourse Phenomena									
	BLEU	METEOR	TER	BERT	COMET	BLONDE	AMBIGUITY		ENTITY		TENSE		PRONOUN		ELLIPSIS	
SMT	6.86	18.73	84.35	32.14	-0.65	15.52	28.92	20.94	30.05	19.67	51.01	44.58	52.21	38.73	34.85	21.66
BD	10.02	26.57	70.72	46.78	-0.16	16.17	41.52	31.47	21.28	12.04	58.36	54.19	56.81	43.65	52.53	41.05
BING	12.23	23.27	89.53	41.45	-0.38	22.47	8.09	3.33	31.37	22.62	58.66	54.93	61.54	52.29	34.88	21.74
GGL	12.81	20.80	81.39	46.80	-0.22	22.86	46.66	36.05	27.46	20.03	61.01	58.58	59.04	48.63	43.04	30.76
SENT	15.24	22.39	78.93	41.58	-0.08	26.79	58.16	37.59	22.27	20.54	69.77	65.77	65.21	62.64	64.01	55.81
DOC	17.45	25.21	89.48	44.78	0.03	31.53	61.95	43.18	28.25	25.67	70.39	67.80	74.20	71.96	73.55	71.93
PE	19.52	22.38	78.47	55.50	0.10	38.18	60.65	59.09	60.94	60.41	71.68	67.56	71.51	61.04	77.45	78.75

Table 7: Results of MT systems and human post-editing on the BWB test set. For discourse phenomena, we report both F1 measure defined in Jiang et al. (2022a) and exact-match accuracy defined in Alam et al. (2021).

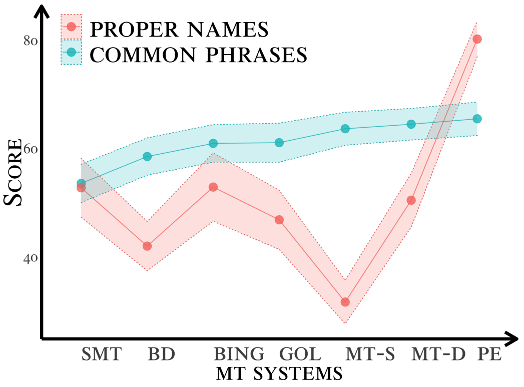

Figure 5: Terminology evaluation results on BWB.

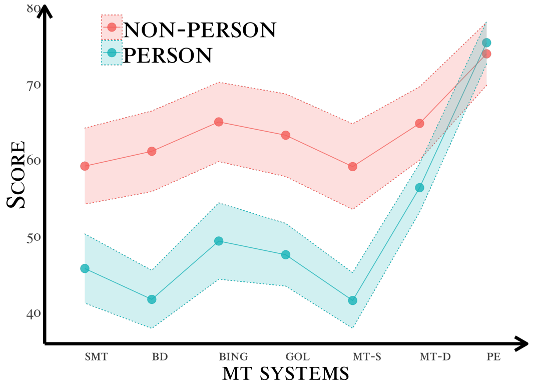

Figure 6: The F1 scores of person entities vs non-person entities on BWB.

our claim that discourse phenomena have a huge impact on human judgment of translation quality.

Terminology Translation. We next turn to evaluating the performance of each system on terminology translation, depicted in Fig. 5. The recall rate of terminology is reported. As can be seen, although the accuracy of each system in the translation of common phrases is not significantly different, the performance of the PE system is superior to that of the MT systems in the translation of proper names. Notably, despite having been trained on in-domain data, the SENT system is unable to accurately translate terms on the test set due to the varied terminologies used in different novels, resulting in a performance that is even lower than that of commercial MT systems.

Entity Translation. For the evaluation of entity translations, we compared translations of PER and translations of other entities. As illustrated in Fig. 6, the translation of personal names and terms exhibits a similar trend, with the MT systems performing significantly worse than PE. This outcome is expected, as both personal names and terms present similar challenges for MT, including (1) the need for consistent translation, (2) the potential for multi-word personal names to be combined in specific ways that may not translate well when each word is translated individually, and (3) the possibility of personal names not having a direct equivalent in other languages. For example, in Fig. 9, the character name **Ye Qing Luo** *(literal: night, clear, fall)* is mistranslated into **The night fell**. Additionally, wordplay and puns based on character names may be difficult to convey in translation, as character names may be chosen for their sound or meaning in the original language, which may not translate effectively into other languages.

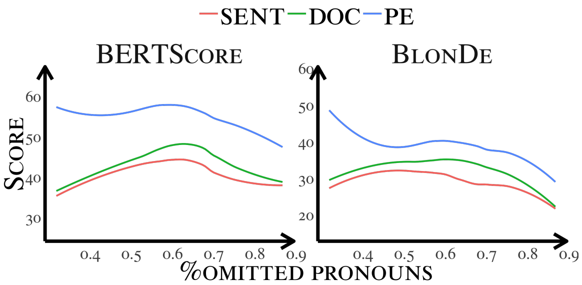

Figure 7: BLONDE and BERTSCORE as a function of omitted pronouns. The x-axis is the ratio of omitted pronouns to the total number of pronouns in each document.

Pronoun Translation. Our evaluation of pronoun translation focuses on the impact of omitted pronouns on translation quality. As depicted

	B^3↑	CEAF↑	COHERENCE↓
SENT	43.3	49.7	4
DOC	63.5	55.1	3*
PE	74.0	67.3	2
HT	88.1	76.7	1*

Table 8: The coreference resolution performance (B^3, CEAF) and the coherence ranking of different translation systems. The Kruskal–Wallis test was applied and the systems that are significantly better than the previous one with p-value $< .05$ are marked with *.

in Fig. 7, an increase in the proportion of omitted pronouns leads to a decrease in the translation quality of the human post-editing system. This finding suggests that pronoun omission is a major issue in Chinese-to-English translation. However, the BERTSCORE scores of the sentence-level and document-level MT systems do not follow a monotonic trend with the increase in the proportion of omitted pronouns. Rather, the highest value is reached when the proportion of omitted pronouns is within the interval of $[0.6, 0.7]$. This result indicates that NMT models process pronoun omissions in a manner that is fundamentally different from that of humans. In contrast, the BLONDE scores exhibit a pattern that is consistent with the trend observed in the PE system, indicating that while the omission of pronouns may not significantly affect the quality of translation at the sentence level, it is crucial for ensuring the overall quality of document-level translation.

Translation Coherence. Finally, the coreference annotations allow us to measure the coherence of translated texts. We first fine-tune a neural coreference model (Lee et al., 2017) trained on OntoNotes (Hovy et al., 2006). This end-to-end model jointly performs mention detection and antecedent linking. We modify the model by replacing the span representations with SpanBERT embeddings (Joshi et al., 2020). Additionally, during inference, we inject mentions that match the reference annotations into the antecedent candidate sets after the mention detection stage. The model then greedily selects the most likely antecedent for each mention in a document, or a dummy antecedent that begins a new coreference chain with the mention. Coreference resolution performance and the coherence ranking are reported in Tab. 8. Coherence is compared at the document level according to the centering theory (Grosz et al., 1995). We use the operationalization of Jiang et al. (2022b) and compare KP scores of each document.

As shown in Tab. 8, HT surpasses MT in terms of entity coherence. Additionally, human post-editing was found to achieve entity coherence comparable to that of HT. These results highlight the challenges that MT currently faces in regard to maintaining contextual coherence when translating longer texts.

6 Comparing to Related Work

Evaluation Test Suites for Context-Aware MT. Several context-aware test suites have been proposed in recent years (Hardmeier et al., 2015; Guillou and Hardmeier, 2016; Burchardt et al., 2017; Isabelle et al., 2017; Rios Gonzales et al., 2017; Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019; Guillou and Hardmeier, 2018). Although they facilitate the development of context-aware machine translation, they are not without limitations: First, the previous test suites are limited in size, with the largest being Parcorfull (Lapshinova-Koltunski et al., 2018) which has 82,379 words from three different domains. Moreover, it is often the case that test sets are annotated based on a small portion of a parallel corpus that comprises documents. However, the parallel corpus from which previous test sets are selected is generally quite small, which makes it impossible to train large in-domain NMT models. In contrast, our test set is based on the ultra-large-scale parallel corpus BWB, which is twice as large as the OpenSubtitles fr-en corpus (Lison et al., 2018) upon which most previous challenge sets were based (see Tab. 9). This enables us to differentiate between the source of mistranslation being underfitting or the model not having the capability to model discourse structure. In addition, the scope of most test sets has been restricted to a single discourse phenomenon (the majority of which focus on pronoun translation), which makes it impossible to objectively compare a model's translation capability for different discourse phenomena in the same language and domain. For example, it is hard to decipher whether the model misinterpreted a pronoun's gender as a result of a coreference resolution error or as a consequence of misjudging the gender of the entity. In sharp contrast to this, we have four annotation layers on the same parallel corpus.

Monolingual Corpora with Discourse Annotations. Linguistically annotated corpora have contributed significantly to the advancement of key natural language technologies such as named entity recognition (Tjong Kim Sang and De Meulder, 2003), coreference resolution (Lee et al., 2017), and discourse parsing (Surdeanu et al., 2015). The majority of evaluation has, however, only been conducted on monolingual corpora such as the BBN named entity and pronoun coreference corpus (Weischedel and Brunstein, 2005), the Penn Discourse Treebank (Miltsakaki et al., 2004; Webber et al., 2019), and the OntoNotes corpus (Hovy et al., 2006). And yet, languages differ considerably regarding the discourse phenomena they exhibit. In particular, different languages have different linguistic features that influence the application of cohesive devices, and there are language-specific constraints governing the choice of referring expressions. For instance, Fig. 1 demonstrates a prominent feature distinguishing Chinese from English. In Chinese, it is a common practice to omit pronouns, and an anaphoric link can be inferred from context without explicit assertion. English, in contrast, does not generally allow the omission of pronouns. Our dataset contains aligned parallel discourse annotations in two languages, allowing analysis of the transferability of current NER and coreference resolution models across languages.

7 Conclusion

In this paper, we introduced the bilingual parallel corpus BWB-test, which includes annotations of various discourse phenomena. Our analysis of these annotations revealed the significant challenges posed by discourse phenomena to MT. Therefore, we advocate for greater attention on discourse coherence and consistency of the outputs of NMT models. The main discourse challenges faced by MT include entity consistency, entity recognition, anaphoric information loss, and coreference. Additionally, the BWB corpus, with its rich discourse annotations, serves as a valuable resource for a variety of purposes, including studying the transferability of named entity recognition and coreference resolution, as well as the development of multilingual structured prediction models.

Limitations

There are several limitations to the current study. Firstly, as for now, the corpus used in this study only consists of a single language pair. Secondly, the coherence of the MT systems was evaluated using a fine-tuned conference model, as no annotations were available for the MT outputs. However, as shown in Tab. 8, the fine-tuned conference model is not perfect and may affect the quality of our coherence evaluation. Thirdly, this paper focuses on using discourse annotations to reveal and analyze discourse phenomena and the challenges they present to machine translation. Using the annotations to improve MT models is beyond the scope of this study and is left for future work.

Ethical Considerations

The annotators were paid a fair wage, and the annotation process did not solicit any sensitive information from the annotators. In regard to the copyright of our dataset, as stated in the paper, the crawling script that we plan to release will allow others to reproduce our dataset faithfully and will not be in breach of any copyright. In addition, the release of our annotated test set will not violate the doctrine of **Fair Use** (US/EU), as the purpose and character of the use is *transformative*. Please refer to https://www.nolo.com/legal-encyclopedia/fair-use-the-four-factors.html for relevant laws.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers), pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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 (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Peter Jan-Thorsten, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Kevin Cohen, Arrick Lanfranchi, Miji Choi, Michael Bada, William Baumgartner Jr, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC Bioinformatics*, 18.

Cleotilde Gonzalez, Brad Best, Alice F. Healy, James A. Kole, and Lyle E. Bourne Jr. 2011. A cognitive modeling account of simultaneous learning and fatigue effects. *Cognitive Systems Research*, 12(1):19–32.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Christian Hardmeier. 2012. Discourse in statistical machine translation. *Discours Revue de linguistique, psycholinguistique et informatique*.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022a. 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.

Yuchen Eleanor Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2022b. Investigating the role of centering theory in the context of neural coreference resolution systems. *arXiv preprint arXiv:2210.14678*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE. Association for Computational Linguistics.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

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

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of 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.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Pro-*

ceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Mihai Surdeanu, Tom Hicks, and Marco Antonio Valenzuela-Escárcega. 2015. Two practical Rhetorical Structure Theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado. Association for Computational Linguistics.

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).

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, UAE.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Hovy Eduard, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2012. OntoNotes Release 5.0.

Miculicich Werlen and Lesly Sadiht. 2021. Discourse phenomena in machine translation. Technical report, EPFL.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Document-Level Parallel Corpora

There are some document-level parallel corpora in the market: TED Talks of IWSLT dataset (Ansari et al., 2020), News Commentary (Tiedemann, 2012), LDC[4] and OpenSubtitle (Lison et al., 2018). The sizes and average length of these corpora are summarized in Tab. 9. Below we review these document-level parallel corpora in detail.

LDC. This corpus consists of formal articles from the news and law domains. The articles ave syntactic structures such as conjoined phrases, which make machine translation challenging. However, the news articles in this corpus are relatively outdated.

IWSLT. This corpus contains the TED Talks that covers the variety of topics. However, it is quite small in scale, which makes training large transformer-based models impractical.

News Commentary. This corpus consists of political and economic commentary crawled from the web site Project Syndicate.[5] However, the scale of this corpus is also quite small. Moreover, there are no parallel Chinese-to-English data available in this corpus.

Europarl. The corpus is extracted from the proceedings of the European Parliament. Only European language pairs are available in this corpus.

OpenSubtitle. This corpus is a collection of translated movie subtitles (Lison and Tiedemann, 2016). Besides being simple and short, the documents in this corpus are usually verbal and informal as well.

BWB. The BWB corpus is the largest corpus in terms of size. Moreover, the sentences and documents in BWB are substantially longer than previous corpora. It is also worth noting that BWB differs from previous corpora in terms of genre – in-depth human analysis shows that it is very challenging for current NMT systems due to its rich discourse phenomena.

Corpus	Genre	Language		Size		Averaged Length		
			#word	#sent	#doc	#word/sent	#sent/doc	#word/doc
IWSLT (Ansari et al., 2020)	TED talk	ZH-EN	4.2M	0.2M	2K	19.5	100	2,100
	TED talk	FR-EN	4.4M	0.2M	2K	20.8	100	2,190
	TED talk	ES-EN	4.2M	0.2M	2K	19.9	100	2,080
	TED talk	DE-EN	4.1M	0.2M	2K	19.3	100	2,070
NewsCom (Tiedemann, 2012)	News	ES-EN	6.4M	0.2M	5K	30.7	40	1,288
	News	DE-EN	6.4M	0.2M	5K	33.1	40	1,288
Europarl (Koehn, 2005)	Parliament	ET-EN	7.3M	0.2M	15K	35.1	13	485
LDC	News	ZH-EN	81.8M	2.8M	61K	23.7	46	1,340
OpenSub (Lison et al., 2018)	Subtitle	FR-EN	219.0M	29.2M	35K	8.0	834	6,257
	Subtitle	EN-RU	183.6M	27.4M	35K	5.8	783	5,245
	Subtitle	ZH-EN	16.9M	2.2M	3K	5.6	733	5,647
BWB (chapter)	Novel	ZH-EN	**460.8M**	**9.6M**	**196K**	**48.1**	49	2,356
BWB (book)	Novel	ZH-EN	**460.8M**	**9.6M**	384	**48.1**	**25.0K**	**1.2M**

Table 9: Statistics of various document-level parallel corpora. The parallel Chinese-English data is highlighted in Cyan .

B Case Study

We provide two example chapters in BWB with coreference annotation in Fig. 8 and Fig. 9. We observe that the BWB dataset poses challenges for NMT in the following ways.

[4] https://www.ldc.upenn.edu

[5] 4https://wit3.fbk.eu

Entity Consistency. There are many named entities in the dataset that have a high repetition rate, such as fictional characters. Therefore, named entity consistency is a significant challenge in machine translation on this dataset. For example, the translations of Weibo and **Qiao Lian** in Fig. 8 are not consistent in context.

Entity Recognition and Retrieval. In addition to the fluency of entity translation, the adequacy of entity translation is another challenge in BWB. In the case of fictional characters with strange names, the NMT model may not correctly *recognize* named entities, resulting in extremely poor translation quality, as in Fig. 9. “Ye Qing Luo” could be literally translated as “night”, “clear”, “fall”; however, it is actually the name of a fictional character. Even though fictional characters are difficult to translate, they are relatively rare throughout the text, so it would be beneficial to abandon the assumption of inter-sentence independence in consideration of global contextual information. One potential way to alleviate this problem is to equip NMT models with an entity recognition module.

Anaphoric Information Loss. Chinese, being one of the pro-drop languages, omits many pronouns, while the English language does not, as shown in Tab. 5. Translating from Chinese to English thus requires context to infer the correct English pronouns to compensate for the anaphoric information loss of sentence-level Chinese-to-English translations.

Morphological Information Loss. Tense information is also frequently absent in Chinese and can only be inferred from context. In general, this problem, which we refer to as morphological information loss, is often encountered when translating from a morphologically poorer language to a morphologically richer one. In the case of Chinese-to-English translation, tense information is often lost, while in other language pairs, such as English-to-French and English-to-German, gender information is often missed since as French and German are morphologically richer than English.

Coreference. In addition, in Fig. 8, we observe that the focus entity of the document is shifting throughout the text (**Qiao Lian** → **Shen Liangchuan** → **Wang Wenhao** → **Shen Liangchuan** → **Song Cheng**), and this information is language-independent, i.e., consistent in source and target. This information could be used to improve the coherence of translation.

C Experiment Setup

We adopt the parameters of Transformer Big (Vaswani et al., 2017) for both SENT and DOC. More precisely, the layers in the big encoders and decoders are $N = 12$, the number of heads per layer is $h = 16$, the dimensionality of input and output is $d_{\text{model}} = 1024$, and the inner-layer of a feed-forward networks has dimensionality $d_{\text{ff}} = 4096$. The dropout rate is fixed as 0.3. We adopt Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$, and set learning rate 0.1 of the same learning rate schedule as Transformer. We set the batch size as 6,000 and the update frequency as 16 for updating parameters to imitate 128 GPUs on a machine with 8 V100 GPU. The datasets are encoded by BPE with 60K merge operations.

D The BWB Corpus

In this section, we describe three stages of the dataset creation process: collecting bilingual parallel documents, quality control and dataset split.

D.1 Bilingual Document Collection

385 Chinese web novels across multiple genres were selected, including action, fantasy, romance, comedy, science fiction, martial arts, etc. The genre distribution is shown in Fig. 2. We then scrape their corresponding English translations from the Internet.[6] The English versions are translated by professional translators who are native speakers of English, and then corrected and aligned by professional editors at the chapter level. The text is converted to UTF-8 and certain data cleansing (e.g., deduplication) is

[6] https://readnovelfull.com

performed in the process. Chapters that contain poetry or couplets in classical Chinese are excluded as they are difficult to translate directly into English. Further, we exclude chapters with less than 5 sentences and chapters where the sequence ratio is greater than 3.0. The titles of each chapter are also removed, since most of them are neither translated properly nor at the document level. The sentence alignment is automatically performed by Bleualign[7] (Sennrich and Volk, 2011). The final corpus has 384 books with 9,581,816 sentence pairs (a total of 461.8 million words).[8]

D.2 Quality Control

We hired four bilingual graduate students to perform the quality control of the aforementioned process. These annotators were native Chinese speakers and proficient in English. We randomly selected 163 chapters and asked the annotators to distinguish whether a document was well aligned at the sentence level by counting the number of misalignment. It is identified as a misalignment if, for example, line 39 in English corresponds to line 39 and line 40 in Chinese, but the tool made a mistake in combining the two sentences. We observed an alignment accuracy rate of 93.1%.

D.3 Dataset Split

We construct the development set and the test set by randomly selecting 160 chapters from 6 novels, which contain 3,018 chapters in total. To prevent any train-test leakage, these 6 novels are removed from the training set. Tab. 1 provides the detailed statistics of the BWB dataset split. In addition, we asked the same annotators who performed the quality control to manually correct misalignments in the development and test sets, and 7.3% of the lines were corrected in total.

E Human Evaluation

We conducted human evaluation on the BWB test set following the protocol proposed by (Läubli et al., 2018, 2020). As stated in §5, we evaluated two units of linguistic context (SENTENCE and DOCUMENT) independently based on their respective FLUENCY and ADEQUACY. We showed raters isolated sentences in random order in the SENTENCE-level evaluation, whereas in the DOCUMENT-level evaluation, we presented entire documents and asked raters to evaluate a sequence of five sequential sentences at a time in order. The ADEQUACY evaluation was based solely on source texts, whereas neither source texts nor references were included in the FLUENCY evaluation.

The ADEQUACY evaluation was conducted by four professional Chinese to English translators, and the FLUENCY evaluation was conducted by four native English revisers. The four translators were different from the professional translators who performed human translation. For human evaluation, we deliberately invited another group of specialists to avoid making judgments biased towards human translation.

We adopted relative ranking because it has been shown to be more effective than direct assessment when conducted by experts rather than crowd workers (Barrault et al., 2019). In particular, raters were presented with the system outputs and were asked to evaluate the system outputs vis-à-vis one another, e.g., to decide whether system A was better than system B (with ties allowed).

By randomizing the order of presentation of the system outputs, we were able to blind the origin of the output sentences and documents. While in the SENTENCE-level evaluation, the system outputs were presented in different orders for each sentence, the DOCUMENT-level evaluation used the same ordering of systems within a document to help raters better assess global coherence.

Additionally, we used spam items for quality control.(Kittur et al., 2008). At the SENTENCE-level, we make one of the five options nonsensical in a small fraction of items by randomly shuffling the order of the translated words, except for 10% at the beginning and end. At the DOCUMENT-level, we randomly shuffle all translated sentences except the first and last sentence at the document level, rendering one of the five options nonsensical. If a rater marks a spam item as better than or equal to an actual translation, this is a strong indication that they did not read both options carefully.

[7] https://github.com/rsennrich/Bleualign

[8] We will release a crawling and cleansing script pointing to a past web arxiv that will enable others to reproduce our dataset faithfully.

	PART1		PART2			PART1		PART2	
ADEQUACY	SENT	DOC	SENT	DOC	FLUENCY	SENT	DOC	SENT	DOC
RATER1		✓	✓		RATER5		✓	✓	
RATER2		✓	✓		RATER6		✓	✓	
RATER3	✓			✓	RATER7	✓			✓
RATER4	✓			✓	RATER8	✓			✓

Table 10: The evaluation units and corresponding raters. RATER1-4 are professional Chinese to English translators and RATER5-8 are native English revisers.

	SENTENCE	DOCUMENT
RATER1-RATER2	.171	.169
RATER3-RATER4	.294	.346
RATER5-RATER6	.323	.402
RATER7-RATER8	.378	.342

Table 11: Inter-rater agreements measure by Cohen's κ.

Each raters evaluated 180 documents (including 18 spam items) and 180 sentences (including 18 spam items). The 180 sentences were randomly sampled from PART1 or PART2. We spited the test set into two non-overlapping subsets, referred to as PART1 and PART2. Note that PART1 and PART2 were chosen from different books. Each rater evaluated both sentences and documents, but never the same text in both conditions so as to avoid repetition priming (Gonzalez et al., 2011). Each document or sentence was therefore evaluated by two raters, as shown in Tab. 10.

We report pairwise inter-rater agreement in Tab. 11. Cohen's kappa coefficients were used:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the proportion of times that two raters agree, and $P(E)$ is the likelihood of agreement by chance.

	SOURCE	REFERENCE	MT
1)	乔恋攥紧了拳头，垂下了头。	**Qiao Lian** clenched **her** fists and lowered **her** head.	**Joe** clenched **his** fist and bowed **his** head.
2)	其实他说得对。	Actually, **he** ***was*** right.	In fact, **he*’s*** right.
3)	自己就是一个蠢货，竟然会相信了网络上的爱情。	⟦**She**⟧ ***was*** indeed an idiot, as only an idiot ***would*** believe that they could find true love online.	⟦**I**⟧ ***am*** a fool, even ***will*** believe the love on the Internet.
4)	她勾起了嘴唇，深呼吸一下，正打算将手机放下，微信上却被炸开了锅。	**She** curled ⟦**her**⟧ lips and took a deep breath. Just when ⟦**she**⟧ was about to put down ⟦**her**⟧ cell phone, a barrage of posts bombarded ⟦**her**⟧ WeChat account.	**She** ticked ⟦**her**⟧ lips, took a deep breath, and was about to put ⟦**her**⟧ phone down, but weChat was blown open.
5)	她点进去，发现是凉粉群，所有人都在@她。	**She** logged into ⟦**her**⟧ account and saw that a large number of fans in the **Shen Liangchuan fan group** had tagged **her**.	She nodded in and found it was a **cold powder group**, and everyone was on **her**.
6)	【乔恋：怎么了?	[**Qiao Lian**: What happened?]	**Joe**: What’s the matter?
7)	【川流不息：乔恋，快看微博头条！微博头条?	[**Chuan Forever**: **Qiao Lian**, look at the headlines on **Weibo**, quickly!]	**Chuan-flowing**: **Joe love**, quickly look at the **micro-blogging** headlines! **Weibo** headlines?
8)	她微微一愣，拿起手机，登陆微博，在看到头条的时候，整个人一下子愣住了！	**She** froze momentarily, then picked up ⟦**her**⟧ cell phone and logged into **Weibo**. When ⟦**she**⟧ saw the headlines, ⟦**her entire body**⟧ immediately froze over again!	**She** took a slight look, picked up the phone, landed on the **micro-blog**, when ⟦**she**⟧ saw the headlines, ⟦**the whole person**⟧ suddenly choked!
9)	剧组发布会。沈凉川应邀出场，导演立马恭敬地迎接过来，客气的跟他说这话，表达着自己对他能够到来的谢意。	**Shen Liangchuan** arrived at the scene after accepting the invitation. **The director** immediately went to greet **him** in a respectful manner, politely welcoming **him** and expressing **his** gratitude for **Shen Liangchuan**’s presence today.	The show’s release. **Shen Liangchuan** was invited to appear, **the director** immediately greeted **him** with respect, politely said this to **him**, expressed **his** gratitude for **his** arrival.
10)	对沈凉川没有说话，看向不远处的王文豪。	**Shen Liangchuan** did not speak. Instead ⟦**he**⟧ looked at **Wang Wenhao**, who was nearby.	**Shen Liangchuan** did not speak, look not far from **Wang Wenhao**.
11)	王文豪出事以后，所有的作品全部下架，而这一部剧还能播出，是因为王文豪在里面友情饰演的男三号戏份很少，几乎可以忽略不计。	After **Wang Wenhao**’s scandal broke, every film ⟦**he**⟧ starred in had been taken down. Only this show ***could*** still be broadcasted, as **Wang Wenhao** ***had*** a supporting role in it and ***was*** practically unnoticeable.	After **Wang Wenhao**’s accident, all the works were off the shelves, and this play ***can*** also be broadcast, because **Wang Wenhao** in the friendship played by the male no. 3 play ***is*** very few, almost negligible.
12)	剧组根本就没有邀请王文豪，可他却不知道从哪里拿到了邀请函，自己堂而皇之的进来了。他当然要进来了。	In fact, the cast and crew hadn’t even invited **Wang Wenhao**. However, **he** had obtained a copy of the invitation letter somehow, and ***strode*** imposingly into the venue anyway.	The crew did not invite **Wang Wenhao**, but **he** did not know where to get the invitation, **his** own entrance. Of course **he*’s*** coming in.
13)	这是他最后的机会了。	After all, this ***was*** **his** final chance.	This ***is*** **his** last chance.
14)	丑闻闹出来，几乎所有的广告商和剧组都跟他毁约。	After ⟦**his**⟧ scandals broke, practically every advertiser and filming crew wanted to break their contracts with **him**.	The scandal broke, and almost all advertisers and crews broke ⟦**his**⟧ contract with **him**.
15)	他现在宁可拍男三号，也不想就此沉寂。	**He** would rather take a supporting role than fade out into obscurity.	**He** would rather shoot the men’s number three now than be silent about it.
16)	因为他的事情，根本就压不下去。	That was because the scandals surrounding **him** ***would*** never disappear.	Because of **his** affairs, there ***is*** no pressure.
17)	所以王文豪在发布会上，到处讨好别人。	Thus, **Wang Wenhao** went around trying to curry favor with everybody at this press conference.	So **Wang Wenhao** tried to please others at the press conference.
18)	沈凉川穿着一身深灰色西装，面色清冷，手里端着一个高脚香槟杯，站在桌子旁边，整个人显得格外俊逸，却也格外的清冷，让周围的人都不敢上前搭讪。	**Shen Liangchuan** was wearing a dark grey suit and ⟦**he**⟧ had a cold expression. ⟦**He**⟧ was holding a champagne glass and was currently standing beside a table. ⟦**He**⟧ looked exceptionally stylish, but also exceptionally icy. As a result, none of the people around ⟦**him**⟧ dared to approach ⟦**him**⟧.	**Shen River** was wearing a dark gray suit, ⟦**his**⟧ face was cold, and ⟦**he**⟧ was holding a tall champagne glass in ⟦**his**⟧ hand. Standing beside the table, the whole person looked extraordinarily handsome, but also extraordinarily cold, so that people around ⟦**him**⟧ did not dare to approach ⟦**him**⟧.
19)	可如果能注意到他，就会发现他的视线，却总是若有似无的飘到王文豪身上。	If anyone ***had*** paid attention to **him**, they ***would*** have noticed that **his** gaze ***kept*** drifting over to **Wang Wenhao**.	But if you ***can*** notice **him**, you ***will*** find **his** sight, but always if there ***is*** nothing floating to **Wang Wenhao** body.
20)	宋城站在他的身边，察觉到这一点以后，就忍不住拽了拽他的胳膊。	**Song Cheng** stood at **his** side. After noticing ⟦**his**⟧ behavior, ⟦**he**⟧ ***could*** not help but pinch **his** arm.	**Songcheng** stood by his side, aware of this, ***can*** not help but pull **his** arm.
21)	沈凉川淡淡回头，看向他，目露询问。	**Shen Liangchuan** turned around and looked at **him** casually, with a questioning face.	**Shen Liangchuan** faint lying back, looked at **him**, blind inquiry.
22)	“沈哥，您到底是要干什么啊？能不能告诉我，好让我有个心理准备。您这样突然跑过来参加这么一个小剧组的发布会，又什么都不说就这么杵着，我心里的慌。”	“**Brother Shen**, what are you planning to do? Can you tell me beforehand so that I can prepare myself mentally. You suddenly decide to come and attend such a small-scale press conference, yet you have been completely silent and are now just standing here and doing nothing? My heart is beating anxiously right now.”	**Shen brother**, what the hell are you doing? Can you tell me so that I have a mental preparation. You suddenly ran over to attend the launch of such a small group, and said nothing so, I panicked.
23)	沈凉川听到这话，抿了一口香槟，接着，将香槟杯放下。	After **Shen Liangchuan** heard him speak, he sipped a mouthful of champagne and put the glass down.	**Shen** Said, took a sip of champagne, and then put the champagne glass down.
24)	旋即，他迈开了修长的步伐。	Then, **he** walked away in long strides.	Immediately, **he** took a slender step.
25)	宋城的心都提了起来，紧跟在他身后。沈凉川一步一步往前，走到了前方。	**Song Cheng** was extremely nervous and followed **him**. **Shen** Liangchuan walked forward, one step at a time, until ⟦**he**⟧ reached the front of the room.	**Song Cheng**’s heart was raised and followed immediately behind **him**. **Shen Liangchuan** step by step forward, walked forward.
26)	王文豪正在跟别的三流小明星套近乎，那个人询问，“听说你打了一个狗仔？”	**Wang Wenhao** ***was*** currently ingratiating himself with a C-list celebrity. The celebrity asked, “Hey, I heard that you beat a paparazzi?”	**Wang Wenhao** ***is*** with other third-rate star-studded sets, the man asked, "I heard you hit a paparazzi?" "
27)	“对啊，现在的狗仔就是惹人厌恶，我早就想动手教训他们了！“你这样，就不怕跟他们结仇啊？”	“Yeah, the paparazzi nowadays are so disgusting. **I** have wanted to teach them a lesson myself for some time now!” "Are not **you** afraid of becoming an enemy of them?"	"Yeah, the paparazzi now are disgusting, **I** have long wanted to teach them a lesson! "If **you** are like this, **you** are not afraid of enmity with them? "
28)	”我都这样了，我怕什么？当初沈影帝以正当防卫为借口，将一名狗仔打了，告到了法庭上去不也不了了之吗？“	“**I**’ve already done it, so what should **I** be scared of? That time **Best Actor Shen** beat up a reporter, ⟦**he**⟧ claimed that it was in self-defence so that ⟦**he**⟧ would have an excuse if he got sued, right? At that time, nobody said anything”	"**I**’m already like this, what am **I** afraid of? Back then, **Yingdi Shen** beat up a paparazzi under the pretext of self-defense, and sued **him** in court, wouldn’t it be over?
29)	王文豪说到这里，嘿嘿一笑。还想说什么，忽然察觉到身后有人靠近。	As **Wang Wenhao** spoke, ⟦**he**⟧ laughed heartily. Just as ⟦**he**⟧ was about to continue speaking, ⟦**he**⟧ suddenly felt a presence approaching ⟦**him**⟧ from behind.	When **Wang Wenhao** said this, ⟦**he**⟧ smiled. ⟦**He**⟧ was about to say something, but suddenly noticed someone approaching behind ⟦**him**⟧.
30)	他回头，就看到⟦沈凉川⟧，眼瞳一缩，舔着脸笑，却见⟦沈凉川⟧往前一步，一把揪住他的衣领，接着一拳头对着他的脸就砸过来！	He turned around and saw **Shen Liangchuan**. ⟦**His**⟧ eyes narrowed and attempted to smile at ⟦**him**⟧. However, **Shen Liangchuan** took a step forward, grabbed **his** tie and threw a punch at **his** face!	**He** looked back, ⟦**he**⟧ saw **Shen Liangchuan**, eyes shrink, licking ⟦**his**⟧ face and smiling, but saw Shen Liangchuan a step forward, a holding **his** collar, and then a fist to **his** face hit!

Figure 8: An example chapter in BWB. The same entities are marked with the same color. Pronoun omissions are marked with ⟦⟧. The mistranslated verbs are marked with ***teal***, and the mistranslated named entities are marked with .

	SOURCE	REFERENCE	MT
1)	夜清落浑身上下都传来剧烈的疼痛感，宛如千万把利刃，切割着她的身体。	**Ye Qing Luo** suddenly felt an excruciating sharp pain tormenting **⟦her⟧** entire body. It seemed as if a million sharp blades were slashing at **her**.	**The night fell** all over the body came a sharp pain, like a thousand sharp blades, cutting **⟦her⟧** body.
2)	尤其心脏那处，像是有着一团烈火，在体内燃烧，肆意的烧灼着她的一切。	**⟦Her⟧** heart felt as if it was burning and that flame threatened to burn everything.	Especially the heart, like a fire, burning in the body, burning **her** everything.
3)	夜清落想要伸手揉揉眉心，却发现自己全身虚软无力，连手指头都没法动弹。	**Ye Qing Luo** wanted to reach out but **⟦she⟧** found that **⟦she⟧** couldn't move. **⟦She⟧** felt so weak that **⟦she⟧** could not even lift a finger.	**Night clear** wants to reach out to rub the eyebrows, but found that **⟦their⟧** whole body is weak, even fingers can not move.
4)	耳边，隐隐约约传来模糊不清的声音。“四少爷，这么做真的没事吗？”	In **⟦her⟧** ears, there was a vague sound. “Fourth Young Master, are you sure it's really alright to do this?”	in my ear, vaguely came a vague sound. "Four Masters, is it really okay to do this?"
5)	“怕什么！”	What are you so afraid of!	"Afraid of what! "
6)	“太子殿下怎么可能看得上这个废物？要不是看她担着三泉宗少宗主的名头，太子殿下早就将她踹了。”	How can His Royal Highness put such a good-for-nothing waste in his eyes? If not for her status as the Three Spring's Lord's daughter, do you think he would even bother with her?	"How could His Highness see this waste? Had it not been for her bearing the name of the Lord of Sanquan Zong, His Royal Highness would have taken her. "
7)	“小废物可是夜四小姐亲自送来的，那碗药也是夜四小姐亲自喂的，太子殿下……说不定也是知情的。”	She's a present personally sent to us by the Fourth Young Miss of the Ye family. That bowl of medicine was also personally fed to her by the Fourth Young Miss. . . so this matter. . . may also have been known by His Highness.	"Small waste is the night four Miss personally sent, that bowl of medicine is also the night four Miss personally fed, His Royal Highness ... ... Maybe it's also informed. "
8)	妈哒！这是个什么鬼情况！	Damn it, what kind of crappy drama is happening?	Damn! What a ghost situation!
9)	夜清落紧蹙着细眉，努力的睁开沉重的眼皮。	**Ye Qing Luo** scrunched **⟦her⟧** brows together, mustering all her energy to lift her heavy eyelids.	**The night fell** with a thin brow, and tried to open the heavy eyelids.
10)	刚一睁开，就被极为耀眼的光芒，刺的她眼皮一痛。	Just as **⟦she⟧** opened them, **her** eyes were stung by a bright light.	As soon as **⟦she⟧** opened, she was the bright light, stabbing **her** eyelids a pain.
11)	一幕幕陌生的画面，宛如走马灯在脑海里不断的回旋。	Suddenly, **⟦her⟧** mind reeled and it felt as if a there was an explosion in **⟦her⟧** head. Fragments of unfamiliar pictures and scenes started to flood **⟦her⟧** mind.	A scene of strange scenes, like walking horse lights in the mind of the constant swing.
12)	斑斓画面一过，那些景象，像是强行插入的记忆，快速的在脑海里重叠，旋即渐渐归于平静。	It continued to flash in **⟦her⟧** mind non stop when **⟦she⟧** suddenly realized that these fragments were forcing themselves into **⟦her⟧** own memories as they melded and fused together. Soon, everything was calm.	The scene, those scenes, like forced insertion of memories, quickly overlapped in the mind, and gradually fell calm.
13)	接收完这些记忆后，夜清落再次睁开了眼睛。	After **⟦she⟧** received these memories, **Ye Qing Luo** tried to pry open **⟦her⟧** eyes again.	After receiving these memories, **the night fell** and opened **⟦his⟧** eyes again.
14)	这一次，她的眼睛适应了屋内的烛光摇动，灯火明耀。	This time, **her** eyes adapted quickly and focused on the candles.	This time, **her** eyes adapted to the candlelight in the house, and the lights lit up.
15)	奢华精致的房间，颇有古风意味，白色纱帐随风浮动。	It was a luxurious room, exquisitely decorated in an ancient flavour.	Luxurious and sophisticated rooms, quite ancient, white yarn book with the wind floating.
16)	四颗夜明珠伫立在房间四个角落，散发着莹莹的光芒。	Four night pearls stood on each corner of the room, illuminating the room aglow.	Four night pearls stand in the four corners of the room, emitting a bright light.
17)	最忒玛坑爹的是，她现在四肢大敞的躺在一张圆木桌上，自己原本那具引以为傲的身材，变成了纤细娇弱的少女身体，只着了一件白色长袍。	What seemed even more shocking was that **she** was now lying atop a wooden table, **⟦her⟧** voluptuous figure was now instead a slender and delicate girl's body, with only a white robe on.	The most beautiful father is that **she** is now lying on a round wooden table with **⟦her⟧** limbs, **⟦her⟧** original proud figure, into a delicate girl's body, only a white robe.
18)	靠！她这样躺在桌子上，是当做食物，等待着别人把她吃掉吃掉吃掉咩?	What the?! Lying on the table like this seemed to be as if **she** was some kind of food and waiting for someone to eat **her** up?	By! **She** lay on the table like this, as food, waiting for others to eat **her** and eat **her**?
19)	尼玛她不就是从飞机上跳了下来，就把自己的灵魂转移到了这么一具小屁孩的身体里了?	Damn it! ***Didn't*** **she** jump off the plane? How did her soul end up in this little girl's body?	***Does*** Neymar just jump off the plane and transfer her soul to such a little?
20)	而且，刚穿越就这么惨，还能不能愉快的玩耍了?	And since **⟦she⟧**'s been given a new lease of life, ***wasn't*** this start a bit too miserable?	And, just through such a miserable, ***can*** not be happy to play?
21)	夜清落想动，使不上劲。	**Ye Qing Luo** tried to move again, but **⟦she⟧** could not muster any strength at all.	**The night fell** and tried to move, so that **⟦it⟧** could not be vigorous.
22)	门外的那两人刚刚说过，她被人下了药。	As what the two men had said earlier, **she** was drugged.	The two men outside the door had just said that **she** had been drugged.
23)	而且下药的那人，是这具身体的四姐，并且亲自把她送到了这里，让人来玷污她！	And what's more, the drug was personally administered by **⟦her⟧** very own sister and even sent **her** here as a present to these men to tarnish **her**!	And the man who took the medicine, is the body of the four sisters, and personally sent **her** here, let people to tarnish **her**!
24)	夜清落快速从记忆中搜寻自己所需要的记忆。	**She** quickly searched through **⟦her⟧** memories.	**Night clearing** quickly searches for the memories **⟦you⟧** need from memory.
25)	门外那人，是玄者四大家之一尉迟世家的四少爷尉迟涯，此人风流成性，游手好闲，就是一个大写的纨绔少爷。	The man who was outside, was from the **Yuchi** clan, one of the four major family clans. **He** was the Fourth Young Master of the **Yuchi** family, a well known foppish playboy who spent, **⟦his⟧** time idling about - **Yuchi Ya**.	The person outside the door, is one of the four people of the **Xuan**, the captain of the late family of the four young captain siaa, this person is a popular, idle, is a capital is a master.
26)	把她送到尉迟涯面前，根本就是送羊入“狼”口！	Sending her to **Yuchi Ya** was simply putting a sheep in front of a wolf's mouth!	Send **her** to the captain in front of the **late ya**, is simply to send sheep into the "wolf" mouth!
27)	四姐？还有那个所谓的未婚夫?	Fourth Sister? And who was her so called fiance?	Four sisters? And that so-called fiance?
28)	呵！	Ah!	Oh!
29)	都给她等着！	They better be good and wait for her to return this favour back many folds!	Just give her a wait!
30)	夜清落微眯起锋锐的眼睛，强压住身体传来的剧痛和麻木，努力的控制着四肢。	**Ye Qing Luo**'s gaze sharpened and she exerted a strong pressure, using all her effort to regain control of her limbs.	**The night fell** slightly with sharp eyes, pressed the body from the sharp pain and numbness, and tried to control the limbs.
31)	“吱呀”一声门响，尉迟涯走了进来。	[Squeak-] The door opened and **Yuchi Ya** strode in.	"Squeaky" a door rang, **the captain** came in late.
32)	听脚步声，少说也有五人以上。	From the sound of the footsteps, **⟦she⟧** gathered that there were at least five or more people with **him**.	Listen to the footsteps, less say there are more than five people.
33)	“小废物，哥哥现在就来疼你！”	**Little Waste**, **brother** is here to dote on you. . .	"Little waste, **brother** is here to hurt you now!"
34)	尉迟涯走到桌子边，直接伸手扯她身上的白袍。	**He** leered and slowly walked over to the table and immediately reached for **her** white robe.	**The captain** walked up to the table and reached directly for **her** white robe.
35)	夜清落冰冷的眼神锐利，沙哑着嗓音，吐出一个字：“滚！”尉迟涯听到她的声音，笑得更是嚣张：“还没昏死过去？也好，也好！”	When **Yuchi Ya** heard **her** voice, **⟦he⟧** laughed even more lasciviously, with a hint of arrogance, "You're awake? Very good, very good!"	**The night clear** cold eyes sharp, hoarse voice, spit out a word: "Roll! When **⟦he⟧** heard her voice, **⟦he⟧** smiled more loudly: "Haven't passed out yet? Good, good!"

Figure 9: Another example chapter in BWB. This example is even more difficult for MT since it fails to recognise the main character “Ye Qing Luo” as a named entity.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2 + Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2 + Appendix

C ☑ Did you run computational experiments?

5

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. we used existing models

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

in lab experts

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Sec. 3.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Sec. 3.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Sec. 3.