Abstract

Machine translation (MT) has almost achieved human parity at sentence-level translation. In response, the MT community has, in part, shifted its focus to document level translation. However, the development of document-level MT systems is hampered by the lack of parallel document corpora. This paper describes BWB, a large parallel corpus first introduced in Jiang et al. (2022), along with an annotated test set. The BWB corpus consists of Chinese novels translated by experts into English, and the annotated test set is designed to probe the ability of machine translation systems to model various discourse phenomena. Our resource is freely available, and we hope that it will serve as a guide and inspiration for more work in the area of document-level machine translation.

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

1 Introduction

Machine translation (MT) has made significant progress in the past few decades. Neural machine translation (NMT) models, which are able to leverage abundant quantities of parallel training data, have been one of the main contributors to this progress (Luong et al., 2015; Vaswani et al., 2017; Zhang et al., 2018, inter alia). Unfortunately, the majority of available parallel corpora contain sentence level translations. As a result, models trained on these corpora translate text quite well at the sentence level, but perform poorly when the entire document translation is seen in context (Voita et al., 2017; Miculicich et al., 2018; Maruf and Haffari, 2018; Voita et al., 2019a, inter alia). Although such approaches have achieved some improvements, they nonetheless suffer from a dearth of document level training data. Take the WMT news translation task as an example. The document level news commentary corpus (Tiedemann, 2012) only contains 6.4M tokens while the available sentence-level training data has around 825M tokens. To alleviate this problem, we collect a large document level parallel corpus that consists of 196K paragraphs from Chinese novels translated into English. As shown in Fig. 1, it is the largest document-level corpus to the best of our knowledge. Additionally, an in-depth human analysis shows, it is very challenging for current NMT systems due to its rich discourse phenomena.

To better evaluate context-aware MT models, we further annotate the test set with characteristic discourse-level phenomena, namely ambiguity and ellipsis. The test set is designed to specifically measure models’ capacity to exploit such long range linguistic context. We then conduct systematic

Figure 1: Comparing sizes of various document-level parallel corpora. BWB is the to-date largest parallel corpus.

vs micro-blog. Qiao Lian vs Joe vs Joe love).

1 This discourse phenomenon is referred as entity consistency. There are other discourse dependencies that MT fails to capture, such as tense cohesion, ellipsis and coreference. We have left explanations of discourse phenomena to §3.1.
We first select 385 Chinese web novels across multiple genres, including action, fantasy, romance, comedy, science fictions, martial arts, etc. The genre distribution is shown in Fig. 3. We then crawled their corresponding English translations from the Internet. 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 performed in the process.

Table 1: Statistics of various document-level parallel corpora. For corpora that do not contain ZH-EN parallel documents (NewsCommand and Europarl), we report the statistics for ZH-EN. For corpora that do not contain ZH-EN parallel documents, we report the statistics of their (largest) available language pairs (DE-EN and ET-EN). w, s and d stand for word, sentence and document, respectively. The full list is in Tab. 5.

| Corpus  | Genre | Size | Averaged Length | Error Type | Description | An |
|---------|-------|------|-----------------|------------|-------------|----|
| IWLSIT | TED talk | 4.2M | 0.2M | 2K | ENTITY | 43.3% | error(s) due to the mistranslation of named entities. |
| NewsCom | news | 6.4M | 0.2M | 5K | TENSE | 38.7% | error(s) due to incorrect tense. |
| Europarl | Parliament | 7.3M | 0.2M | 15K | ZEROPRO | 17.3% | error(s) caused by the omission of pronoun(s). |
| LDC | News | 81.8M | 2.8M | 61K | AMBIGUITY | 7.3% | there are some ambiguous span(s) that is(are) correct in the stand-alone sentence but wrong in context. |
| OpenSub | Subtitle | 16.9M | 2.2M | 3K | ELLIPSIS | 4.0% | error(s) caused by the omission of other span(s). |
| BWB | novel (chapter) | 460.8M | 9.6M | 196K | SENTENCE | 51.3% | sentence-level error(s). |
| BWB | novel (book) | 460.8M | 9.6M | 384 | NO ERROR | 17.1% | no errors. |

Figure 2: Part of a chapter in BWB. The same entities are marked with the same color. Pronoun omissions are marked with [ ]. The mistranslated verbs are marked with teal. The same entities are marked with the same color. Pronoun omissions are marked with [ ].

Table 2: The types of NMT errors and their description. # represents the proportion of the error in the BWB test set. ✓ indicates “with annotation”.

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. The final corpus has 384 books with 9,581,816 sentence pairs (a total of 461.8 million words). 4

2.2 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%.

2.3 Dataset Split

We construct the development set and the test set by randomly selecting 80 and 79 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. 6 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.

3 Dataset Analysis and Annotation

As part of this section, we analyze the types of translation errors that can occur in sentence-level NMT outputs, as well as annotate the BWB test set. We also provide analysis on coherence-related properties: numbers of named entities, numbers of pronouns in both English and Chinese, and the relationships of those factors. The annotation was conducted by eight professional translators.

3.1 Translation Errors

The annotators were asked to identify and categorize discourse-level translation errors made by a state-of-the-art commercial NMT system, i.e. errors that are only visible in context larger than individual sentences. The annotators followed the following guideline for this error annotation:

1. Identify cases that have translation errors: label examples as NO ERROR only if they meet both the criteria of adequacy and fluency as well as the global criterion of coherence.
2. Identify whether the translation error is at the sentence level or document level (or both): SENTENCE are examples that are already not adequate or fluent as stand-alone sentences.
3. Categorize the DOCUMENT examples in accordance with the discourse phenomena, mark the corresponding spans in the reference (English) that cause the MT output to be incorrect, and provide the correct versions.

The types of errors are summarized in Tab. 2.

| Lang | MASCULINE | FEMININE | NEUTER | EPICENE |
|------|-----------|----------|--------|---------|
| EN   | 1633      | 2521     | 608    | 391     |
| ZH   | 654       | 967      | 14     | 118     |

Table 3: The distributions of different types of pronouns in both English and Chinese in the BWB test set.

3.2 Named Entities

Named entities (NEs) are an essential part of sentences in terms of human understanding and readability. The mistranslation of NEs can significantly impact translation output, although evaluation scores (e.g. BLEU) may not be adversely affected. Therefore, we also annotate named entities in the reference documents, following a similar procedure to OntoNotes (Hovy et al., 2006). In total, 2,234 entities are annotated in the BWB test.

3.3 Pronouns

Pronoun translation has been the focus of discourse-level MT evaluation (Hardmeier, 2012; Miculicich Werlen and Popescu-Belis, 2017). As shown in Tab. 3, there are significantly fewer pronouns in Chinese due to its pronoun-dropping property. This poses extra challenges for NMT since the skill of anaphoric resolution is required.

4 Experiments

We carry out evaluation of both baseline and state-of-the-art MT models on BWB and also provide human post-editing performance PE for comparison. The following 6 baselines are adapted:

- SMT: phrase-based baseline (Chiang, 2007).
- BING, GooGLE, BaiDu: commercial systems.
- MT-S: the Transformer baseline that translates sentence by sentence (Vaswani et al., 2017).
- MT-D: the document-level NMT model that adopts two-stage training (Zhang et al., 2018).

Evaluation Metrics Systems are evaluated with automatic standard sentence-level MT metrics (BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), BERT (Zhang* et al., 2020)) and a document-level metric (BLONDE, ?). We also performed evaluation targeted at specific discourse-phenomena.

5 Professional translators were hired to conduct post-editing on the BING outputs. They were instructed to correct only discourse-level errors with minimal modification. MT-S and MT-D are trained on BWB by fairseq (Ott et al., 2019), and the training details are in App.C.
**Table 4**: Results of MT systems and human post-editing on the BWB test set. For discourse phenomena, we report both F1 measure defined in (2) and Exact-Match Accuracy defined in Alam et al. (2021).

**Human Evaluation** The human evaluation on BWB is conducted by 8 professional assessors, i.e. 4 Chinese to English translators and 4 native English revisers. Following the recommendations of Läubli et al. (2020), we evaluate two units of linguistic context (SENTENCE and DOCUMENT) independently based on their respective FLUENCY and ADEQUACY. Span items are used for quality control (Kittur et al., 2008). The results and rankings of the aforementioned systems and human translation (HT) are listed in Fig. 4. The large gap between the performances of HT and MT indicates that the genre of BWB, i.e., literary translation, is very challenging for MT, and NMT systems are far from human parity. MT-D performs significantly better than MT-S, 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-editing of the poorly performing system BING, it is still surprisingly able to achieve better performance than MT-D at the document level. This observation confirms the claim that the discourse phenomena contained in BWB have a huge impact on human judgment of translation quality.

**Evaluation Test Suites for Document-Level MT** Evaluating document-level translation quality is difficult for metrics such as BLEU. Therefore, many test suites that perform context-aware evaluation have been proposed (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., 2019b; Guillou and Hardmeier, 2018, inter alia). However, the scope of most test suites has been restricted to pronouns and limited in size. In contrast, BWB annotates not only pronouns but also other context-sensitive spans that cannot be translated correctly by context-agnostic systems.

**5 Related Work**

**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 and OpenSubtitle (Lison et al., 2018). The sizes and average length of these corpora are summarized in Tab. 1. Detailed descriptions of these corpora are in App. A. BWB 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.

**6 Conclusion**

We presented a newly constructed document level parallel corpus BWB with annotations of various discourse phenomena such as discourse-sensitive spans and named entities in the test set to allow for a fine-grained analysis of document level translation. Experiments show that BWB is challenging for existing NMT models and could serve as a good benchmark for coherent document level translation.
Limitations

As of now, this corpus consists of only Chinese-to-English data. In addition, as illustrated in App. B, coreference resolution is also crucial to document level machine translation. We have left coreference annotation to 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

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

Ebrahim Ansari, Amitai 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.

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.

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.

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. 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, Minminya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

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.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qilan 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.

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).

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Annette Rios Gonzales, Laura Mascarrell, 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.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and
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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. 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.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Miculicich Werlen and Lesly Sadiht. 2021. Discourse phenomena in machine translation.

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 Dataset Statistics

A.1 Existing Corpora

Here we review the existing document-level parallel corpora mentioned in §5 in detail. The full list of their statistics is provided in Tab. 5.

**LDC** This corpus consists of formal articles from the news and law domains. The articles have 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. 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.

A.2 Statistics of BWB

The statistics of the training, development and test sets of BWB is provided in Tab. 6.

B Case Study

We provide two example chapters in BWB with coreference annotation in Fig. 5 and Fig. 6. We observe that the BWB dataset poses challenges for NMT in the following ways.

**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. 5 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. 6. “Ye Qing Luo” could be literally translated as “night”, “clear”, “fall”; however, it is actually a fictional characters.

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. 3. 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. 5, 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 MT-S and MT-D. More precisely, the layers in the big encoders and decoders are \( N = 12 \), the number of heads per
Table 5: Statistics of various document-level parallel corpora. The parallel Chinese-English data is highlighted in Cyan.

Table 6: Statistics of the training, development and test sets of BWB.

layer is $h = 16$, the dimensionality of input and output is $d_{\text{model}} = 1024$, and the inner-layer of a feedforward 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 Human Evaluation**

We conducted human evaluation on the BWB test set following the protocol proposed by (Läubli et al., 2018, 2020). As stated in §4, 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 invite 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.

Each rater 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 split 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. 7.

We report pairwise inter-rater agreement in Tab. 8. Cohen’s kappa coefficients were used:

\[
\kappa = \frac{P(A) - P(E)}{1 - P(E)}
\]  

(1)

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

### Table 7: The evaluation units and corresponding raters.

RATER1-4 are professional Chinese to English translators and RATER5-8 are native English revisers.

| ADEQUACY | PART1 | PART2 |
|----------|-------|-------|
| RATER1   | ✓     | ✓     |
| RATER2   | ✓     | ✓     |
| RATER3   | ✓     | ✓     |
| RATER4   | ✓     | ✓     |

| FLUENCY  | PART1 | PART2 |
|----------|-------|-------|
| RATER5   | ✓     | ✓     |
| RATER6   | ✓     | ✓     |
| RATER7   | ✓     | ✓     |
| RATER8   | ✓     | ✓     |

### Table 8: Inter-rater agreements measure by Cohen’s \(\kappa\).

| RATER     | SENTENCE | DOCUMENT |
|-----------|-----------|----------|
| RATER1-RATER2 | .171     | .169     |
| RATER3-RATER4 | .294     | .346     |
| RATER5-RATER6 | .323     | .402     |
| RATER7-RATER8 | .378     | .342     |
Figure 5: An example chapter in BYVB. The same entities are marked with the same color. Pronoun omissions are marked with [ ]. The mistranslated verbs are marked with ted, and the mistranslated named entities are marked with .
Ye Qing Luo suddenly felt an excruciating sharp pain tumbling [her] entire body. It seemed as if a million sharp blades were slashing at her. [Her] heart felt as if it was burning and that flame threatened to burn everything. 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. In [her] ears, there was a vague sound. “Fourth Young Master, are you sure it’s really alright to do this?”

What are you so afraid of? 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?

Ye Qing Luo was also personally fed to her 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. Damn it, what kind of crap is happening?

Ye Qing Luo scratched [her] brows together, muttering. They all the energy to lift her heavy eyelids. Just as [she] opened them, her eyes were stuffed by a bright light. Suddenly, [her] mind reeled and it felt as if there was an explosion in [her] head. Fragments of unfa- miliar pictures and scenes started to flood [her] mind. 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.

After [she] received these memories, Ye Qing Luo tried to pry open [her] eyes again. This time, [her] eyes adapted quickly and focused on the candlelight!

It was a luxurious room, exquisitely decorated in an ancient flavour.

As soon as [she] was woken up, there was no waste in his eyes? If not for her status as His Highness, how can His Royal Highness put such a good-for-nothing waste in his eyes? If not for her status as His Highness, how can His Royal Highness put such a good-for-nothing waste in his eyes?

K

Ye Qing Luo glanced in the mirror, as well as to other places.

Young Master, are you sure it’s really alright to do this?

Ye Qing Luo just jump off the plane and transfer her soul to such a little girl’s body?

What the?! Lying on the table like this seemed to be as good as she was dead.

After receiving these memories, the night fell and opened [her] eyes again. This time, [her] eyes adapted quickly and focused on the candlelight!

The night fell all over the body came a sharp pain, like a knife piercing all over the body.

The night clear and the lights lit up. Luxurious and sophisticated rooms, quite ancient, white yarn book with the wind floating. Four night pearls stand in the four corners of the room, emitting a bright light.

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?

By! She lay on the table like this, as food, waiting for others to eat her and eat her?

Does Neymar just jump off the plane and transfer her soul to such a little girl’s body?

And, just through such a miserable, can’t be happy to play?

The night fell and tried to move, so that [it] could not move?

The two men outside the door had just said that she had been drugged.

And the man who took the medicine, is the body of the four sisters, and personally sent her here, let people to tamish her!

Ye Qing Luo quickly searches for the memories [you] need from memory.

The person outside the door, is one of the four people of the Xiayan, the captain of the late family of the four young captain sias, this person is a popular, idle, is a capital is a master.

Send her to the captain in front of the late ya, is simply to send sheep into the “wolf” mouth!

Four sisters! And that so-called fiancée!

Oh! Just give her a wait!

The night fell slightly with sharp eyes, pressed the head on the sharp pain and numbness, and tried to control the limbs.

“Squirmy” a door rang, the captain came in late. Listen to the footsteps, less say there are more than five people.

“Little waste, brother is here to hurt you now!”

The captain walked up to the table and reached directly for her white robe.

When Ye Qing Luo heard her voice, [he] laughed even more lasciviously, with a hint of arrogance, “You’re awake? Very good, very good!”