BLOD E: An Automatic Evaluation Metric for Document-level Machine Translation

Yuchen Eleanor Jiang1∗ Tianyu Liu1 Shuming Ma2 Dongdong Zhang2 Jian Yang2 Haoyang Huang2 Rico Sennrich3 Mrinmaya Sachan1 Ryan Cotterell1 Ming Zhou4

1ETH Zürich 2Microsoft Research Asia 3Universität Zürich 4Langboat.com

{yuccjiang,tianyu.liu,ryan.cotterell,mrinmaya.sachan}@inf.ethz.ch
{shuming.ma,dongdong.zhang,t-jianya,haohua}@microsoft.com
sennrich@cl.uzh.ch ming.zhou@chuangxin.com

Abstract

Standard automatic metrics, e.g., BLEU, are not reliable for document-level MT evaluation. They can neither distinguish document-level improvements in translation quality from sentence-level ones, nor identify the discourse phenomena that cause context-agnostic translations. This paper introduces a novel automatic metric BLOD E to widen the scope of automatic MT evaluation from the sentence to the document level. BLOD E takes discourse coherence into consideration by categorizing discourse-related spans and calculating the similarity-based F1 measure of categorized spans. We conduct extensive comparisons on a newly constructed document-level translation dataset BWB. The experimental results show that BLOD E possesses better selectivity and interpretability at the document-level, and is more sensitive to document-level nuances. In a large-scale human study, BLOD E also achieves significantly higher Pearson’s r correlation with human judgments compared to previous metrics.

https://github.com/EleanorJiang/BlonDe

1 Introduction

Over the past years neural machine translation (NMT) models have become the models of choice in Machine Translation (MT; Luong et al., 2015; Vaswani et al., 2017; Zhang et al., 2018, inter alia). Although some recent work (Hassan et al., 2018; Popel, 2018; Bojar et al., 2018) suggests that NMT has achieved human parity at the sentence level, the reliability of these human-parity claims was quickly contested by Lüüblle et al. (2018, 2020), showing that there is a larger difference between human and machine translation quality when inter-sentential context is taken into account.

Therefore, document-level machine translation has received increased attention in the MT community. However, despite various modeling advances, the MT community still lacks an efficient and effective evaluation metric for document-level translation. Standard evaluation metrics for MT, e.g., BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005), focus on the quality of translations at the sentence level and do not consider discourse-level features.

Thus, test suites that perform context-aware evaluation by targeting characteristic discourse-level phenomena 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., 2019; Guillou and Hardmeier, 2018, inter alia) for document-level MT. However, such test suites need to be re-created for new domains or even language pairs, and their construction can be very labor-intensive. We still lack an
easy-to-use automatic metric that can reliably discriminate the quality of document-level translation.

In this paper, we curate a large-scale document-level parallel corpus (BWB) from heterogeneous data sources, and quantify document-level translation mistakes by performing a large human study. On this dataset, we found that inconsistency, ellipsis, and ambiguity were the most noticeable phenomena critical for document-level MT, together amounting to 86.73% of MT mistakes. Based on this analysis, we propose BLONDE, an automatic metric that evaluates translation quality at the document level. At the core of the metric is the similarity-based bijection between subsets of reference and system categories, e.g., pronouns, inflected forms, discourse relations and lexicons, and phrases, e.g., named entities. It computes recall, precision and F1, along with the corresponding measure of n-grams. Furthermore, BLONDE can incorporate human annotation easily by computing scores of human-annotated categories in the same way.

We compare BLONDE with 11 other metrics and demonstrate that BLONDE is better at distinguishing between context-aware and context-agnostic MT systems. We also observe that the degree to which BLONDE correlates with sentence-level metrics (e.g., BLEU) is lower than the degree to which the sentence-level metrics correlate with each other. This signals that BLONDE indeed captures additional aspects of translation quality beyond the sentence-level. Finally, our human evaluation also reveals significantly higher Pearson’s r correlation coefficients between BLONDE and human assessments compared to other metrics.

2 BWB: Bilingual Web Book Dataset

To design a metric that is sensitive to document-level phenomena, we first curate a document-level Chinese–English parallel corpus, called BWB (Bilingual Web Books). BWB consists of Chinese web novels across multiple genres (sci-fi, romance, action, fantasy, comedy, inter alia) and their corresponding English translations crawled from the Internet.

Dataset Creation. The novels are translated by professional native English speakers, and are corrected by editors. The sentence alignment of the training set is done by Bluealign (Sennrich and Volk, 2011). We hired four bilingual graduate students to manually evaluate 163 randomly selected documents from the resulting BWB parallel corpus and observe an alignment accuracy rate of 93.1%. We further asked the same batch of annotators to correct such misalignments in both the development and the test set. The details of the corpus creation and quality control are described in Appendix A.

Statistics. Table 1 summarizes the statistics of the BWB dataset. It is a much larger dataset, and contains longer documents and richer discourse phenomena compared to all previous document-level datasets (Lison and Tiedemann, 2016; Koehn and Knowles, 2017; Barrault et al., 2019; Koehn, 2005; Liu and Zhang, 2020). To the best of our knowledge, this is the largest Chinese–English document-level translation dataset to date.

Dataset Split. We treat chapters in our books as documents. The maximum, median, and minimum number of sentences per document are 46, 30 and 18, respectively. To prevent any train–test leakage, we split the dataset into a training, development and a test set such that chapters from the same book are part of the same split. We use 377 books for training, and randomly select 80 and 79 documents from the 3,018 documents in the remaining 6 books as the development and test sets, respectively.

3 Analyzing Discourse Errors

Next, we conduct a human study on the test set of BWB, in which we identify and categorize the discourse errors made by MT systems that are not captured in sentence-level evaluation. This human study is conducted by eight professional translators. The annotators are asked to classify translation errors into DOCUMENT-level and SENTENCE-level errors (some cases can be both). SENTENCE-level

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Train</th>
<th>Test</th>
<th>Dev</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Docs</td>
<td>196,304</td>
<td>80</td>
<td>79</td>
<td>196,463</td>
</tr>
<tr>
<td>#Sents</td>
<td>9,576,566</td>
<td>2,632</td>
<td>2,618</td>
<td>9,581,816</td>
</tr>
<tr>
<td>#Words</td>
<td>325.4M</td>
<td>68.0K</td>
<td>67.4K</td>
<td>460.8M</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the BWB dataset.

3 Analyzing Discourse Errors

Next, we conduct a human study on the test set of BWB, in which we identify and categorize the discourse errors made by MT systems that are not captured in sentence-level evaluation. This human study is conducted by eight professional translators. The annotators are asked to classify translation errors into DOCUMENT-level and SENTENCE-level errors (some cases can be both). SENTENCE-level

3 https://github.com/rsennrich/Bleualign

4 One document in the development set was dropped due to its poor annotation quality.
errors refer to those errors that render the translations to be inadequate or not fluent as stand-alone sentences, while DOCUMENT-level errors reflect a coherence violation across multiple sentences in the document. DOCUMENT-level errors are further categorized according to the linguistic phenomena leading to the lower performance in the context-dependent translation.\(^5\)

Table 2 shows the result of our error analysis. A substantial proportion of translations have document-level errors (71.9%). This supports the claim that BWB contains rich discourse phenomena that common MT systems cannot address. We observe that three categories, i.e., inconsistency (64.4%), ellipsis (20.3%) and ambiguity (7.3%), account for the vast majority of document-level errors. Below we discuss these three categories of DOCUMENT-level errors and the design intuitions behind BLOND\(\text{E}\.\)

**Inconsistency.** We consider two kinds of consistency in translation: lexical and grammatical. Lexical consistency is defined as a repetitive term that keeps the same translation throughout the whole document (Carpuant and Simard, 2012). Inconsistent translation of named entities can significantly impact translation output, although BLEU may not be adversely affected (Agrawal and Singla, 2012; Hermjakob et al., 2008). Therefore, in the design of BLOND\(\text{E}\), we also focus on the reiteration of named entities (e.g., Qiao in Figure 3). On the other hand, typical examples of grammatical consistency are tense and gender consistency. Tense consistency refers to the tense being *compatible* with the context, rather than being exactly the same across the whole document. Tense inconsistency can arise when the source language is an isolating language and does not mark tense explicitly, e.g., in Chinese, and the target language is a synthetic language, e.g., English (teal in Figure 3), where tense is marked explicitly. In the same spirit, the same entity should maintain a consistent grammatical gender.\(^6\)

**Ellipsis.** Ellipsis denotes the omission from a clause of one or more words that are nevertheless inferred in the context of the remaining elements (Yamamoto and Sumita, 1998; Voita et al., 2019). Translation errors arise when there are elliptical constructions in the source language while the target language does not allow for the same types of ellipsis. For example, the ellipsis of subjects or objects is very common in Chinese while it is ungrammatical in English—especially for pronouns. In Figure 3, *she* (Qiao) is omitted in Chinese. However, it is hard to guess the gender of *Qiao* from this stand-alone sentence: the correct pronoun choice can only be inferred from context (there is a *her* in the previous sentence). Another common form of ellipsis is the omission of discourse markers, especially when the source language has more zero connective structures (Po-Ching and Rimmington, 2004) than the target language. In the example, *However* and *So* are ignored in SRC, which misleads the sentence-level system MTA to ignore the discourse relations between sentences.

**Ambiguity.** Translation ambiguity occurs when a word in one language can be translated in more than one way into another language (Tokowicz and Degani, 2010). The cross-language ambiguity comes from several sources of within-language ambiguity including lexical ambiguity, polysemy, and near-synonymy. A unified feature of these is that ambiguous terms satisfy the form of one-to-many mappings. For the example in Figure 2, the word 看(kan) can be translated to *look, see, watch* or *view*. Without access to the context, all the lexical choices are sensible.

\(^5\)The annotation guidelines are described in Appendix B.

\(^6\)It is worth noting that the metric proposed in this study can be applied to a wider range of language pairs by extending the definition of grammatical consistency.

<table>
<thead>
<tr>
<th>Error Type</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO ERROR</td>
<td>451</td>
<td>17.1%</td>
</tr>
<tr>
<td>SENTENCE</td>
<td>1351</td>
<td>51.3%</td>
</tr>
<tr>
<td>DOCUMENT</td>
<td>1893</td>
<td>71.9%</td>
</tr>
<tr>
<td>INCONSISTENCY</td>
<td>1695</td>
<td>64.4%</td>
</tr>
<tr>
<td>NAMED ENTITY</td>
<td>1139</td>
<td>43.3%</td>
</tr>
<tr>
<td>TENSE</td>
<td>1018</td>
<td>38.7%</td>
</tr>
<tr>
<td>ELLIPSIS</td>
<td>534</td>
<td>20.3%</td>
</tr>
<tr>
<td>PRONOUN</td>
<td>456</td>
<td>17.3%</td>
</tr>
<tr>
<td>OTHER</td>
<td>103</td>
<td>4.0%</td>
</tr>
<tr>
<td>AMBIGUITY</td>
<td>193</td>
<td>7.3%</td>
</tr>
</tbody>
</table>

Table 2: The statistics of translation errors in human analysis.
4 BLONDE

The aforementioned document-level phenomena have little impact on the n-gram statistics of translations. However, as is shown in Section 3, they can be key considerations for human readers when evaluating translations at the document level. Standard automatic metrics ignore the importance of contextual coherence of translations, which implies that the document-level nuances are not being properly modeled (Zhou et al., 2008; Xiong and Zhang, 2014). In this section, we describe BLONDE, an automatic metric that explicitly tracks discourse phenomena.

4.1 Document-Level Evaluation

We first give the formulation of measuring discourse phenomena. We define a document $\mathcal{D} = [S_1, \ldots, S_N]$ as a sequence of N sentences. We take a sentence $S$ of length $T$ to be a string of tokens $t_1 \cdots t_T$ where each token $t_i$ is taken from the vocabulary $\mathcal{V}$. Let spans $(S) = \{m_1, m_2, \ldots\}$ be the set of spans in the sentence $S$. Here, a span is a subsequence of the tokens in $S = t_1 \cdots t_T$.

Let us assume that we are interested in $K$ discourse categories. Each of these categories capture a discourse phenomenon of interest. As shown in Section 3, named entity inconsistency, tense inconsistency and pronoun ellipsis make up the majority of discourse errors (67.8%) on the data analyzed. We therefore introduce three types of categories: ENTITY, TENSE and PRONOUN. In addition, we introduce discourse markers DM as a category, which are the essential contextual links between the various discourse segments (See Figure 3).

For a certain discourse category of interest, $k$, we assume that there are $L_k$ features. In our case, the features of ENTITY $\mathcal{E}$ are a list of named entities in $\mathcal{D}$; the features of TENSE are $\mathcal{V} = \{\text{MD}, \text{VBD}, \text{VBN}, \text{VBP}, \text{VBZ}, \text{VB}\}$, the features of PRONOUN are $\mathcal{P} = \{\text{masculine}, \text{feminine}, \text{neuter}, \text{epicene}\}$, and the features of DM are $\mathcal{M} = \{\text{contingency}, \text{temporal}, \text{expansion}, \text{comparison}\}$. Note that different categories can have different numbers of features and the number of features can be dynamic: $\mathcal{E}$ depends on $\mathcal{D}$ while $\mathcal{V}$ and $\mathcal{P}$ are fixed. The intuition behind this is that we want to encourage the system output to keep consistent tense and pronouns as well as the consistent translation for a specific named entity.

Let us now define $C_{k,t}(S) \subseteq \text{spans}(S)$ as the set of spans in $S$ that share the $t^{th}$ feature in the $k^{th}$ discourse category. To give a concrete example, let us assume that TENSE is the $k^{th}$ category and VBD is the $t^{th}$ feature in this category. The corresponding $C_{k,t}(S)$ is the set of the spans (in this case, unigrams) tagged with VBD in the sentence.

 maid: Modal; VBD: Verb (past tense verb); VBN: Verb (past participle); VBP: Verb (non-3rd person singular present); VBZ: Verb (3rd person singular present); VB: Verb (gerund or present participle); VB: Verb (base form).

7 masculine: he, him, his, himself; feminine: she, her, hers, herself, neuter: it, its, itself; epicene: they, them, their, theirs, themselves.

8 A detailed explanation is provided in Table 6.

<table>
<thead>
<tr>
<th>SRC</th>
<th>Qiao 看着(look)相片回憶(recall)起了二十年前。</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>Qiao looked at the photo and recalled twenty years ago.</td>
</tr>
<tr>
<td>b)</td>
<td>This bearded man was her newlywed husband.</td>
</tr>
<tr>
<td>c)</td>
<td>[yet] this was the first time they were meeting with each other.</td>
</tr>
<tr>
<td>d)</td>
<td>[So] Qiao’s heart jolted as soon as [she] saw him, and [she] quickly stood up.</td>
</tr>
<tr>
<td>REF</td>
<td>Qiao looked at the photo and recalled twenty years ago.</td>
</tr>
<tr>
<td>a)</td>
<td>Qiao looked at the photo and recalled twenty years ago.</td>
</tr>
<tr>
<td>b)</td>
<td>This bearded man is her newlywed husband.</td>
</tr>
<tr>
<td>c)</td>
<td>This is the first time they met with each other.</td>
</tr>
<tr>
<td>d)</td>
<td>[So] Joe’s heart jolted as soon as [he] saw him, and [he] quickly stood up.</td>
</tr>
<tr>
<td>MTA</td>
<td>Qiao looked at the photo and recalled the past twenty years ago.</td>
</tr>
<tr>
<td>a)</td>
<td>Qiao looked at the photo and recalled the past twenty years ago.</td>
</tr>
<tr>
<td>b)</td>
<td>This man with the beard was her newlywed husband.</td>
</tr>
<tr>
<td>c)</td>
<td>[However], that was the first time they met.</td>
</tr>
<tr>
<td>d)</td>
<td>[So] as soon as Qiao saw him, [her] heart became squeaky, and [she] swiftly stood up.</td>
</tr>
<tr>
<td>MTB</td>
<td>Qiao looked at the photo and recalled the past twenty years ago.</td>
</tr>
<tr>
<td>a)</td>
<td>Qiao looked at the photo and recalled the past twenty years ago.</td>
</tr>
<tr>
<td>b)</td>
<td>This man with the beard was her newlywed husband.</td>
</tr>
<tr>
<td>c)</td>
<td>[However], that was the first time they met.</td>
</tr>
<tr>
<td>d)</td>
<td>[So] as soon as Qiao saw him, [her] heart became squeaky, and [she] swiftly stood up.</td>
</tr>
</tbody>
</table>

Figure 3: An example containing inconsistency and ellipsis in BWB. For inconsistency, the same entities are marked in the same color (Qiao and Husband), and verbs are marked in teal. For ellipsis, omissions are marked with [ ]. DM stands for discourse markers ( ). The translation mistakes are underlined. MTB is intuitively a better system than MTA to human readers.
tence $S$. In Figure 3, all the spans are colored. We then let $C_k(S)$ refer to a vector of size $L_k$ where each element of that vector is the set $C_{k,l}(S)$. The sets of spans for ENTITY, TENSE, PRONOUN and DM can be produced by an NER model, a POS tagger, a rule-based string match and a discourse marker, respectively. We also define a weight vector $w_k = [w_{k,l} : l \in \{1, \ldots, L_k\}]$ for each discourse category $k$, where each entry $w_{k,l}$ corresponds to the weight given to a feature.

We then define the discourse representation of sentence $S$ as the concatenation of all categories:

$$C(S) = \{C_k(S) : k \in \{1, \ldots, K\}\}$$  \hspace{1cm} (1)

**Similarity.** Let $\text{sim} : C(S^s) \times C(S^r) \to \mathbb{R}^K$ represent a similarity vector which measures category-wise similarity between the discourse representations of two sentences $S^r$ and $S^s$. Each entry of the vector $\text{sim}$ takes non-negative values: The entry being zero if $S^s$ and $S^r$ have no shared spans with the same discourse category.

The similarity vector $\text{sim}$ defined here can be implemented in several ways. A possible implementation of $\text{sim}$ can be achieved by counting the number of functionally similar spans for each feature and then taking a weighted sum over all features:

$$\text{sim}(S^s, S^r) = \sum_{k=1}^{K} w_k \cup \min\{\text{count}(C_k(S^s)), \text{count}(C_k(S^r))\}$$  \hspace{1cm} (2)

where each entry $\text{sim}_k$ is defined as follows:

$$\text{sim}_k(S^s, S^r) = \sum_{l=1}^{L_k} w_{k,l} \cup \min\{\text{count}(C_{k,l}(S^s)), \text{count}(C_{k,l}(S^r))\}$$  \hspace{1cm} (3)

where

$$\text{count}(C_k(\cdot)) = |\{C_{k,l}(\cdot) : l \in \{1, \ldots, L_k\}\}|$$  \hspace{1cm} (4)

denotes the cardinality of $C_{k,l}$ applied entry-wise and $\cup$ denotes the minimum function applied element-wise. Intuitively, $\text{sim}_k$ measures the number of functionally similar spans shared by $S^s$ and $S^r$. Assume that TENSE is the $k^{th}$ category.\footnote{In Figure 3, $\text{sim}(\cdot)_{\text{MTA}}, U_{\text{MTA}} = 0$ since MTA mistranslated the verbs as present tense due to the exclusion of context. The total similarity $\text{sim}(D^s_{\text{MTA}}, D^r)$ is the number of functionally similar spans across all features: $(1, 2, 4, 0)$ for $(\mathcal{E}, \mathcal{V}, \mathcal{P}, \mathcal{M})$. Here, we assume all category weights are 1.}

It is worth noting that there are many other reasonable ways to operationalize $\text{sim}$. For ENTITY, partial credit could be assigned to two named entities if they have overlapping tokens; for TENSE and PRONOUN, partial credit could be assigned to two similar categories, e.g., VBP and VB; for DM, partial credit could be assigned according to the sense hierarchy and the confidences in the detected discourse markers. We leave the expansion of the $\text{sim}$ definition to future work.

**A Document-level Similarity Measure.** Now we turn from measuring the similarity at the sentence level to the document level. We first lift $\text{sim}(\cdot, \cdot)$ to measure the similarity between two documents:

$$\text{sim}(D^s, D^r) = \sum_{S^r \in D^r} \text{sim}(S^s, S^r)$$  \hspace{1cm} (5)

where the sum is applied element-wise.

We then define $\text{sim}(\cdot, \cdot)$ for a system document $D^s$ and a set of reference documents $D^r = \{D^{r_1}, D^{r_2}, \ldots\}$ by aggregating the $\text{sim}$ of all sentences in $D^s$ and $D^r$:

$$\text{sim}(D^s, D^r) = \sum_{S^s \in D^s, S^r \in D^r} \text{sim}(S^s, S^r)$$  \hspace{1cm} (6)

Here, $\oplus$ is a generic aggregator over multiple references, e.g., $\oplus = \max$, if we take the reference which has the maximum similarity with the system output; or $\oplus = \sum$ if we sum up the similarity scores of all references. Again, $\oplus$ is applied element-wise.\footnote{In Figure 3, since we only have one reference, $\text{sim}(D^s_{\text{MTA}}, D^r) = \text{sim}(D^s_{\text{MTA}}, D^{r_1}) = (1, 2, 4, 0)$.}

We also reuse the notation $\text{sim}(\cdot, \cdot)$ for two sets of documents $D^s$ and $D^r$:

$$\text{sim}(D^s, D^r) = \bigoplus_{D^s \in D^s} \text{sim}(D^s, D^r)$$  \hspace{1cm} (7)

Note that the similarity vector can also be computed for the same (set of) documents. For example, if $\text{sim}$ is implemented as counting the number of functionally similar spans for each feature, then, $\text{sim}(D^s, D^s)$ and $\text{sim}(\mathcal{D}^s, \mathcal{D}^r)$ denote the total number of spans of each category in the system output and the reference, respectively.\footnote{In Figure 3, $\text{sim}(\mathcal{D}^s_{\text{MTA}}, \mathcal{D}^s_{\text{MTA}}) = (1, 7, 6, 0)$ and $\text{sim}(\mathcal{D}^s, \mathcal{D}^r) = \text{sim}(\mathcal{D}^{r_1}, \mathcal{D}^{r_1}) = (2, 7, 5, 2)$.}

**Scoring.** We are now ready to define the “goodness” of a system output with respect to our discourse phenomenon of interest. We compute the
precision, recall and F1 for all $K$ discourse categories defined as follows:

$$p(\mathcal{D}^s, \mathcal{D}^r) = \frac{\text{sim}(\mathcal{D}^s, \mathcal{D}^r)}{\text{sim}(\mathcal{D}^s, \mathcal{D}^r)}$$

(8)

$$r(\mathcal{D}^s, \mathcal{D}^r) = \frac{\text{sim}(\mathcal{D}^s, \mathcal{D}^r)}{\text{sim}(\mathcal{D}^r, \mathcal{D}^s)}$$

(9)

$$F(\mathcal{D}^s, \mathcal{D}^r) = 2 \cdot \frac{P \circ R}{P + R}.$$  (10)

Here, $p(\mathcal{D}^s, \mathcal{D}^r)$, $r(\mathcal{D}^s, \mathcal{D}^r)$ and $F(\mathcal{D}^s, \mathcal{D}^r)$ are all $K$-dimensional vectors; where, the $k^{th}$ element of these vectors represents the precision, recall and F1-score for the $k^{th}$ category. Thus, the addition, multiplication and division operations above are also defined element-wise.\(^{13}\)

**BLOND-d.** Further, we combine the scores of all categories into an overall score with a simple weighted average, named BLOND-d. By computing BLOND-d, one can distill the document-level translation quality from translation quality at the sentence level. More formally, we have

$$\text{BLOND-d.P}(\mathcal{D}^s, \mathcal{D}^r) = \left( \prod_{k=1}^{K} (p_k(\mathcal{D}^s, \mathcal{D}^r))^{a_k} \right)^{1/\sum_{k=1}^{K} a_k}$$

(11)

$$\text{BLOND-d.R}(\mathcal{D}^s, \mathcal{D}^r) = \left( \prod_{k=1}^{K} (r_k(\mathcal{D}^s, \mathcal{D}^r))^{a_k} \right)^{1/\sum_{k=1}^{K} a_k}$$

(12)

where $a_k$ denotes the importance weight of the $k^{th}$ category, and $p_k$ and $r_k$ denote the $k^{th}$ entry of $p$ and $r$, respectively.\(^{14}\) Therefore, BLOND-d.F1 is defined as follows:

$$\text{BLOND-d.F1}(\mathcal{D}^s, \mathcal{D}^r) = 2 \cdot \frac{\text{BLOND-d.P} \cdot \text{BLOND-d.R}}{\text{BLOND-d.P} + \text{BLOND-d.R}}$$

(13)

Whenever not otherwise specified, we simply use BLOND-d to refer to BLOND-d.F1.\(^{15}\)

### 4.2 BLONDE: Combining BLOND-d with $n$-grams

However, focusing on discourse phenomena solely is not enough to provide comprehensive MT evaluation that correlates strongly with human judgments. Consider the following example:

(1) **REF** Qiao lifted her heavy eyelids.

**MT** Qiao scrunched her brows together.

The output of MT is far from “good” in terms of adequacy, whereas BLOND-d(MT) = 1, since MT translates both named entities and tenses correctly. Thus, in order to account for sentence-level adequacy of our final metric BLONDE, we augment the set of categories and features to include each $n$-gram (for a value of $n$) as a category and each span of $n$-tokens as a feature for the $n$-gram category. Formally, we have

$$C^n(S) = \{C_k(S) : k \in \{1, \ldots, K + N\}\}$$

(14)

where we define

$$C_{K+n} = \{n\text{-gram} : n \in \{1, \ldots, N\}\}$$

(15)

The calculation of BLONDE.P, BLONDE.R and BLONDE.F1 is then done exactly in the same manner as BLOND-d. Whenever not specified, we simply use BLONDE to refer to BLONDE.F1.

\(^{15}\)E.g., the BLOND-d scores of mta in Figure 3 are:

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>BLONDE</th>
<th>BLOND-d</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>41.5</td>
<td>10.5</td>
<td>17.4</td>
</tr>
<tr>
<td>R</td>
<td>35.9</td>
<td>60.6</td>
<td>58.9</td>
</tr>
<tr>
<td>F1</td>
<td></td>
<td></td>
<td>97.7</td>
</tr>
</tbody>
</table>

Table 3: The BLEU and BLONDE scores of the two system outputs in Figure 3. P, R and F1 represent precision, recall and F1, respectively.

\(^{16}\)In Figure 3, recall that $\text{sim}(\mathcal{D}^{mta}, \mathcal{D}^r) = (1, 2, 4, 0)$, $\text{sim}(\mathcal{D}^{mta}, \mathcal{D}^{mta}) = (1, 7, 6, 0)$ and $\text{sim}(\mathcal{D}^r, \mathcal{D}^r) = (2, 7, 5, 2)$. Thus, we have $p(\mathcal{D}^{mta}, \mathcal{D}^r) = \left(\frac{1}{2}, \frac{2}{7}, \frac{1}{5}, \frac{1}{0}\right)$ and $p(\mathcal{D}^{mta}, \mathcal{D}^{mta}) = (1, \frac{2}{7}, \frac{2}{5}, NA)$ where NA denotes a missing value. Furthermore, we have $r(\mathcal{D}^{mta}, \mathcal{D}^r) = \left(\frac{1}{2}, \frac{2}{7}, \frac{1}{5}, \frac{1}{2}\right)$ and $r(\mathcal{D}^{mta}, \mathcal{D}^{mta}) = (1, \frac{2}{7}, \frac{2}{5}, \frac{1}{2})$ where $\delta$ denotes a small value ($0.0001$) for smoothing. Finally, we have $F(\mathcal{D}^{mta}, \mathcal{D}^r) = \left(\frac{1}{2}, \frac{2}{7}, \frac{1}{5}, \frac{1}{0}\right)$.

\(^{14}\)BLOND-d adopts uniform weights.
BLONDE covers both discourse coherence features and sentence-level adequacy, thus providing a comprehensive measurement of translation quality. Table 3 compares BLONDE with BLEU using the two MT outputs found in Figure 3. It is striking that BLEU rates MTA higher than MTB given that MTB is clearly better than MTA to human readers. In sharp contrast, their BLONDE scores reflect the correct ranking in translation quality.

4.3 BLOND+: Combining BLONDE with Human Annotations
BLONDE is easy to generalize—for instance, it would be easy to incorporate human annotations, e.g., one could annotate spans related to discourse errors and treat them as categories. The automatically inferred categories and human annotated categories are then combined by adopting the same weighted averaging approach, which we call BLOND+. We hired the same translators who analyzed discourse errors in Section 3 to annotate ambiguous and omitted word/phrases on the test set of BWB.\(^{17}\)

5 Experiments
In this section, we examine the effectiveness of BLONDE at the document-level MT evaluation through experiments. We answer the following question: Do differences in BLONDE reliably reflect differences in the document-level translation quality of different MT systems? To answer this question, we run several MT baselines and compare their BLONDE scores to eleven other metrics:

Standard Sentence-level Metrics. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015).

Document-level Metrics. LC and RC (Wong and Kit, 2012)—these are ratios between the number of lexical cohesion devices (repetition and collocation) and repeated content words over the total number of content words in a target document, which are direct measurements of lexical cohesion.

Embedding-based Metrics. We consider four embedding-based metrics in this work: SkipThought cosine similarity (SKIP; Kiros et al., 2015), embedding average cosine similarity (AVER; Sharma et al., 2017), Vector extrema cosine similarity (VECTOR; Forgues et al., 2014), Greedy Match (GREEDY; Rus and Lintean, 2012).

5.1 MT Systems
We test BLONDE on the following system outputs: an SMT system (Chiang, 2007), three well-known commercial NMT systems (OMT-A, OMT-B, OMT-C), a sentence-level transformer-based system (MT-S) and a document-level system (MT-D) trained on BWB. MT-D (Zhang et al., 2018) trains sentence-level model parameters and then estimates document-level model parameters while keeping the sentence-level Transformer model parameters fixed. We adopt Transformer Big (Vaswani et al., 2017) for both MT-S and MT-D. The final “system” is a human post-editing (PE) on OMT-C provided by professional translators, so it is supposed to be the strongest baseline.\(^{18}\)

5.2 The BLONDE Evaluation
Firstly, we leverage the test set of BWB and evaluate the above-mentioned systems by BLONDE and other metrics. Figure 4 presents the means of all metrics along with the 95% confidence interval estimated from bootstrap resampling. We observe that the BLONDE scores demonstrate an exponentially increasing trend from sentence-level towards document-level and human post-editing, while the trends of standard metrics are mostly linear. Specifically, the difference between the BLONDE scores of MT-S and MT-D (denoted as \(\Delta(MT-S, MT-D)\)) is significantly higher than the difference between the \(\Delta(MT-S, MT-D)\) in their BLEU scores. An even larger \(\Delta\) between MT-D and PE in their BLONDE scores is observed, indicating MT-D is still far away from achieving human parity. Note that the trend of BLOND-D scores is even more exponential, which indicates that BLOND-D indeed distills document-level translation quality.

The \(t\)-statistics of the paired sample \(t\)-tests of individual documents are given in Table 4. Unlike BLEU, METEOR and other metrics, which either fails to distinguish human and machine translation or has lower discriminative power compared to distinguishing different machine translations, the BLONDE family maintain similar discriminative

\(^{17}\)We also make this annotated test set publicly available as a testbed for evaluating the ability of MT systems to disambiguate word senses and to predict coherent pronouns or discourse markers in the case of omission.

\(^{18}\)We trained models by fairseq (Ott et al., 2019). Model parameters and the post-editing details are in Appendix F.2 and C, respectively.
Figure 4: The mean scores of different system outputs given by different metrics on the BWB test set. Shaded region represents 95% confidence interval.

Table 4: The paired t-statistics of different MT systems. The cells with p-value > .05 are marked in gray. While BLEU distinguishes SMT and the sentence-level MT-S significantly, it fails to possess the same discriminative power towards document-level and human translations. BLONDe maintains similar discriminative power across the three t-tests.

![Figure 5: Absolute Pearson correlation pairs of automatic metrics. Computed over the scores of individual documents in BWB test set.](image)

In addition, the t-statistics of BLOND-D categories provide rich diagnostic information. As can be seen, although transformer-based NMT models have substantially higher BLEU scores than SMT systems, MT-S is not statistically superior to SMT in terms of named entity translation. However, human post-editing scores significantly better on entity translation—meaning that named entity translation accounts for a substantial part of quality differences between machine and human. In terms of TENSE and and DM translation, MT-D does not significantly out-perform MT-S, which could be taken into consideration in future document-level MT model designs.

We also show the pairwise Pearson correlations between different metrics in Figure 5. It illustrates the homogeneity/heterogeneity of different metrics. We report the absolute value of the correlation for TER. We see that while sentence-level metrics (BLEU, METEOR and ROUGE-L) have strong correlations with each other, BLONDe correlates less well with those metrics.
We then evaluate BLONDE along with other metrics in terms of their Pearson correlation with human assessment. Our human assessment is provided by four professional Chinese to English translators and four native English revisers. Two experimental units (SENTENCE vs DOCUMENT) are assessed independently in terms of FLUENCY and ADEQUACY, respectively. In the SENTENCE-level evaluation, we show the raters isolated sentences, while in the DOCUMENT-level evaluation, we show them entire documents and we only ask raters to evaluate the overall quality of sequential blocks of sentences (5 sentences per block) as used in the Relative Ranking (RR) evaluation (Bbojar et al., 2016). We use the Williams significance test (Williams, 1959; Graham and Baldwin, 2014) to identify correlation differences that are statistically significant. The detailed protocol is presented in Appendix D.

The results are shown in Table 5. BLONDE obtains the highest correlation with human assessment at both the sentence level and the document level. However, BLONDE correlates remarkably better with human assessment when context is taken into account, and it only significantly outperforms all other metrics at document level.

It is worth noting that BLONDE also correlates well with FLUENCY assessment, even though it is, in essence, still a reference-based metric. One possible explanation for this unexpected positive result is that it tracks span categories that directly relate to cohesion and coherence. Another important observation is that the recall-based BLONDE variants generally correlate better with human assessment, yet appears to be less selective compared to the precision-based variants (see MT-D→PE in Table 4). This provides support for adopting the F1 in order to get the best of both worlds.

### 6 Related Work

There have been a few studies on automatic evaluation metrics for specific discourse phenomena.

#### Pronoun Translation

Hardmeier and Federico (2010) measured the precision and recall of pronouns directly and Miculicich Werlen and Popescu-Belis (2017) proposed to estimate the accuracy of pronoun translation (APT) by aligning source and target texts. However, as shown in Guillou and Hardmeier (2018), APT does not take the antecedents of an anaphoric pronoun into account. They cannot handle the mismatches in the numbers of pronouns either. Jwalapuram et al. (2019) also proposed a specialized measure for pronoun evaluation which involves training. In comparison, BLONDE does not rely on any alignment or training.

#### Lexical Cohesion

Wong and Kit (2012) proposed LC and RC. Gong et al. (2015) described a cohesion function to measure text cohesion via lexical chain and a gist consistency score based on topic model. However, they fail to distinguish accidental repetition from document-level cohesion.

#### Discourse Relations

Hajlaoui and Popescu-Belis (2013) proposed to assess the accuracy of connective translation (ACT). However, such an assessment requires a bilingual dictionary of all possible DM translations, whereas BLONDE only requires a list of monolingual DMs. Guzmán et al. (2014) and Joty et al. (2014) compute a metric based on the similarity between the discourse trees of reference and system output. Those discourse-representation-based metrics are indirect, and rely on discourse parsing tools, which are much more inaccurate than syntactic and semantic parsing tools used in BLONDE. Unlike previously proposed metrics, BLONDE does not only focus on one specific discourse phenomenon, and thus has significantly

### 5.3 Human Evaluation

We then evaluate BLONDE along with other metrics in terms of their Pearson correlation with human assessment. Our human assessment is provided by four professional Chinese to English translators and four native English revisers. Two experimental units (SENTENCE vs DOCUMENT) are assessed independently in terms of FLUENCY and ADEQUACY, respectively. In the SENTENCE-level evaluation, we show the raters isolated sentences, while in the DOCUMENT-level evaluation, we show them entire documents and we only ask raters to evaluate the overall quality of sequential blocks of sentences (5 sentences per block) as used in the Relative Ranking (RR) evaluation (Bojar et al., 2016). We use the Williams significance test (Williams, 1959; Graham and Baldwin, 2014) to identify correlation differences that are statistically significant. The detailed protocol is presented in Appendix D.

The results are shown in Table 5. BLONDE obtains the highest correlation with human assessment at both the sentence level and the document level. However, BLONDE correlates remarkably better with human assessment when context is taken into account, and it only significantly outperforms all other metrics at document level.

<table>
<thead>
<tr>
<th>Unit</th>
<th>SENTENCE</th>
<th></th>
<th>DOCUMENT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADE</td>
<td>FLU</td>
<td>ADE</td>
<td>FLU</td>
</tr>
<tr>
<td>BLONDE.R</td>
<td>.363</td>
<td>.327</td>
<td>.436†</td>
<td>.371†</td>
</tr>
<tr>
<td>BLONDE.P</td>
<td>.331</td>
<td>.296</td>
<td>.383†</td>
<td>.344†</td>
</tr>
<tr>
<td>BLONDE.F1</td>
<td>.35</td>
<td>.314</td>
<td>.417†</td>
<td>.358†</td>
</tr>
<tr>
<td>BLOND+.R</td>
<td>.364</td>
<td>.329†</td>
<td>.44†</td>
<td>.373†</td>
</tr>
<tr>
<td>BLOND+.P</td>
<td>.334</td>
<td>.3</td>
<td>.39†</td>
<td>.349†</td>
</tr>
<tr>
<td>BLOND+.F1</td>
<td>.351</td>
<td>.318</td>
<td>.422†</td>
<td>.362†</td>
</tr>
<tr>
<td>BLEU</td>
<td>.325</td>
<td>.308</td>
<td>.343</td>
<td>.266</td>
</tr>
<tr>
<td>METEOR</td>
<td>.338</td>
<td>.31</td>
<td>.339</td>
<td>.278</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>.275</td>
<td>.262</td>
<td>.29</td>
<td>.211</td>
</tr>
<tr>
<td>TER</td>
<td>.063</td>
<td>.027</td>
<td>.044</td>
<td>.092</td>
</tr>
<tr>
<td>CIDER</td>
<td>.139</td>
<td>.116</td>
<td>.114</td>
<td>.087</td>
</tr>
<tr>
<td>SKIP</td>
<td>.213</td>
<td>.174</td>
<td>.163</td>
<td>.171</td>
</tr>
<tr>
<td>AVER</td>
<td>.163</td>
<td>.163</td>
<td>.16</td>
<td>.111</td>
</tr>
<tr>
<td>VECTOR</td>
<td>.25</td>
<td>.243</td>
<td>.248</td>
<td>.218</td>
</tr>
<tr>
<td>GREEDY</td>
<td>.323</td>
<td>.3</td>
<td>.307</td>
<td>.265</td>
</tr>
<tr>
<td>LC</td>
<td>.086</td>
<td>.061</td>
<td>.153</td>
<td>.116</td>
</tr>
<tr>
<td>RC</td>
<td>.096</td>
<td>.07</td>
<td>.109</td>
<td>.13</td>
</tr>
</tbody>
</table>

Table 5: Absolute Pearson correlation with human judgments on BBW. The highest correlations are in bold. Correlation of metrics not significantly outperformed by any other metrics are highlighted with †. The BLONDE family are not tested against each other.
higher Pearson correlation coefficients with human assessments.

7 Conclusion

In this paper, we build a large-scale parallel dataset for document-level translation, BWB. We analyze it for common document-level translation errors in practice and propose BLONDE, an interpretable automatic metric for document-level MT evaluation. We further improve BLONDE by diagnosing and distilling discourse-related errors in MT outputs and human-annotations to obtain two improved metrics BLOND-D and BLOND+. These metrics were shown to have better selectivity than various sentence-level metrics and correlate better with human judgments.

Ethical Considerations

The annotators were paid a fair wage and the annotation process did not solicit any sensitive information from the annotators. Finally, while our approach is not tuned for any specific real-world application, the approach could be used in sensitive contexts such as legal or health-care settings, and any work must use our approach undertake extensive quality-assurance and robustness testing before using it in their setting.

Replicability. As part of our contributions, we will release the annotated BWB test set, and release the crawling script of the training set under Fair Use rules. The BLONDE package is also publicly available at https://github.com/EleanorJiang/BlonDe.

Acknowledgements

Mrinmaya Sachan acknowledges support from an ETH Zürich Research grant (ETH-19 21-1) and a grant from the Swiss National Science Foundation (project #201009) for partially supporting this work.

References


A Dataset Creation

The Background of Translators. The original Chinese books are translated by professional native English speakers, and are corrected by editors.

Data Collection. This process is implemented by a python web crawler, and certain data cleaning is also done in the process. We crawl the books chapter by chapter, and convert the text to UTF-8. After deduplication, we remove the chapters with less than 5 sentences. We further remove the titles of each chapter, because most of them are neither translated properly nor in the document-level.

Alignment and Quality Control. After collecting the web books, we align the bilingual books chapter by chapter according to the indices, while removing those chapters without parallel data. Then, we use Bluealign, which is an MT-based sentence alignment tool, to align the chapters into parallel sentences, while retaining the document-level information. We further deduplicate the parallel corpus and filter the pairs with a sequence ratio of 3.0. The scale of the final corpus is 384 books with 9,581,816 sentence pairs (a total of 460 million words). To estimate the accuracy of this process, we hired 4 bilingual graduate students to manually evaluate 163 randomly selected documents from the resulting BWB parallel corpus. These students are native Chinese speakers who are proficient in English. More specifically, they were asked to distinguish whether a document is well aligned at the sentence level by counting the number of misalignment. For example, if Line 39 in English actually corresponds to Line 39 and Line 40 in Chinese, but the tool made a mistake that it combines the two sentences, it is identified as a misalignment. We observe an alignment accuracy rate of 93.1%.

We further asked the same batch of annotators to correct such misalignments in both the development and the test set. The annotation result shows that 7.3% lines are corrected.

B Error Analysis and BLOND+ Annotation

Error analysis and BLOND+ annotation are conducted together. This task is conducted by eight professional Chinese-English translators who are native in Chinese and fluent in English.

The guideline is as follows:

- First, identify cases which have translation errors. The annotators are instructed to mark examples as “translations with no error” only if it satisfies the criteria of both adequacy and fluency as well as satisfies the criterion that it is coherent in the context.
- Second, identify whether the translation contains document-level error or sentence-level error (or both). The annotators are instructed to mark examples as “cases with sentence-level errors” when they are not adequate or fluent as stand-alone sentences; while “document-level errors” mean those errors that cause the example violating the global criterion of coherence.
- Third, categorize the examples with document-level errors into mainly into 3 categories, namely inconsistency, ellipsis, and ambiguity. According to this observation, we instruct annotators to mark document-level errors as inconsistency, ellipsis, and ambiguity, or other document-level error during the annotation process for the entire test set.

We first conduct a test annotation and observe that the annotators categorize document-level errors into mainly into 3 categories, namely inconsistency, ellipsis, and ambiguity. According to this observation, we instruct annotators to mark document-level errors as inconsistency, ellipsis, and ambiguity, or other document-level error during the annotation process for the entire test set.

In the formal annotation process, we also added the requirement to annotate BLOND+ spans. The detailed requirement is as follows:

- Third, categorize the examples with document-level into 4 categories: inconsistency, ellipsis, and ambiguity, or other document-level error which cannot be categorized.
- Fourth, if the example is categorized as ambiguity, mark the specific word/phrase in the reference (English) that cause ambiguity and give the correct word/phrase.
- Fifth, if the example is categorized as ellipsis and it is not related to pronouns or discourse markers, mark the omitted word/phrase in the reference (English).

C Human Post-Editing

This task is conducted by the same eight professional Chinese-English translators who carry out the annotation in Appendix B. We asked them to follow guidelines for achieving “good enough”
quality at the sentence-level (comprehensible, accurate but as not being stylistically compelling) but especially pay attention to document-level errors and correct them.

### D The Human Evaluation Protocol

The human evaluation is conducted on outputs of four systems (OMT-B, MT-S, CTX, PE) and human translation. We follow the protocol proposed by (Läubli et al., 2018, 2020). We conduct the evaluation experiment with a $2 \times 2$ mixed factorial design, carrying both DOCUMENT-level and SENTENCE-level evaluation in terms of ADEQUACY and FLUENCY. In the SENTENCE-level evaluation, we show raters isolated sentences in random order; while in the DOCUMENT-level evaluation, entire documents are presented and we only ask raters to evaluate a sequence of 5 sequential sentences at a time in order.

To avoid reference bias, the ADEQUACY evaluation is only based on source texts, while no source texts nor references are presented in the FLUENCY evaluation. We adopt Relative Ranking (RR): raters are presented with outputs from the aforementioned five systems, which they are asked to evaluate relative to each other, e.g., to determine system A is better than system B (with ties allowed).

We use source sentences and documents from the BWB test set, but blind their origins by randomizing both the order in which the system outputs are presented. Note that in the DOCUMENT-level evaluation, the same ordering of systems is used within a document. The order of experimental items is also randomised. Sentences are randomly drawn from these documents, regardless of their position.

We also use spam items for quality control ( Kittur et al., 2008): In a small fraction of items, we render one of the five options nonsensical by randomly shuffling the order of all translated words, except for 10% at the beginning and end. 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. At the DOCUMENT-level, we render one of the five options nonsensical by randomly shuffling the order of all translated sentences, except for the first and the last sentence.

We recruit four professional Chinese to English translators and four native English revisers for the adequacy and fluency conditions respectively. Note that the eight translators are different from those professional translators who carry out the human translation PE. We deliberately invite another group of specialists for human evaluation to avoid making unreasonable judgments biased towards PE. In each condition, each rater evaluates 162 documents (plus 18 spam items) and 162 sentences (plus 18 spam

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>#Docs</th>
<th>#Sents</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT (Barrault et al., 2019)</td>
<td>News</td>
<td>68.4k</td>
<td>3.63M</td>
</tr>
<tr>
<td>OpenSubtitles (Lison et al., 2018)</td>
<td>Subtitles</td>
<td>29.1k</td>
<td>31.2k</td>
</tr>
<tr>
<td>TED (Ansari et al., 2020)</td>
<td>Talks</td>
<td>1K</td>
<td>219M</td>
</tr>
<tr>
<td>BWB</td>
<td>Books</td>
<td>196k</td>
<td>9M</td>
</tr>
</tbody>
</table>

Table 6: Explanations of the discourse marker types (discourse relations) in DM.

Table 7: Comparison of different document-level datasets.
<table>
<thead>
<tr>
<th>RATER1-RATER2</th>
<th>SENTENCE</th>
<th>DOCUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>.171</td>
<td>.169</td>
<td></td>
</tr>
<tr>
<td>RATER3-RATER4</td>
<td>.294</td>
<td>.346</td>
</tr>
<tr>
<td>RATER5-RATER6</td>
<td>.323</td>
<td>.402</td>
</tr>
<tr>
<td>RATER7-RATER8</td>
<td>.378</td>
<td>.342</td>
</tr>
</tbody>
</table>

Table 8: Inter-rater agreements measure by Cohen’s κ, where RATER1-4 are professional translators whose native language is Chinese, RATER5-8 are native English revisers.

items). We use two non-overlapping sets of documents and two non-overlapping sets of sentences, and each is evaluated by two raters. Specifically, we refer the first half of the test set as PART1 and the second half as PART2. Note that PART1 and PART2 are chosen from different books. Each rater evaluates both sentences and documents, but never the same text in both conditions so as to avoid repetition priming (Gonzalez et al., 2011): RATER1 and RATER2 conduct the DOCUMENT-level ADEQUACY evaluation on 180 documents sampled from PART1 and the SENTENCE-level ADEQUACY evaluation for PART2; RATER3 and RATER4 conduct the SENTENCE-level FLUENCY evaluation on 180 documents sampled from PART1 and the DOCUMENT-level FLUENCY evaluation for PART2; RATER5 and RATER6 conduct the DOCUMENT-level FLUENCY evaluation on 180 documents sampled from PART1 and the SENTENCE-level FLUENCY evaluation for PART2; RATER7 and RATER8 conduct the SENTENCE-level FLUENCY evaluation on 180 documents sampled from PART1 and the DOCUMENT-level FLUENCY evaluation for PART2.

E Statistical Analysis of Human Evaluation

We calculate Cohen’s kappa coefficient:

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

where \( P(A) \) is the proportion of times that two raters agree, and \( P(E) \) is the likelihood of agreement by chance. We report pairwise inter-rater agreement in Table 8.

F Experiment Settings

F.1 BLOONDE

We use the named entity recognition module and the POS tagger of spaCy (Honnibal and Montani, 2017) to implement the categorizing function cat for ENTITY and TENSE, respectively. We use the script provided by Sileo et al. (2019) as the discourse marker minor.

F.2 Model Hyperparameters

We follow the setup of Transformer big model for BWB experiments. More precisely, the parameters 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 \( d_{model} = 1024 \), and the inner-layer of a feed-forward networks has dimensionality \( d_{ff} = 4096 \). 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.