BLONDE: An Automatic Evaluation Metric for Document-level Machine Translation

Anonymous NAACL submission

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 BLONDE[1] to widen the scope of automatic MT evaluation from sentence to document level. BLONDE 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 dataset $\mathcal{BWB}$. The experimental results show that BLONDE possesses better selectivity and interpretability at the document-level, and is more sensitive to document-level nuances. In a large-scale human study, BLONDE also achieves significantly higher Pearson's r correlation with human judgments compared to previous metrics.

1 Introduction

Over the past few 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) suggest that NMT has achieved human parity at the sentence level, the reliability of these human-parity claims was quickly contested by Läubli 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 growing attention in the MT community. However, despite various modeling advances, we still lack an efficient and effective evaluation

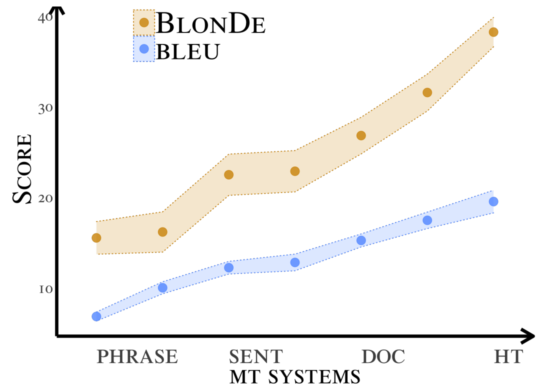

Figure 1: BLONDE is a lot more selective than BLEU for document-level MT, and also shows a larger quality difference between human and machine translations.

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 performs 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 the construction of such test suites can be very labor-intensive. We still lack a 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 ($\mathcal{BWB}$) from heterogeneous data sources, and quantify document-level translation mistakes by performing a large human study. We found that on this dataset, inconsistency[2], ellipsis and ambiguity were the most noticeable phenomena critical for document-level MT, together

[1] BLONDE: **Bil**ingual Evaluati**on** of **D**ocum**e**nt Translation. The package and data will be publically available.

[2] By inconsistency we mean the mistakes related to coreference and lexical cohesion (Carpuat, 2009; Guillou, 2013).

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 then computes recall, precision and F-measure, 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) are 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. Human evaluation also reveals significantly higher Pearson's r correlation coefficients between BLONDE and human assessments.

2 $\mathcal{BWB}$: Bilingual Web Book Dataset

To design a metric that is more sensitive to document-level phenomena, we first curate a document-level Chinese–English parallel corpus, called $\mathcal{BWB}$ (**B**ilingual **W**eb **B**ooks). $\mathcal{BWB}$ consists of Chinese online novels and their corresponding English translations crawled from the Internet. Table 1 summarizes the statistics of the $\mathcal{BWB}$ 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. [3]

Statistic	Train	Test	Dev	Total
#Docs	196,304	80	79	196,463
#Sents	9,576,566	2,632	2,618	9,581,816
#Words	325.4M	68.0K	67.4K	460.8M

Table 1: Statistics of the proposed $\mathcal{BWB}$ dataset.

[3] The details of the corpus creation and quality control are described in Appendix A.

Dataset Split We treat chapters as *documents*. The maximum, median, and minimum number of sentences per document are 46, 30 and 18, respectively. We split the dataset into a training, development and a test set in units of *books*. 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 set.

3 Analyzing Discourse Errors

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

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

In this section, we conduct a human study on the test set of $\mathcal{BWB}$, in which we identify and categorize the discourse errors made by MT systems that are invisible 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 (in some cases, both). SENTENCE-level errors refer to those errors that cause the translations to be inadequate or not fluent as stand-alone sentences, while DOCUMENT-level errors lead to coherence violation across multiple sentences in the document. DOCUMENT-level errors are further categorized according to the linguistic phenomena leading to a discrepancy in context-dependent translations.[4]

Table 2 shows the result of error analysis. A substantial proportion of translations have document-level errors (71.9%). This verifies that $\mathcal{BWB}$ contains rich discourse phenomena that current 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 BLONDE.

Inconsistency We consider two kinds of consistencies in translation: lexical and grammatical.

[4] The annotation guidelines are described in Appendix B.

		ENTITY $\mathcal{E}$	TENSE $\mathcal{V}$	PRONOUN $\mathcal{P}$	DM $\mathcal{M}$
SRC	a) 小乔**(Qiao)** 看着***(look)*** 相片回忆***(recall)*** 起了二十年前。 b) 那个满脸胡须的男人**(man)** 正是***(be)***她**(she)** 的新婚丈夫。 c) 那却是***(be)*** 他们之间初次见面***(meet)***。 d) 小乔**(Qiao)**一见到他**(he)** 心里就咯噔***(jolt)*** 了一下， 噌的站***(stand)*** 起来。	⟦Qiao⟧	[VBD, VBZ]	[masculine, feminine, epicene, neuter]	[contigency, temporal, expansion, comparison]
REF	a) **Qiao** ***looked*** at the photo and ***recalled*** twenty years ago. b) This bearded **man** ***was*** **her** newlywedded **husband**, c) ⟦**yet**⟧ this ***was*** the first time they ***were meeting*** with each other. d) ⟦**So**⟧ **Qiao**'s heart ***jolted*** as soon as ⟦**she**⟧ saw **him**, and ⟦**she**⟧ quickly ***stood*** up.	[1] [0] [0] [1]	[2, 0] [1, 0] [2, 0] [2, 0]	[0, 0, 0, 0] [0, 1, 0, 0] [0, 0, 1, 0] [1, 2, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 1] [1, 0, 0, 0]
MTA	a) **Qiao** ***looked*** at the photo and ***recalled*** twenty years ago. b) The bearded **man** ***is*** **her** newlywed **husband**. c) This ***is*** the first time they ***meet*** with each other. d) **Joe**'s heart ***is*** squeaky as soon as ⟦**he**⟧ saw **him**, and ⟦**he**⟧ quickly ***stands*** up.	[1] [0] [0] [0]	[2, 0] [0, 1] [0, 2] [0, 2]	[0, 0, 0, 0] [0, 1, 0, 0] [0, 0, 1, 0] [3, 1, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 0] [0, 0, 0, 0]
MTB	a) **Qiao** ***looked*** at the photo and ***recalled*** the past twenty years ago. b) This bearded **man** ***was*** **her** newlywed **husband**. c) ⟦**However**⟧, that was the first time they ***met***. d) ⟦**So**⟧ as soon as **Qiao** saw **him**, ⟦**her**⟧ heart ***became*** squeaky, and ⟦**she**⟧ swiftly ***stood*** up.	[1] [0] [0] [1]	[2, 0] [1, 0] [2, 0] [2, 0]	[0, 0, 0, 0] [0, 1, 1, 0] [0, 0, 1, 0] [1, 2, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 1] [1, 0, 0, 0]

Figure 2: An example containing inconsistency and ellipsis in $\mathcal{BWB}$. 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.

	BLEU	BLONDE			BLOND-D
	P	P	R	F1	F1
MTA	47.6	47.6	11.9	19.1	3.0
MTB	41.1	69.1	68.1	68.6	98.5

Table 3: The BLEU and BLONDE scores of the two system outputs in Figure 2. P, R and F1 represent precision, recall and F-measure, respectively.

Lexical consistency is defined as a repetitive term keeping the same translation throughout the whole document (Carpuat 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 BLONDE, we mainly focus on the reiteration of named entities (e.g. **Qiao** in Figure 2). Typical grammatical consistency includes tense consistency and gender consistency. Tense consistency refers to the tense being *compatible* (rather than keeping exactly the same tense) with the context. It is prominent when the source language is an isolating language, e.g. Chinese, and the target language is synthetic language, e.g. English (***teal*** in Figure 2). In the same spirit, the same entity should maintain a consistent grammatical gender. [5]

Ellipsis Ellipsis denotes the omission from a clause of one or more words that are nevertheless understood in the context of the remaining elements (Voita et al., 2019; Yamamoto and Sumita, 1998).

[5] 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.

SRC	你在看***(kan)***什么？《复仇者联盟》。
REF	What are you ***watching***? The Avengers.
MT	What are you ***looking*** at? The Avengers.

Figure 3: An example of ambiguity. 看(kan) corresponds to *look*, *see*, *watch* and *view*. The correct translation can only be inferred from the next sentence (The Avengers).

Confusion arises when there are elliptical constructions in the source language while the target language does not allow 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 2, *she (Qiao)* is omitted in Chinese. However, it is hard to know 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 ellipsis that cannot be ignored 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 3, the word

看(kan) can be translated to *look*, *see*, *watch* or *view*. Without the context indicating *what* refers to a movie, all the lexical choices are coherent.

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, causing the document-level nuances to be overlooked (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** $\mathfrak{D} = [S_1, \ldots, S_N]$ of length N as a sequence of N sentences, each denoted as S_n. We take a **sentence** S_n of length T to be a string of tokens $t_1 \cdots t_T$ where each token t_i is taken from the target vocabulary $\mathcal{V}$. Let $\mathcal{M}(S_n) = \{m_1, m_2, \ldots\}$ be the set of spans in the sentence S_n. A **span** is a subsequence of the tokens that comprise $S_n = t_1 \cdots t_T$. We then define a **category** $C_k(S_n) \subseteq \mathcal{M}(S_n)$ as an equivalence class of spans that share a certain property k, e.g. POS-tag (`VBD`) or discourse relation (`contigency`).[6] Let a sequence of mutually exclusive categories[7] in a sentence S_n be

$$\mathbf{C}(S_n) = [C_k(S_n) : k = 1, 2, \ldots, K].$$

In our scenario, $\mathbf{C}(\cdot)$ corresponds to a certain discourse phenomenon.

Categorization Let $\mathsf{cat} : \mathcal{M}(S_n) \to \mathbf{C}(S_n)$ implement a categorization function from the set of spans in utterance S_n to the sequence of categories. The simplest implementation is strict string match. For example, the span *he* could be directly matched to the pronoun category `masculine`.

Similarity Measure Let $\mathsf{sim} : \mathbf{C}(S_n^s) \times \mathbf{C}(S_n^r) \to \mathbb{R}$ be an abstract similarity measure between two sequences of categories. sim takes non-negative value: zero means that $\mathbf{C}(S_n^s)$ and $\mathbf{C}(S_n^r)$ have nothing in common. It is desirable that the sim increases when $\mathbf{C}(S_n^s)$ and $\mathbf{C}(S_n^r)$ become similar. One possible choice is the number of common spans shared by $\mathbf{C}(S_n^r)$ and $\mathbf{C}(S_n^s)$. Now we turn from measuring the similarity at the sentence level to the document level. We lift the similarity measure to a system document $\mathfrak{D}^s$ and a set of reference documents $\mathbb{D}^r = \{\mathfrak{D}^{r_1}, \mathfrak{D}^{r_2}, \cdots\}$ by summing up the sim of all sentences in $\mathfrak{D}$ and $\mathbb{D}$. For clarity's sake, we abuse notation and write $\mathsf{sim}(\mathbf{C}(S_n^s), \mathbf{C}(S_n^r))$ as $\mathsf{sim}(S_n^s, S_n^r)$:

$$\mathsf{sim}\left(\mathfrak{D}^s, \mathbb{D}^r\right) = \sum_{S_n^s \in \mathfrak{D}^s} \bigoplus_{S_n^r \in \mathbb{U}_n^r} \mathsf{sim}(S_n^s, S_n^r) \quad (1)$$

where $\mathbb{D}^r = \{\mathfrak{D}_n^{r_i} : \mathfrak{D}^{r_i} \in \mathbb{D}^r; i = 1, 2, \cdots\}$, and $\oplus$ 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. Here we also reuse the notation $\mathsf{sim}(\cdot, \cdot)$ for two documents and for two sets of documents:

$$\mathsf{sim}\left(\mathfrak{D}^s, \mathfrak{D}^{s'}\right) = \sum_{n=1}^{N} \mathsf{sim}(S_n^s, S_n^{s'}) \quad (2)$$

$$\mathsf{sim}\left(\mathbb{D}^r, \mathbb{D}^{r'}\right) = \sum_{n=1}^{N} \bigoplus_{S_n^r \in \mathbb{U}_n^r, S_n^{r'} \in \mathbb{U}_n^{r'}} \mathsf{sim}(S_n^r, S_n^{r'})$$

Scoring We are now ready to define the "goodness" of a system output in regards to a certain discourse phenomenon $\mathbf{C}(\cdot)$. The precision, recall and F-measure are defined as follows:

$$p = \frac{\mathsf{sim}\left(\mathfrak{D}^s, \mathbb{D}^r\right)}{\mathsf{sim}\left(\mathfrak{D}^s, \mathfrak{D}^s\right)}, \qquad r = \frac{\mathsf{sim}\left(\mathfrak{D}^s, \mathbb{D}^r\right)}{\mathsf{sim}\left(\mathbb{D}^r, \mathbb{D}^r\right)}$$

$$F = \frac{2pr}{p + r}. \quad (3)$$

4.2 BlonD-d

So far, we have tacitly kept cat and sim abstract. In this section, we introduce one operationalization of these two functions and a way of combining multiple categories that are related to discourse phenomena.

Categories 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 $\mathcal{E}$, TENSE $\mathcal{V}$ and PRONOUN $\mathcal{P}$ (See Figure 3). $\mathcal{E}$ is a sequence of named entities in

[6] We have left the definition of categories intentionally open-ended and will revisit this point in Section 4.2

[7] Each mention m can only belong to one C_k in $\mathbf{C}$.

$\mathfrak{D}$, $\mathcal{V}$ = [MD, VBD, VBN, VBP, VBZ, VBG, VB],[8] and $\mathcal{P}$ = [masculine, feminine, neuter, epicene][9]. In addition, we introduce **discourse markers** DM as categories: $\mathcal{M}$ = [contigency, temporal, expansion, comparison][10]. The categorizing function cat then can be operationalized as a NER model, a POS parser, a rule-based string match and a discourse marker miner for ENTITY, TENSE, PRONOUN and DM, respectively. Note that the number of ENTITY categories depends on $\mathfrak{D}$ while the numbers of TENSE and PRONOUN categories 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.

Similarity How similar two sequences of categories are can be measured by the counts of their matched spans:

$$\begin{aligned}&\mathsf{sim}(S_n^s, S_n^r) = \\ &\quad \boldsymbol{w} \odot \min(\mathsf{count}(\mathbf{C}(S_n^s)), \mathsf{count}(\mathbf{C}(S_n^r)))\end{aligned} \tag{4}$$

where $\mathsf{count}(\mathbf{C}(\cdot)) = [|C_k(\cdot)| : k = 1, 2, \cdots, K]$, and $\boldsymbol{w} \in \Delta^{K-1}$ is a weight vector. Intuitively, this measures how many functionally similar spans they share. For example, in Figure 2, $\mathsf{sim}(U_b^{\text{MTA}}, U_b^{\text{REF}}) = 0$ since MTA mistranslated all the verb into the present tense due to the ignorance of context. The total similarity $\mathsf{sim}(\mathfrak{D}^{\text{MTA}}, \mathfrak{D}^{\text{REF}})$ is the (weighted) total shared spans in all subcategories: (1, 2, 4, 0) for ($\mathcal{E}$, $\mathcal{V}$, $\mathcal{P}$, $\mathcal{M}$). The denominators of Equation (3) are thus the numbers of detected spans in the system output (1, 6, 6, 0) and in the reference (2, 7, 5, 2), respectively. F1 is then the ratio of spans that are in the correct category.

It is worth noting that there are many other reasonable ways to operationalize 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 sim definition to future work.

BLOND-D Further, we combine these four scores into an overall score by a simple weighted averaging approach, named as BLOND-D. By computing BLOND-D, one can distill the document-level translation quality from the sentence-level one.

$$\begin{aligned}&\text{BLOND-D}(\mathfrak{D}^s, \mathbb{D}^r) = \\ &\quad \Bigg(\prod_{\mathbf{C}\in\{\mathcal{E},\mathcal{V},\mathcal{P},\mathcal{M}\}} (F(\mathfrak{D}^s, \mathbb{D}; \mathbf{C})^{w_i}\Bigg)^{1/\sum_i w_i}\end{aligned} \tag{5}$$

where w_i is the weight corresponding to a certain type of categories $\mathbf{C}$, and $F(\cdot, \cdot; \mathbf{C})$ is the scoring in Equation (3), as described in Section 4.1.[11] The BLOND-D.R, BLOND-D.P and BLOND-D.F1 of MTA are $(\frac{1}{2})^{\frac{1}{4}}(\frac{2}{7})^{\frac{1}{4}}(\frac{4}{5})^{\frac{1}{4}}(\frac{\delta}{2})^{\frac{1}{4}} = .03$, $(\frac{1}{1})^{\frac{1}{3}}(\frac{2}{6})^{\frac{1}{3}}(\frac{4}{6})^{\frac{1}{3}} = .04$ and $\frac{2\cdot.03\cdot.04}{.03+.04} = .036$, respectively.

4.3 BLONDE: Combining 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. On account of that, we further calculate the same statistics of n-grams by simply treating each n-gram (span) as a singleton category, and combine them all together:

$$\begin{aligned}&\text{BLONDE}(\mathfrak{D}^s, \mathbb{D}^r) = \\ &\quad \Bigg(\prod_{\substack{\mathbf{C}\in\{\mathcal{E},\mathcal{V},\mathcal{P},\mathcal{M}\}\cup \\ \{n\text{-gram}:n=1,2,\cdots,N\}}} (F(\mathfrak{D}^s, \mathbb{D}; \mathbf{C})^{w_i}\Bigg)^{1/\sum_i w_i}\end{aligned} \tag{6}$$

BLONDE covers both discourse coherence features and the sentence-level adequacy, thus provides 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.

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

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

[10] A detailed explanation is provided in Table 6

[11] BLOND-D adopts uniform weights. The weighted arithmetic mean can also be applied.

4.4 BLOND+: Combining with Human Annotations

BLONDE is highly extensible and it is easy to incorporate human annotations: we can annotate spans related to a discourse errors and treat them as categories. The automated detected categories and human annotated categories is then combined by adopting the same weighted averaging approach. We name it as BLOND+. We hired the same translators who conducted Section 3 to annotate ambiguous and omitted word/phrases on the test set of $\mathcal{BWB}$. This annotated test set is also publicly available as a testbed for evaluating MT system's capacity to disambiguate word senses and to predict coherent pronouns or discourse markers in the case of omission.

5 Experiments

In this section, we examine the effectiveness of BLONDE at the document through experiments. The following question needs to be answered:

- 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). LC and RC are the ratios between the number of lexical cohesion devices (i.e. repetition and collocation) and repeated content words over the total number of content words in a target document. They are direct measurements of lexical cohesion.

Embedding-based metrics SKIP (SkipThought cosine similarity (Kiros et al., 2015)), AVER (Embedding average cosine similarity (Sharma et al., 2017)), VECTOR (Vector extrema cosine similarity (Forgues et al., 2014)), GREEDY (Greedy Match (Rus and Lintean, 2012)).

5.1 MT Systems

We test BLONDE on the following system outputs: a SMT system (Chiang, 2007), three well-known online 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 $\mathcal{BWB}$. MT-D (Zhang et al., 2018) trains sentence-level model parameters first and then estimates document-level model parameters while keeping the learned original 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.[12]

5.2 The BLONDE Evaluation

Firstly, we leverage the test set of $\mathcal{BWB}$ 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 Δ(MT-S, MT-D)) is significantly higher than the difference between the Δ(MT-S, MT-D) in their BLEU scores. An even larger Δ 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", indicating BLOND-D indeed distills document-level translation quality.

The paired t-statistics 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 power across the three pair comparisons. Interestingly but not surprisingly, the non-reference-based LC and RC fail to distinguish neither (MT-S, MT-D) nor (MT-D, PE), since sentence-level MT is by nature more repetitive than human translation, thus hard to distinguish accidental repetition from document-level cohesion.

In addition, the t-statistics of BLOND-D categories provide rich diagnostic information. As can be seen, although transformer-based NMT models have way higher BLEU scores than SMT, MT-S is

[12] We trained models by fairseq (Ott et al., 2019). Model parameters and the post-editing details are in Appendix F.2 and C, respectively.

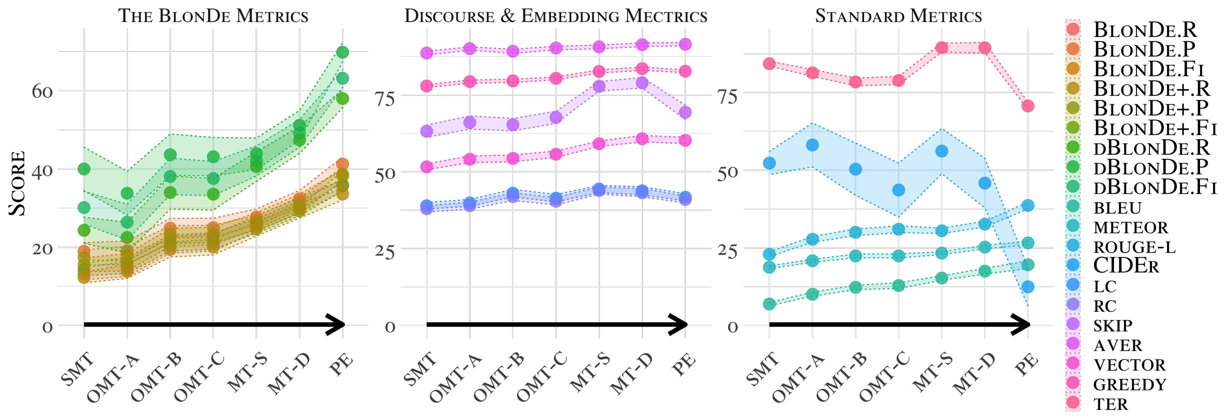

Figure 4: The mean scores of different system outputs given by different metrics on the $\mathcal{BWB}$ test set. Shaded region represents 95% confidence interval.

	BLEU	BLONDE R	BLONDE P	BLONDE F1	BLOND+ R	BLOND+ P	BLOND+ F1	BLOND-D R	BLOND-D P	BLOND-D F1	Categories $\mathcal{E}$	Categories $\mathcal{V}$	Categories $\mathcal{P}$	Categories $\mathcal{M}$
SMT→ MT-S	25.8	13.5	7.42	10.9	**14.5**	8.51	12	8.02	1.32	5.1	-2.12	23.6	11.4	13.6
MT-S→ MT-D	8.97	6.32	5.45	5.92	**6.58**	5.61	6.13	4.85	4.57	4.79	4.93	1.88	7.43	1.62
MT-D→ PE	2.6	4.51	7.77	6.06	4.2	7.27	5.66	6.44	**11.1**	8.58	12.9	2.44	2.76	5.35

	Other Standard Metrics METEOR	ROUGE-L	TER	CIDER	Discourse Cohesion LC	RC	Embedding-based Metrics SKIP	AVER	VECTOR	GREEDY
SMT→ MT-S	25.3	19.8	-8.28	.853	11.7	12.9	12.2	9.5	18	22.3
MT-S→ MT-D	13.4	11.8	.148	-3.03	-1.23	-1.45	1.62	3.13	5.05	5.83
MT-D→ PE	3.58	9.65	19.9	-6.67	-4.23	-4.44	-6.23	.628	-1.03	-3.15

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.

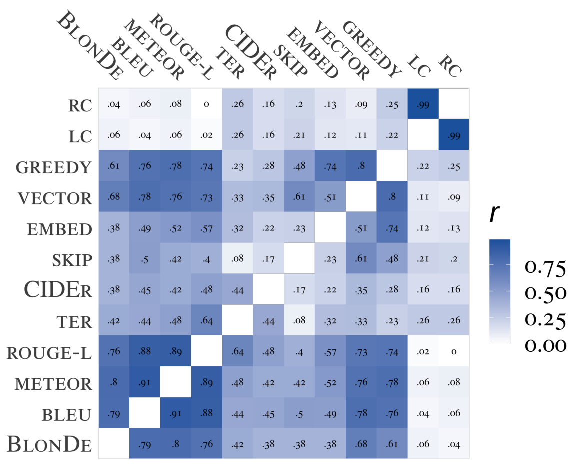

Figure 5: Absolute Pearson correlation pairs of automatic metrics. Computed over the scores of individual documents in $\mathcal{BWB}$ test set.

not statistically superior to SMT in terms of named entity translation. However, human post-editing scores significant better machine translations in 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 is not doing significantly better than 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 correlation for TER as it aims for a strong negative correlation with human assessment. 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, suggesting their heterogeneity.

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 raters isolated sentences, while in the DOCUMENT-level evaluation, entire documents are presented and we only ask raters to evaluate the overall quality of sequential blocks of sentences (5 sentences per block). We adopt

Unit	SENTENCE ADE	SENTENCE FLU	DOCUMENT ADE	DOCUMENT FLU
BLONDE.R	**.363**	.327	.436†	.371†
BLONDE.P	.331	.296	.383†	.344†
BLONDE.F1	.35	.314	.417†	.358 †
BLOND+.R	.364	**.329**	**.44**†	**.373**†
BLOND+.P	.334	.3	.39†	.349†
BLOND+.F1	.351	.318	.422†	.362†
BLEU	.325	.308	.343	.266
METEOR	.338	.31	.339	.278
ROUGE-L	.275	.262	.29	.211
TER	.063	.027	.044	.092
CIDER	.139	.116	.114	.087
SKIP	.213	.174	.163	.171
AVER	.163	.163	.16	.111
VECTOR	.25	.243	.248	.218
GREEDY	.323	.3	.307	.265
LC	.086	.061	.153	.116
RC	.096	.07	.169	.13

Table 5: Absolute Pearson correlation with human judgments on $\mathcal{BWB}$. 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.

Relative Ranking (RR) (Bojar et al., 2016). The detailed protocol is presented in Appendix D. We employ Williams significance test (Williams, 1959; Graham and Baldwin, 2014) following the practice adopted by WMT (Mathur et al., 2020).

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 correlates better with human assessment, yet work worse in selectivity compared to the precision-based variants (see MT-D→ PEin Table 4). This provides support for adopting the F-measure in terms to get the best of both worlds.

6 Related Work

There have been a few works on automatic evaluation metrics for specific discourse phenomena. For **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. Compared to those metrics, BLONDE does not rely on any alignment or training. For **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. For **discourse relations**, Hajlaoui and Popescu-Belis (2013) proposed to assessing the accuracy of connective translation (ACT). However, it needs a bilingual dictionary of all possible DM translations, while BLONDE only demands a list of monolingual DMs. Moreover, BLONDE has higher tolerance of valid drop (Zufferey and Cartoni, 2012), where ACT suffers due to its recalled-based exact match. Guzmán et al. (2014) and Joty et al. (2014) exploited the discourse structure by computing a similarity measure 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, thus has significantly higher Pearson's r correlation coefficients with human assessments.

7 Conclusion

In this paper, we build a large-scale parallel dataset for document-level translation, $\mathcal{BWB}$. 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 $\mathcal{BWB}$ test set, and release the crawling script of the training set under Fair Use rules. The BLONDE package will also be public available.

References

N Agrawal and A Singla. 2012. Using named entity recognition to improve machine translation. *Technical report, Standford University, Natural Language Processing*.

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.

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.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. 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.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Marine Carpuat and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada. Association for Computational Linguistics.

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

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal. Association for Computational Linguistics.

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.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference* on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 236–247. Springer.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

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.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of ACL-08: HLT*, pages 389–397, Columbus, Ohio. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA. Association for Computational Linguistics.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. 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 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

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 *MTSUMMIT*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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).

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France. European Language Resources Association.

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.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. 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 NAACL-HLT 2019: Demonstrations*.

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.

Yip Po-Ching and Don Rimmington. 2004. Chinese: A comprehensive grammar. *NY: Routledge*.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. 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.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. 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).

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Natasha Tokowicz and Tamar Degani. 2010. Translation ambiguity: Consequences for learning and processing. *Research on second language processing and parsing*, pages 281–293.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

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.

Evan James Williams. 1959. *Regression analysis*, volume 14. wiley.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2014. Semantics, discourse and statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*, pages 11–12, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kazuhide Yamamoto and Eiichiro Sumita. 1998. Feasibility study for ellipsis resolution in dialogues by machine-learning technique. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

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.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1121–1128, Manchester, UK. Coling 2008 Organizing Committee.

Sandrine Zufferey and Bruno Cartoni. 2012. English and french causal connectives in contrast. *Languages in contrast*, 12(2):232–250.

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[13] (Sennrich and Volk, 2011), 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 $\mathcal{BWB}$ 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.

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

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 according to the linguistic phenomena that lead to errors in MT outputs when considering context.

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+M. 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

CATEGORIES	DESCRIPTION	MARKERS
contingency	only consider "cause"	["but", "while", "however", "although", "though", "still", "yet", "whereas", "on the other hand", "in contrast", "by contrast", "by comparison", "conversely"]
comparison	combine "concession" and "contrast"	["if", "because", "so", "since", "thus", "hence", "as a result", "therefore", "thereby", "accordingly", "consequently", "in consequence", "for this reason"]
expansion	only consider "conjunction"	["also", "in addition", "moreover", "additionally", "besides", "else,", "plus"]
temporal	"synchronous" "asynchronous"	["meantime", "meanwhile", "simultaneously"] ["when", "after", "then", "before", "until", "later", "once", "afterward", "next"]

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

Dataset	Domain	#Docs	#Sents
WMT (Barrault et al., 2019)	News	68.4k	3.63M
OpenSubtitles (Lison et al., 2018)	Subtitles	29.1k	31.2k
TED (Ansari et al., 2020)	Talks	1K	219M
$\mathcal{BWB}$	Books	196k	9M

Table 7: Comparison of different document-level datasets.

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 five systems (SMT, OMT-B, MT-S, CTX, PE). We follow the protocol proposed by (Läubli et al., 2018, 2020). We conduct the evaluation experiment with a 2×2 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 by 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 $\mathcal{BWB}$ 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. On 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 raters evaluate 162 documents (plus 18 spam items) and 162 sentences (plus 18 spam items). We use two non-overlapping sets of docu-

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

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.

ments 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:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{7}$$

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 BLONDE

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 $\mathcal{BWB}$ 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.