**Abstract**

Standard automatic metrics (such as BLEU) are problematic for document-level MT evaluation. They can neither distinguish document-level improvements in translation quality from sentence-level ones, nor can they identify the specific discourse phenomena that caused the translation errors. To address these problems, we propose an automatic metric BlonD\(^1\) for document-level machine translation evaluation. BlonD takes discourse coherence into consideration by calculating the recall and distance of check-pointing phrases and tags, and further provides comprehensive evaluation scores by combining with n-gram. Extensive comparisons between BlonD and existing evaluation metrics are conducted to illustrate their critical distinctions. Experimental results show that BlonD has a much higher document-level sensitivity with respect to previous metrics. Human evaluation also reveals high Pearson R correlation values between BlonD scores and manual quality judgments.\(^2\)

**1 Introduction**

Some recent works (Wu et al., 2016; Hassan et al., 2018) suggest that neural machine translation (NMT) approaches have achieved comparable accuracy to average bilingual human translators or even professional human translators. Nevertheless, most of the current NMT models operate at the sentence level, ignoring the coherence of the text. Recently, document-level machine translation has received a lot of attention in the machine translation (MT) community. However, the progress and widespread adoption of document-level MT approaches is hampered by the lack of efficient document-level metrics.

\(^{\ast}\)Most of the work was done while the first author was an intern at Microsoft Research.

\(^{1}\)BlonD: Bilingual Evaluation of Document Translation.

\(^{2}\)The code and data will be publicly released.

\(^{3}\)By inconsistency, we mean the mistakes related to coreference and lexical cohesion (Carpuat, 2009; Guillou, 2013).
Table 1: Statistics information on the training, testing, and development sets.

| 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 | 68.0M | 67.4M | 460.8M |

We also observe that the correlation between BlonD and sentence-level metrics is lower than the correlations between the metrics belonging to the same sentence-level category, indicating that BlonD captures new features of MT quality that are different from those metrics. We further show that unlike previously proposed metrics for specific discourse phenomena, BlonD scores has a high Pearson R correlation value with the human assessment.

2 Analysis of Discourse Errors

To design a metric that is more sensitive to the document-level improvements of MT systems, we first prepare a corpus that contains rich discourse phenomena and manually analyze discourse errors made by machine translation systems that are invisible in sentence-level evaluation.

2.1 Dataset

First, we construct a large Bilingual Web Book (BWB) dataset, which contains many scenarios that are suitable for displaying the system’s document-level translation capabilities (e.g. emails, books, stories, novels, subtitles).

Dataset Construction  We crawl down the Chinese books and their corresponding English translations from the Internet. Then, we use the tools provided by Sennrich and Volk (2011) to align these corresponding books. We randomly select 163 documents to evaluate the quality of the dataset, and observe a translation accuracy rate of over 90%. We further hire professionals to proofread the test&dev set to ensure their alignment and translation quality.

Dataset Split  We treat a chapter as a document, which usually contains more than 30 sentences (The number of sentences per document varies between 18 and 46). We divide the training set, development set and test set in units of books. To avoid overfitting, documents from the same book will not appear in the train and test sets at the same time. We use 377 books as training sets, and selected 79 and 80 documents from the remaining 6 books as the test and dev set, respectively. Table 1 presents the statistics of BWB. To the best of our knowledge, this is to be the largest Chinese-English document-level dataset.

2.2 Error Analysis

During the assessing process, the annotators are asked to distinguish between document-level error and sentence-level error (or 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” denotes those errors causing the 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.

2.3 Analysis Results

Table 2 shows the statistics of manual analysis. Firstly, a substantial proportion of translations have document-level errors (71.9%). This verifies that BWB contains rich discourse phenomena that current common MT systems cannot address. Secondly, it shows that three major categories (inconsistency (64.4%), ellipsis (20.3%) and ambiguity (7.3%)) account for almost all document-level errors. We will discuss them in more detail below.

Table 2: Human analysis statistics of translation errors.

| Type                | #   | %    |
|---------------------|-----|------|
| No Error            | 451 | 17.1%|
| Sentence-level      | 1351| 51.3%|
| Document-level      | 1893| 71.9%|
| Inconsistency       | 1695| 64.4%|
| Named Entity        | 1139| 43.3%|
| Tense               | 1018| 38.7%|
| Ellipsis            | 534 | 20.3%|
| Ambiguity           | 193 | 7.3%  |

Inconsistency  Lexical consistency is defined as a repeated term keeping the same translation throughout the whole document (Carpuat and Simard, 2012), also known as lexical cohesion. Guillou (2013) finds whether lexical consistency should be encouraged depends on parts-of-speech: the consistent translation of nouns proves beneficial while encouraging the consistency of verbs would be undesirable. We focus on the most common cases in repetition: reiteration of named entities (underline in Figure 1).
there is an ellipsis in the source language, choosing either gender/number pronoun might be reasonable. However, with context, there usually exists only one right choice. It shares some similarities with the linguistic phenomenon focused by pronoun prediction tasks (Guillou, 2013; Loiiciga et al., 2017). However, following Voita et al. (2019), only the ellipsis that can only be understood and translated with context beyond the sentence-level are considered in this study.

**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). Cross-language ambiguity phenomenon comes from several sources of within-language ambiguity including lexical ambiguity, polysemy, and near-synonymy. The unified feature of them is that all ambiguous terms satisfy the form of one-to-many mappings (for example “looking at” in Figure 2). Translation ambiguities exist extensively but we only focus on ambiguities which are caused by lack of context.

3 **BlonD**

The document-level phenomena mentioned above may have an impact on relatively few word forms, but they are the key considerations when manually evaluating document translations. However, standard automatic metrics ignores their importance for contextual coherence, causing the document-level improvements being overlooked. This is also pointed out in Xiong and Zhang (2014) and Zhou et al. (2008). In this section, we describe BlonD, an automatic metric that explicitly tracks discourse phenomena.

3.1 The Discourse Level Evaluation

We first give the formulation of measuring identified discourse checkpoints, being named entities, tense and pronouns, since they make up the majority of discourse errors. Formally, we define a document $D$ as a list of sentences. $E$ is the set of named entities corresponding to $D$. $V$ is the set of POS-tags related to tense: $V = \{MD, VBD, VBN, VBP, VBZ, VBG, VB\}^5$. $P$ is the

| BLEU | BlonD | dB-D | BD-d | dB-D-d |
|------|------|------|------|-------|
| MT1  | 21.37| 11.66| 13.14| 98.99 |
| MT2  | 11.50| 41.63| 79.05| 71.00 |

Figure 1: An example containing inconsistency and ellipsis along with its BlonD scores. For inconsistency, named entities are underlined and verbs are bold. Ellipsis is marked in bold Italic. For the table, $BD$ is the shortcut for BlonD. “$d$” means distance-based variations (lower is better).

We also consider another subcategory of consistency: grammatical consistency. Typical grammatical consistency includes tense consistency and gender consistency. Tense consistency means the tense should be compatible(rather than keeping exactly the same tense) with the context (bold in Figure 1). Tense inconsistency is conspicuous when the source language is an isolating language (e.g. Chinese) and the target language is synthetic language (e.g. English). Similarly, gender consistency means the same entity maintains a consistent grammatical gender. It is worth noting that the analysis and the metric proposed in this study can be applied to a wider range of language pairs by extending the definition of grammatical consistency.

**Ellipsis** Ellipsis is 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). 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 and objects is very common in Chinese while it is illegal in English, especially for pronouns (bold Italic in Figure 1). In this example, “she” is omitted in Chinese. However, it is hard to know the gender of Qiao Lian from this stand-alone sentence, the correct pronoun choice can only be inferred from context (Qiao Lian had a husband so “she” is more likely to be the right choice). For stand-alone sentences, when

**SRC**  这个人是乔恋的新婚丈夫。但是这却是他们之间初次见面。乔恋心里咯噔一下，嘈的站起来。

**REF**  This person was Qiao Lian’s newlywedded husband, yet this was the first time they were meeting with each other. Qiao Lian’s heart jolted, and she quickly stood up.

**MT1**  This person is Qiao Lian’s newlywedded husband. However, this is the first time they meet with each other. Joe’s heart is squeaky and he quickly stands up.

**MT2**  This man was Qiao Lian’s newlywedded husband. However, they met for the first time. Qiao Lian’s heart became squeaky and she swiftly stood up.

**Figure 2:** An example of ambiguity.
set of pronouns: $P = \{he/him/his, she/her/hers, it/its, they/them/their/their\}$.

Suppose there are $n$ references, we define $c^E_r$ as the count of names entities $e \in E$ in the $i$-th reference, and $c^s_r$ is defined as the count of this named entity in the system translation. We then define $c^v_r$ as the count of verbs which belong to the tense $v \in V$ in the $i$-th reference, and $c^w_r$ is defined as the same count in the system translation. Similarly, $c^P_r$, $c^V_r$ is the count of the pronoun $p$.

Then we define $C^E_{ri}$ as:

$$C^E_{ri} = w^E \odot c^E_r$$  \hspace{1cm} (1)

where $w^E$ is the weights of the named entities. Similar definitions for $C^V_{ri}$ and $C^P_{ri}$, where $w^V$ is a fix vector of length 7 denoted as the weights of 7 POS-tag categories and $w^P$ is a fix vector of length 4 denoted as the weights of 4 pronoun categories.

REF  He rejected the call irritatedly and cursed, “This Wang Wenhao is just neurotic!”

MT  She snaps the phone and curses, “This Wang Wenhao is just neurotic!”

For this example, the different counts corresponding to $E, V, P$ are listed as follows:

$E = \{"Wang Wenhao"\}, c^E_r = [1], c^E_s = [1]$

$c^V_r = [0, 2, 0, 0, 1, 0, 0], c^V_s = [0, 0, 0, 3, 0, 0]$

$c^P_r = [1, 0, 0, 0], c^P_s = [0, 1, 0, 0]$

We further compute the match between the candidate and the reference $i$:

$$C^E_{mi} = \min(C^E_{ri}, C^E_s)$$  \hspace{1cm} (2)

and compute the scores respectively by:

$$S^E_i = \frac{||\hat{C}^E_{mi}||_1}{||\hat{C}^E_{ri}||_1}$$  \hspace{1cm} (3)

where $\hat{C}^E_{mi}, \hat{C}^E_{ri}$ denote the concatenation of $C^E_{mi}$ and $C^E_{ri}$ corresponding to all the sentences in $D$, $|| \cdot ||_1$ denotes the Euclidean 1-norm (i.e. the sum of the absolute values of the elements). When there are multiple references, we always choose the reference which provides the largest $S^E_i$. We compute $S^V$ and $S^P$ in a similar way.

The intuition here is that for a specific named entity (for example, “Qiao Lian”) we want the count of it in the candidate to be as close as possible to that in the references. Similarly, we want to encourage the system translation to keep consistent tense and pronouns (especially gender pronoun) to the references. On account of that, the weights of POS-tags (pronouns) are deliberately designed to value more on those tags (pronouns) which have higher correlation with human evaluation. After extensive experiments and evaluations, we assign $w^E$ to uniform weights, $w^V$ to $(0.2, 0.2, 0.05, 0.2, 0.15, 0.05, 0.15)$, $w^P$ to $(0.45, 0.45, 0.05, 0.05)$.

**dBlonD**  Further, we combine these three scores into an overall score by a simple weighted mean approach. We name it dBlonD, which is a comprehensive reflection of discursive coherence and cohesion. By computing the dBlonD score, one can distill the translation quality at the document level from the sentence level one.

$$dBlonD = \left(\prod_{i \in \{E, V, P\}} (S^i_{wi})^{1/\sum_i w_i}\right)$$  \hspace{1cm} (4)

In practice, we adopt uniform weights for $w_i$. The weighted arithmetic mean can also be applied.

### 3.2 Combining with n-gram Recall

Focusing on discourse phenomena alone is not enough to provide comprehensive MT evaluation results.

**REF**  Ye Qing Luo lifted her heavy eyelids.

**MT**  Ye Qing Luo scrunched her brows together.

The dBlonD score of the MT translation in this example is 1, but obviously this translation is far from “good” in terms of adequacy. On account of that, we further calculate the recall rate of n-grams in the same way as the calculation of $S^E, S^V, S^P$, and combine them all together. This allows all aspects of translation quality (discursive coherence and cohesion as well as sentence-level adequacy and fluency) to be taken into account.

Since recall-based calculations naturally encourage long sentences, similar to BLEU’s use of the short penalty, we added a long penalty to BlonD. The overall computation is shown as follows:

$$S = \{S^i | i \in \{1, 2, 3, 4\} \cup \{S^E, S^V, S^P\}\}$$

$$LP = \exp(1 - \frac{c}{r}), \text{ if } c \geq r, \text{ else } 1$$
\[ \text{BlonD} = LP \cdot \left( \prod_{S \in S} (S^i)^{w_i} \right)^{1/\sum_i w_i} \]  

(5)

where \( S^i \) refers to the recall of \( i \)-gram, \( c \) and \( r \) refer to the length of the candidate and the corresponding reference, respectively. We assign \( w_E, w_V, w_P \) to a value greater than the \( n \)-gram weights. In practice, we adopt uniform weights for \( w_i \).

When calculating BlonD, pronouns can be directly counted, and named entity recognition and POS-tags can be easily acquired by language processing tools such as spacy (Honnibal and Montani, 2017). One may concern the impact of NER accuracy on BlonD performance. But after extensive evaluation experiments, we prove that since BlonD only assigns weights to a few NER tags (PERSON, NORP, GPE, FAC, ORG, WORKOFART) and merges them into only two very coarse-grained categories (PERSON, NON-PERSON), the requirements for NER reliability are actually very low.

Example Figure 1 shows two versions of MT output of a segment selected from a document in BWB. Humans can easily judge that MT2 is better than MT1, but we observe that the BLEU score of MT2 is lower. In sharp contrast, their BlonD scores reflect the true difference in translation quality.

4 BlonD Extensions

We further explored a distance-based BLOND variant and a method of incorporating human annotation into BLOND.

4.1 BlonD-d: A Distance-based Variation

Since minimizing the distance between the counts of the candidate and the reference is the most intuitive way to keep their number as close as possible, we compute the distance between the candidate and the reference \( i \):

\[ C_{d_i} = |C_{r_i} - C_s| \]

(6)

and compute \( S^E_i, S^P_i, S^V_i \) respectively by:

\[ S_i = \frac{\|\hat{C}_{d_i}\|_\alpha}{\|\hat{C}_{r_i}\|_\alpha} \]

(7)

where \( \hat{C}_{d_i}, \hat{C}_{r_i} \) denote the concatenation of \( C_{d_i} \) and \( C_{r_i} \), corresponding to all the sentences in \( D \), \( \| \cdot \|_\alpha \) denotes the Euclidean \( \alpha \)-norm. The calculation of distances for \( n \)-grams is the same. When there are multiple references, we always choose the reference which provides the smallest \( S_i \). In the distance-based computation, \( \alpha \) determines the sensitivity. It is not hard to prove that when \( \alpha > 1 \) it is more sensitive to replacement (for example, replace “he” with “she”) while when \( \alpha < 1 \) it is more sensitive to absolute changes in counts. In practice, we assign \( \alpha \) to 2.

4.2 BlonD+: Combining with Human Annotation

BlonD is highly extensible. When we want to consider some discourse phenomena that are difficult to be automatically evaluated (e.g. the ambiguity of certain content words or phrases), we can use an extremely simple annotation protocol to annotate ambiguous terms, and then use a calculation method similar to the above-mentioned method based on the recall (or distance) to integrate the manual annotation into BlonD. Specifically, the calculation of ambiguity scores is similar to named entity scores (taking ambiguity terms as named entities). The method of incorporating into BlonD also adopts the same weighted mean.

Annotating Protocol We conduct a human annotation for ambiguous terms on BWB following the mentioned protocol. The annotating instruction is straightforward: mark those terms that are translated correctly when only a single sentence is considered, but wrong when considering the context. Its detailed guideline is listed in the Appendix. We also make the annotated set public as a test set for discursive ambiguity.

Remark: Why recall? Unlike BLEU, BlonD calculates the recall (or distance) based on the reference, instead of the precision based on the candidate. Choosing the count in the reference as the denominator, there are two main desiderata: One is the extensibility. Since the calculation of BlonD is based on the reference translation, when expanding BlonD to incorporate human annotations, only one simple checkpoint annotation is needed, and then the entire community can use this annotation to calculate BlonD + scores. Calculating precision does not allow repeated evaluations with new candidate sentences. Another is the reliability of NER. The quality of the candidate translation cannot be guaranteed, and poor translation quality will lead to unreliable NER results. The NER reliability of the reference translation is much higher.
We evaluate 9 systems in total: an SMT system (Chiang, 2007), several well-known online commercial NMT systems (OMT1-OMT3), a sentence-level baseline (MTs) and a document-level baseline (MTd) trained on BWB. We adopt Transformer Big (Vaswani et al., 2017) for both MTs and MTd; MTd uses a special token at both the source and target side to separate sentences. Besides, two other document-level systems were evaluated. CADec (Voita et al., 2019): First train a context-agnostic model, then train a context-aware decoder which refines context agnostic translations using context. CTX (Zhang et al., 2018): Train sentence-level model parameters first and then estimate document-level model parameters while keeping the learned original sentence-level Transformer model parameters fixed. These two models are also trained on in-domain data. The final system is a human translation system (HT) provided by professional translators, so it is supposed to be the strongest baseline.

5.2 The BlonD Evaluation

Firstly, we leverage the test set of BWB (6232 sentences) and evaluate the above-mentioned 9 systems mentioned above by BlonD and other standard metrics.

Table 3 presents the means and variances of scores computed by different metrics on the test set of BWB.

5.1 MT Systems

We evaluate 9 systems in total: an SMT system (Chiang, 2007), several well-known online commercial NMT systems (OMT1-OMT3), a sentence-level baseline (MTs) and a document-level baseline (MTd) trained on BWB. We adopt Transformer Big (Vaswani et al., 2017) for both MTs and MTd; MTd uses a special token at both
Table 4: The paired t-statistics of different MT systems. The p-value is denoted as † or *: † < .05, †† < .01, ††† < .001, * > .05, ** > .1, *** > .5.

Table 5 presents the correlation rates of different evaluation metrics under different settings, with their 95% confidence intervals (CI) provided. BlonD obtains the highest correlation with human assessments at both the sentence level and the document level.

5.3 Manual Evaluation

We then evaluate BlonD along with other metrics in terms of their correlation with human assessments. Human assessments are provided by professional English-Chinese translators and two experimental units (sentence vs document) are assessed independently. We follow the standard Relative Ranking (RR) (Bojar et al., 2016). For document-level evaluation, human assessors first read the entire document in the source language and take context into account.

Human evaluation shows that there is no significant difference at the sentence level between CTX and HT, but HT is still superior to CTX at the document level, consistent with the conclusion in Läußli et al. (2018) that MT does not achieve human parity at the document level. Details are in the Appendix.

Table 5 presents the correlation rates of different evaluation metrics under different settings, with their 95% confidence intervals (CI) provided. BlonD obtains the highest correlation with human assessments at both the sentence level and the doc-
Table 5: The correlation rates of evaluation metrics with human assessments along with the 95% CI.

| Metric | Sentence | Document |
|--------|----------|----------|
| BLEU   | .810 (.711, .909) | .797 (.701, .895) |
| METEOR | .867 (.794, .940) | .856 (.782, .925) |
| ROUGE-L | .790 (.700, .879) | .798 (.701, .895) |
| CIDEr  | .588 (.407, .769) | .597 (.410, .783) |
| BlonD  | **.878** (.784, .971) | **.884** (.815, .957) |
| dBD    | .716 (.568, .864) | .698 (.535, .864) |
| BlonD+ | .877 (.793, .960) | .866 (.815, .957) |
| dBD+   | .735 (.608, .861) | .706 (.573, .838) |
| LC     | -.559 (.379, .739) | -.564 (.373, .758) |
| RC     | .706 (.553, .858) | .704 (.543, .865) |
| Skip   | .633 (.469, .797) | .631 (.484, .818) |
| Aver   | .377 (.127, .627) | .364 (.125, .602) |
| Vector | .758 (.597, .919) | .764 (.608, .920) |
| Greedy | .535 (.347, .722) | .542 (.357, .726) |
| TER    | .485 (.379, .680) | .484 (.374, .683) |
| BD-d   | -.746 (-.858, -.633) | -.736 (-.844, -.628) |
| dBD-d  | -.524 (-.763, -.284) | -.517 (-.749, -.284) |
| BD-d+  | -.753 (-.900, -.606) | -.739 (-.891, -.586) |
| dBD-d+ | -.678 (-.860, -.495) | -.663 (-.849, -.476) |

Table 6: Means and Variances of scores computed by different metrics on TED along with their t-scores.

| Metric | MTs | MTd | t   |
|--------|-----|-----|-----|
| BLEU   | 19.13 (4.16) | 21.62 (4.89) | **6.54** |
| METEOR | 27.23 (2.89) | 28.59 (2.96) | 2.81 |
| ROUGE-L| 42.19 (5.82) | 45.07 (5.48) | 4.81 |
| BlonD  | **30.07** (11.21) | **34.49** (5.46) | **8.94** |
| dBD    | 51.87 (6.28) | 55.12 (6.10) | 7.31 |
| LC     | 57.77 (6.23) | 58.64 (6.34) | **0.75** |
| RC     | 57.74 (6.23) | 58.61 (6.35) | 0.30 |
| TER    | 62.56 (8.88) | 68.97 (9.19) | **-0.83** |
| BD-d   | 81.16 (5.07) | 79.13 (5.46) | **-6.54** |
| dBD-d  | 81.09 (7.75) | 78.76 (8.01) | -3.67 |

5.4 Evaluation on German Translation Task

Blond can be used not only to evaluate the translation quality of English documents. In any language, as long as the NER and tagging markup tools are available in the target language, the Blond score can be calculated. In order to evaluate the generalization ability of Blond, we also conduct experiments on the TED English-to-German dataset. MTs and MTd have the same definition as above. The POS-tag set $V$ is \{\text{VMFIN}, \text{VMINF}, \text{VMPP}, \text{VVFIN}, \text{VVIMP}, \text{VVIZ-}\text{U}, \text{VVPP}\} and the pronoun set $P$ is \{\text{er, sie, es, man}\}.

As shown in Table 6, the difference between the Blond scores of MTs and MTd is more statistically significant than their BLEU scores, which is consistent with the results in the Chinese-to-English translation task. More detailed results on TED can be found in the Appendix.

6 Related Work

MT outputs are almost always evaluated using standard sentence-level metrics like BLEU and METEOR (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017; Maruf and Haffari, 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019a; Junczys-Dowmunt, 2019). However, errors related to discourse phenomena remain invisible in a sentence-level evaluation (Maruf et al., 2019b; Popescu-Belis, 2019).

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). Jwalapuram et al. (2019) also proposed a specialized measure for pronoun evaluation which involves training. Besides, Wong and Kit (2012) proposed LC and RC to measure lexical cohesion. Hajlaoui and Popescu-Belis (2013) proposed to assessing the accuracy of connective translation (ACT). Comelles et al. (2010) and Joty et al. (2014) exploited the discourse structure to provide more informed evaluation scores.

Unlike previous work, Blond does not focus on one specific phenomenon but has comprehensive consideration of the overall document-level quality.

7 Conclusion

In this paper, we describe Blond, an automatic metric for document-level evaluation. During the development of Blond, we build a large-scale document-level Chinese-English parallel dataset BWB. In addition, we also propose a new method to diagnose discourse errors in MT translations, and to identify the source of improvement of MT systems (dBlond). We also propose a Blond extension scheme (Blond+) that is easy to incorporate manual annotations.
References

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, 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.

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.

Elisabet Comelles, Jesús Giménez, Lluís Márquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.

Gabriel Forgues and Joelle Pineau. 2014. Bootstrapping dialog systems with word embeddings.

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).

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.

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.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? arXiv preprint arXiv:1704.05135.

Shafiq Joty, Francisco Guzmán, Lluís Márquez, 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.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1, pages 225–233.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. arXiv preprint arXiv:1909.00131.
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.

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.

Sharid Loáciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 Discourse MT shared task on cross-lingual pronoun prediction. In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 1–16, Copenhagen, Denmark. Association for Computational Linguistics.

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

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019a. Selective attention for context-aware neural machine translation. 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019b. A survey on document-level machine translation: Methods and evaluation. ArXiv, abs/1912.08494.

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

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

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

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

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. CoRR, abs/1901.09115.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, volume 200.

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

Natasha Tokowicz and Tamar Degani. 2010. Translation ambiguity, pages 281–294.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 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.

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

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

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

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 COLING.