Context-aware Decoder for Neural Machine Translation using a Target-side Document-Level Language Model

Amane Sugiyama
The University of Tokyo*
sugi@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

Although many end-to-end context-aware neural machine translation models have been proposed to incorporate inter-sentential contexts in translation, these models can be trained only in domains where parallel documents with sentential alignments exist. We therefore present a simple method to perform context-aware decoding with any pre-trained sentence-level translation model by using a document-level language model. Our context-aware decoder is built upon sentence-level parallel data and target-side document-level monolingual data. From a theoretical viewpoint, our core contribution is the novel representation of contextual information using point-wise mutual information between context and the current sentence. We demonstrate the effectiveness of our method on English to Russian translation, by evaluating with BLEU and contrastive tests for context-aware translation.

1 Introduction

Neural machine translation (NMT) has typically been explored in sentence-level translation settings. Such sentence-level NMT models inevitably suffer from ambiguities when a source sentence has multiple plausible interpretations. Examples of such ambiguities include anaphora, ellipsis, and lexical coherence (Voita et al., 2019b); although resolving these ambiguities has only a minor impact on the translation performance measured by BLEU scores (Papineni et al., 2002), they are vital in smoothly reading the translated documents.

To address this issue, context-aware NMT models which incorporate document-level information in translation have recently been explored (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Voita et al., 2018; Bawden et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Voita et al., 2019b; Yu et al., 2020; Kang et al., 2020; Zhang et al., 2020). Most of these models are end-to-end models that require document-level parallel data with sentential alignments for training. However, this data is available in only a few domains (Sugiyama and Yoshinaga, 2019). Researchers have therefore started to utilize target-side monolingual data to construct auxiliary models which help a sentence-level NMT model perform context-aware translation (Voita et al., 2019a; Stahlberg et al., 2019; Yu et al., 2020).

In this study, we propose a simple yet effective approach to context-aware NMT using two primitive components, a sentence-level NMT model and a document-level language model (LM). We can independently train the two components on common sentence-level parallel data and document-level monolingual data, respectively, without using document-level parallel data. Our approach thereby makes it possible to perform context-aware translation with any pre-trained sentence-level NMT model, using a pre-trained document-level LM.

To give a probabilistic foundation to this combination of two independent models, we exploit the probabilistic nature of NMT decoding. When generating a sequence, a left-to-right decoder outputs a categorical probability distribution over the vocabulary at every time step. The decoder assigns higher probabilities to the tokens that would be more suitable at that step. Therefore, when multiple valid translations are possible for the source sentence, the decoder just gives a higher probability to the translation that is plausible without considering contexts. We thus adjust the probability distributions in a context-aware manner using a target-side document-level LM which models inter-sentential dependencies in the target-side document.

We evaluate our methods on English to Russian translations with the OpenSubtitles2018 corpus (Lison et al., 2018) in terms of the BLEU scores and contrastive discourse test sets (Voita et al., 2019b). Experimental results confirm that

*Currently at Mitsubishi UFJ Morgan Stanley Securities

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5781–5791
June 6–11, 2021.

our method achieved comparable performance with existing context-aware NMT models that require either document-level parallel data (Zhang et al., 2018; Sugiyama and Yoshinaga, 2019) or more than one additional model (Voita et al., 2019a; Yu et al., 2020) for capturing contexts in translation.

The contributions of this paper are as follows:

- We theoretically derived C-SCORE, a score to **qualify context-aware translation without the need for document-level parallel data**.
- Two formulations with C-SCORE **turn any pre-trained sentence-level NMT model into a context-aware model**, if it generates n-best outputs or performs left-to-right decoding.
- A comparison between our approach and shallow fusion (Gulcehre et al., 2015) reveals that our approach **reformulates shallow fusion while adding a probabilistic foundation**.

2 Context-aware Decoding using Document-level Language Model

In this section, assuming a sentence-level encoder-decoder model (Bahdanau et al., 2015; Vaswani et al., 2017), we first derive *context-aware score* (C-SCORE for short), a context-aware objective function of outputs to be maximized in decoding. We then describe how to compute the C-SCORE using the decoder with a document-level language model (D-LM) (§ 2.1). We finally detail how to perform context-aware decoding based on C-SCORE (§ 2.2).

2.1 C-SCORE: objective function for context-aware NMT decoding

Let us consider the problem of finding a translation $\boldsymbol{y}$ of a source sentence $\boldsymbol{x}$ in a document. The target-side context sentence(s) preceding $\boldsymbol{y}$, $\boldsymbol{c}^{(\boldsymbol{y})}$, are to be given by the past translations. We formulate context-aware translation conditioned on $\boldsymbol{c}^{(\boldsymbol{y})}$ as the maximization of the conditional probability $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})})$,

$$\begin{aligned}\hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} \log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})}) \\ &= \arg\max_{\boldsymbol{y}} \log \frac{p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{x})} \\ &= \arg\max_{\boldsymbol{y}} \log p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{x}). \quad (1)\end{aligned}$$

Assuming that $\boldsymbol{x}$ and $\boldsymbol{y}$ are semantically similar, we make the following approximation,

$$p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{y}, \boldsymbol{x}) \approx p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{y}). \quad (2)$$

From Eq. 1 and Eq. 2, we obtain

$$\begin{aligned}\hat{\boldsymbol{y}} &\approx \arg\max_{\boldsymbol{y}} \log p(\boldsymbol{c}^{(\boldsymbol{y})}|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{x}) \\ &= \arg\max_{y} \log \frac{p(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})}{p(\boldsymbol{c}^{(\boldsymbol{y})})p(\boldsymbol{y})} p(\boldsymbol{y}|\boldsymbol{x}) \\ &= \arg\max_{\boldsymbol{y}} \text{C-SCORE}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})})\end{aligned}$$

where

$$\text{C-SCORE}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})}) = \log p(\boldsymbol{y}|\boldsymbol{x}) + \text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y}) \quad (3)$$

$$\text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y}) = \log \frac{p(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})}{p(\boldsymbol{c}^{(\boldsymbol{y})})p(\boldsymbol{y})} = \log \frac{p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})}{p(\boldsymbol{y})} \quad (4)$$

$\text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})$ is the point-wise mutual information of $\boldsymbol{c}^{(\boldsymbol{y})}$ and $\boldsymbol{y}$ which represents the degree of co-occurrence of $\boldsymbol{y}$ and $\boldsymbol{c}^{(\boldsymbol{y})}$. Given $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{c}^{(\boldsymbol{y})}$, we can evaluate the C-SCORE by computing the two terms in Eq. 3 using a sentence-level NMT (S-NMT) and a document-level LM (D-LM), respectively.

Notations We first introduce some notation to explain the computation in Eq. 3 and Eq. 4 using (auto-regressive) neural sequence generation models in NMT and LM. For a sequence $\boldsymbol{s}$ ($|\boldsymbol{s}| \geq 0$) and token w, a neural sequence generation model parameterized by θ can compute the log probability that w follows $\boldsymbol{s}$, which we denote by $\log p_\theta(w|\boldsymbol{s})$):

$$\log p_\theta(w \text{ folows } \boldsymbol{s}) = \log \frac{p_\theta(\boldsymbol{s} \cdot w)}{p_\theta(\boldsymbol{s})} = \log p_\theta(w|\boldsymbol{s})$$

where "$\cdot$" denotes sequence concatenation. Applying this auto-regressively, for any sequence $\boldsymbol{s}^{(1)}$ ($|\boldsymbol{s}^{(1)}| \geq 0$) and $\boldsymbol{s}^{(2)}$ ($|\boldsymbol{s}^{(2)}| \geq 1$), the probability that $\boldsymbol{s}^{(2)}$ follows $\boldsymbol{s}^{(1)}$ is thereby computed as:

$$\begin{aligned}&\log p_\theta(\boldsymbol{s}^{(2)} \text{ follows } \boldsymbol{s}^{(1)}) \\ &= \log p_\theta(\boldsymbol{s}^{(2)}|\boldsymbol{s}^{(1)}) = \sum_{t=1}^{|\boldsymbol{s}^{(2)}|} \log p_\theta(\boldsymbol{s}^{(2)}_t|\boldsymbol{s}^{(1)} \cdot \boldsymbol{s}^{(2)}_{<t}),\end{aligned}$$

where $\boldsymbol{s}^{(2)}_{<t} = [s_1, \ldots, s_{t-1}]$. (5)

$p(\boldsymbol{y}|\boldsymbol{x})$ computed by sentence-level NMT Computing $\log p(\boldsymbol{y}|\boldsymbol{x})$ using an S-NMT is straightforward. Suppose $\boldsymbol{y}$ to be a sequence of raw tokens, $\boldsymbol{y} = [y_1, \ldots, y_T]$. Then $\log p(\boldsymbol{y}|\boldsymbol{x})$ is computed by

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \log p_{\text{S-NMT}}(\tilde{\boldsymbol{y}}; \boldsymbol{x}) \quad (6)$$

where $\tilde{\boldsymbol{y}} = [y_1, \ldots, y_T, \texttt{</s>}]$ and </s> is a special token to indicate the end of sentence.

PMI computed by document-level LM To compute the components of $\text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})$, $p(\boldsymbol{y})$ and $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$, we use a document-level language model (D-LM) which can handle long text spans containing multiple sentences.

We generate training examples for D-LM from a document as follows. We assume D-LM explicitly models sentence boundaries. We first insert the special token </s> into every sentence boundary including the start and end of the document. With this preprocessing, all the sentences start immediately after an </s> token and end immediately before an </s> token. We then sample text spans from the document using a sliding window, where the start and end of the span do not have to match sentence boundaries. The sliding window's size is larger than the stride size, so adjacent spans may overlap. The resulting sequence is fed to the D-LM for training. Note that </s> for D-LM indicates sentence boundaries, in other words, both the start and end of the sequence.

Using D-LM, $p(\boldsymbol{y})$ is computed by

$$p(\boldsymbol{y}) = p_{\text{D-LM}}(\tilde{\boldsymbol{y}}|\text{</s>}). \tag{7}$$

where $\tilde{\boldsymbol{y}} = [y_1, \ldots, y_T, \text{</s>}]$.

To compute $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$, we first obtain the context sequence $\tilde{\boldsymbol{c}}^{(\boldsymbol{y})}$ by concatenating all the sentences in $\boldsymbol{c}^{(\boldsymbol{y})}$ with </s>. We then compute the conditional probability $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$ by

$$p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})}) = p_{\text{D-LM}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{c}}^{(\boldsymbol{y})}) \tag{8}$$

where $\tilde{\boldsymbol{y}} = [y_1, \ldots, y_T, \text{</s>}]$.

Let us explain why we use the boundary-aware D-LM rather than boundary-agnostic D-LM.[1]

Firstly, boundary-agnostic LMs cannot compute the probability that a sentence is closed with a certain length, namely, Eq. 7 cannot be computed. Secondly, they also cannot compute $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$ correctly. For example, suppose the context $\boldsymbol{c}^{(\boldsymbol{y})}$ is "he's my friend" (with the punctuation "." omitted), and the current target sentence $\boldsymbol{y}$ is "he's nice." In this case, Eq. 8 is computed by

$$p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})}) = p_{\text{D-LM}}([\text{he,'s,nice}]|[\text{he,'s,my,friend}]).$$

However, this estimation of $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$ can underestimate the actual $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$ because Eq. 8 inevitably gives significant probabilities to other $\boldsymbol{y}$ such as "'s father" as well, since "He's my friend's father" is fluent as a sequence. This behavior is unsuitable for $\boldsymbol{y}$,[2] since "'s father" is not a complete sentence.

[1] We cannot rely on punctuations to know sentence boundaries, since they can be omitted in some domains.

2.2 Searching for the optimal solution

Searching for the optimal output $\boldsymbol{y}$ that maximizes the C-SCORE is not trivial since there are $\mathcal{O}(V^T)$ candidate sequences where V is the vocabulary size and T is the maximum length of sequences to be searched. We investigate two approaches to obtain approximate solutions: reranking (§ 2.2.1) and context-aware beam search (§ 2.2.2).

2.2.1 Reranking with C-SCORE

We first generate B hypotheses of the translation $\mathcal{H}_B = \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^B\}$ with beam search of beam size B using the sentence-level NMT model. We then choose the one that maximizes the C-SCORE.

$$\hat{y} = \arg\max_{\boldsymbol{y} \in \mathcal{H}_B} \text{C-SCORE}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})}) \tag{9}$$

An issue with reranking is that we need to set B to a large value when the diversity of models' outputs is limited (Yu et al., 2020), which increases the cost of decoding. We therefore attempt to integrate C-SCORE into the decoding with beam search.

2.2.2 Context-aware beam search

Context-aware beam search (C-AWARE beam) is beam search that is extended to work with C-SCORE. C-SCORE (Eq. 3) can be decomposed into token-wise C-SCOREs (Eq. 5 through Eq. 8).

$$\begin{aligned}\text{C-SCORE}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{c}^{(\boldsymbol{y})}) &= \log p(\boldsymbol{y}|\boldsymbol{x}) + \text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y}) \\ &= \sum_{t=1}^{T+1} \text{C-SCORE}_w(\tilde{y}_t|\tilde{\boldsymbol{y}}_{<t})\end{aligned} \tag{10}$$

where

$$\begin{aligned}\text{C-SCORE}_w(\tilde{y}_t|\tilde{\boldsymbol{y}}_{<t}) = &\log p_{\text{S-NMT}}(\tilde{y}_t|\tilde{\boldsymbol{y}}_{<t}; x) \\ &+ \log \frac{p_{\text{D-LM}}(\tilde{y}_t|\tilde{\boldsymbol{c}}^{(\boldsymbol{y})} \cdot \tilde{\boldsymbol{y}}_{<t})}{p_{\text{D-LM}}(\tilde{y}_t|\text{</s>} \cdot \tilde{\boldsymbol{y}}_{<t})}\end{aligned} \tag{11}$$

By this decomposition, C-SCORE_w is conditioned on the partial sequence generated by time step t. We can therefore apply beam search to generate sequences in an auto-regressive manner.

The first term of Eq. 11 represents the translation probability for the t-th token. The second term can

[2] Strictly speaking, we assume $\boldsymbol{y}$ to be a realization of a random variable Y which is a sentence sampled from the space of an infinitely large document.

be interpreted as PMI between the t-th token and the context, that is, how consistent the t-th token is with the context. Compared to the reranking approach, C-AWARE beam can be considered to maximize the C-SCORE more directly in the sense that disambiguation and token selection based on the context are performed at every step in beam search. Thus C-AWARE beam will more space-efficiently consider diverse hypotheses with the same beam size B than C-AWARE rerank.

2.2.3 Smoothing probabilities for PMI

In our preliminary experiments, we observe that the original C-AWARE beam significantly improves contrastive tests but deteriorates BLEU at the same time. By analyzing contextual PMI correlation between source and target texts, we find the PMI term in the C-SCORE sometimes takes an excessively large value against the translation probability term, which destroys the C-SCORE. This is understood intuitively by the fact that the calculation of PMI includes subtraction of log probability, and log probability may take a very small negative value to represent a probability close to zero.

To alleviate this problem, we adopt a smoothing method for probabilities. For simplicity, in this paper, we only present the temperature scaling (T-scaling, for short) (Guo et al., 2017). T-scaling replaces $p_{y=w}$ by

$$\bar{p}_{y=w} = \frac{p_{y=w}^{1/T}}{\sum_{w'} p_{y=w'}^{1/T}} \tag{12}$$

where T is a hyper-parameter. $T = 1$ is equivalent to no smoothing. We choose T from $[1, \infty)$ to flatten the probability distribution. T-scaling is applied to both the numerator and denominator using the same T.

2.2.4 On the relation to shallow fusion

Shallow fusion (Gulcehre et al., 2015) is a method to integrate probability distribution outputs obtained by NMT and LM at sentence level to form a new translation objective that is expected to promote fluency of translations. The original shallow fusion score is computed using a sentence-level NMT (S-NMT) and language model (S-LM). The token-wise formula of the computation is

$$\log p(y_t) = \log p_{\text{S-NMT}}(y_t; \boldsymbol{x}) + \beta \log p_{\text{S-LM}}(y_t), \tag{13}$$

where β is a hyper-parameter. In our notation with the document-level LM, this is written as

$$\log p(y_t) = \log p_{\text{S-NMT}}(\tilde{y}_t|\tilde{\boldsymbol{y}}_{<t}; \boldsymbol{x}) + \beta \log p_{\text{S-LM}}(\tilde{y}_t|\texttt{</s>} \cdot \tilde{\boldsymbol{y}}_{<t}). \tag{14}$$

A natural extension of this objective to the context-aware scenario should be

$$p(y_t|\boldsymbol{c}^{(\boldsymbol{y})}) = \log p_{\text{NMT}}(\tilde{y}_t|\tilde{\boldsymbol{y}}_{<t}; x) + \beta \log p_{\text{D-LM}}(\tilde{y}_t|\tilde{\boldsymbol{c}}^{(\boldsymbol{y})} \cdot \tilde{\boldsymbol{y}}_{<t}), \tag{15}$$

where context $\tilde{c}^{(\boldsymbol{y})}$ is integrated into the condition. We call this *conditional (document-level) shallow fusion*. Obviously, this is what we obtain from Eq. 11 by ignoring the discount of the unconditional LM probability $p_{\text{D-LM}}(\tilde{y}_t|\texttt{</s>} \cdot \tilde{y}_{<t})$.

Due to the absence of discounting with the unconditional LM, conditional shallow fusion would prefer tokens which frequently occur regardless of the context. It is also worth noting that, when the context is empty, conditional shallow fusion falls back to the original shallow fusion, whereas our C-SCORE falls back to sentence-level NMT. Therefore, we view C-SCORE as a reformulation of shallow fusion for context-aware translation.

3 Experimental Setup

We evaluate our methods on English to Russian translation, in terms of BLEU scores (Papineni et al., 2002) and contrastive tests (Voita et al., 2019b).

3.1 Datasets and preprocessing

We use the OpenSubtitles2018 corpus (Lison et al., 2018) for parallel and monolingual data. Following the criteria for document segmentation and filtering on sentence pairs presented by (Voita et al., 2019b), we build monolingual and parallel data as follows. To build monolingual data, we add document boundary information into each document such that they consist of contiguous subtitle sentences from the same movie and the timestamp difference of any two adjacent sentences is no more than seven seconds. To build parallel data, we pick subtitle pairs where the time overlap between the source and target language subtitles is at least 0.9 (to reduce alignment errors). For the training of multi-encoder NMT models, document boundary information is added to the parallel data based on the source-side timestamps as with the monolingual data. Prior to building the Russian data, we

	Train			Dev.			Test	
	src	trg	(mono)	src	trg	(mono)	src	trg
# sentences	5.8M		30M	6.0k		23k	15.5k	
avg. # tokens	9.9	9.4	8.5	10.1	9.6	8.9	9.8	9.1

Table 1: Statistics of the parallel and monolingual data.

remove the movies from which the contrastive test sets (§ 3.4) were made.

We perform punctuation normalization, tokenization, and truecasing on the source and target texts using Moses toolkit v4.0.[3] We then encode the texts into subwords using SentencePiece (v0.1.81)[4] with unigram LM. The subword vocabularies are of 16,000 tokens and trained for each language. The statistics of the datasets are listed in Table 1.

3.2 Models

We compare our methods to one sentence-level translation model (**SentTransformer**) (Vaswani et al., 2017) and three context-aware translation models: Document transformer (Zhang et al., 2018), DocRepair (Voita et al., 2019a), and Bayes Document Reranker (Yu et al., 2020). All the context-aware models use the previous three sentences as context.

Document Transformer (**DocTransformer**, for short) is a multi-encoder document-level NMT model which takes source-side context as an auxiliary input and can be thus trained from document-level parallel data. We follow (Zhang et al., 2018)'s configuration for DocTransformer.

DocRepair is a sequence-to-sequence post-editing model. It repairs document-level inconsistencies in a text, each sentence of which has been translated separately by a sentence-level NMT model. DocRepair is trained on a pseudo parallel data made by pairing a monolingual corpus and its round-trip translations obtained using a back-translation model and a forward-translation model.

Bayes Document Reranker (hereafter, **Bayes DocReranker**) performs document-level translation on a document containing D sentences in the following steps. First, it produces B-best translations for each sentence in the document and then produces a lattice of width B and depth D, where each node corresponds to a candidate sentence. It then performs document-level beam search of beam size B' on the lattice using the following score:

$$\begin{aligned}&\text{Score}(\boldsymbol{y}_i; \boldsymbol{y}_{<i}, \boldsymbol{x}_i) = \\ &\quad p_{\text{D-LM}}(\boldsymbol{y}_i|\boldsymbol{y}_{<i}) + \text{Score}(\boldsymbol{y}_{i-1}; \boldsymbol{y}_{<i-1}, \boldsymbol{x}_{i-1}) \\ &\quad + \lambda_1 p_{\text{NMT}}(\boldsymbol{y}_i|\boldsymbol{x}_i) + \lambda_2 p_{\text{BACK-NMT}}(\boldsymbol{x}_i|\boldsymbol{y}_i) + \lambda_3|\boldsymbol{y}_i|\end{aligned} \tag{16}$$

Note that this document-level beam search is equivalent to the reranking procedure (§ 2.2.1) when $B' = 1$. Therefore, the essential difference between Bayes DocReranker and our C-SCORE reranking is the score function.

SentTransformer, the post-editing model of DocRepair, and the back-translation models are based on the same configuration of Transformer base (see (Vaswani et al., 2017) for hyperparameter settings). The SentTransformer is trained using the 5.8M sentence pairs and is also used as the sentence-level NMT model in DocRepair, Bayes DocReranker, and our methods. For the training of DocTransformer, we use the 5.8M sentence pairs with document-level source context, which share the target-side sentences with the training data of SentTransformer. Consequently, scores obtained from the model are for reference.[5] We also evaluate DocTransformer and SentTransformer using back-translation (BT) (Sennrich et al., 2016) with the same monolingual data as the other models.

We use no pre-existing document-level parallel data to train the neural networks of DocRepair, Bayes DocReranker, and our methods, although we use a small amount of document-level parallel data as the development set to tune hyperparameters in the methods that combine multiple models. Instead, document-level information is fed to the models via the round-trip augmented data (DocRepair) or language models (Bayes DocReranker and our methods).

Hyper-parameters We tune the models' hyper-parameters based on BLEU score on the development set in the evaluation with BLEU, while we tune these hyper-parameters in the evaluation of contrastive tests by maximizing the coefficient of D-LM under the constraint that it does not deteriorate BLEU compared to the SentTransformer.

For beam search to produce B-best outputs in Bayes DocReranker and our C-AWARE Rerank, we

[3] http://www.statmt.org/moses/

[4] https://github.com/google/sentencepiece

[5] Although we can train DocTransformer only on pseudo document-level parallel data generated by back-translation, we confirmed in preliminary experiments that the resulting model exhibited poor performance.

Models	para only	monolingual data 6M	15M	30M
SentTransformer (w/ BT)	32.36	32.32	32.40	32.40
Shallow Fusion	n/a	32.39	32.56	32.52
baselines				
DocTransformer (w/ BT)	32.50	32.36	31.88	31.59
DocRepair	n/a	32.13	32.36	32.35
Bayes DocReranker	n/a	**32.80***	**33.58****	**33.75****
w/o context	n/a	32.53	33.44**	33.67**
proposed				
C-AWARE Rerank	n/a	32.74*	33.01**	32.93*
C-AWARE Beam	n/a	32.26	32.28	32.27
Cond. Shallow Fusion	n/a	32.38	32.55	32.55

Table 2: Test set BLEU scores. '*' and '**' indicate that gains from SentTransformer in the same column are statistically significant ($p < 0.05$ and $p < 0.01$) by bootstrap resampling with 1000 samples, respectively.

Models		deixis	lex.c	ell.infl	ell.vp
SentTransformer		50.0	45.9	53.2	27.0
w/ BT		50.0	45.9	51.6	26.8
baselines					
Doc-Transformer		50.0	45.9	56.0	57.2
w/ BT		50.0	45.9	64.4	68.2
DocRepair		**89.1**	75.8	**82.2**	67.2
Bayes DocReranker		65.2	72.2	59.6	44.6
proposed					
C-SCORE		86.9	**94.9**	78.2	**77.0**
Cond. Shallow Fusion		54.7	55.3	53.4	32.4
D-LM	PMI($\boldsymbol{c}^{(\boldsymbol{y})}, y$)	**96.8**	**97.8**	75.8	**90.6**
	$p(\boldsymbol{y}\|\boldsymbol{c}^{(\boldsymbol{y})})$	89.7	95.7	**77.4**	81.6

Table 3: Results on contrastive test sets.

use a beam size of $B = 20$. For document-level beam search of Bayes DocReranker, we use a beam size $B' = 5$. For beam search of SentTransformer, DocTransformer, C-AWARE beam, and shallow fusion, we use a beam size of $B = 4$.

3.3 Document-level Language models

The architecture of the document-level LM is the decoder part of a Transformer. The number of decoder blocks is 12. The model size is 768 with 12 attention heads, and the inner layer of the feed-forward networks has 3072 units. We use position embeddings to represent position information.

As described in § 2.1, when training the language models, a special control symbol </s> is inserted at every sentence boundary. Each training mini-batch contains text spans each of which is a randomly sampled fragment of a document with a maximum span length of $W = 384$. Text spans are batched such that about 32,000 tokens are in a training batch.

3.4 Evaluation methods

The existing automatic metrics are not adequate to evaluate gains from additional contexts (Bawden et al., 2018; Läubli et al., 2018; Müller et al., 2018; Voita et al., 2019b; Sugiyama and Yoshinaga, 2019). We thus adopt a contrastive test set (Voita et al., 2019b) to evaluate the model's ability to capture contextual information in translation, in addition to the evaluation by BLEU scores (Papineni et al., 2002) to confirm that the methods do not sacrifice general translation performance. BLEU is computed using `multi-bleu.perl` from the Moses Toolkit after decoding the subword representation of the models' outputs into words using SentencePiece.

The contrastive test set consists of contrastive questions for context-aware NMT models to answer. Each question has a source sentence $\boldsymbol{x}$, a source context $\boldsymbol{c}^{(\boldsymbol{x})}$, a target context $\boldsymbol{c}^{(\boldsymbol{y})}$, and translation candidates $\mathcal{Y} = \{y^1, \ldots, y^M\}$. Models must answer with a candidate $\hat{y} \in \mathcal{Y}$ which would be the most appropriate translation of $\boldsymbol{x}$, i.e.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(\boldsymbol{y})})$$

The test sets consist of 6000 examples in total.

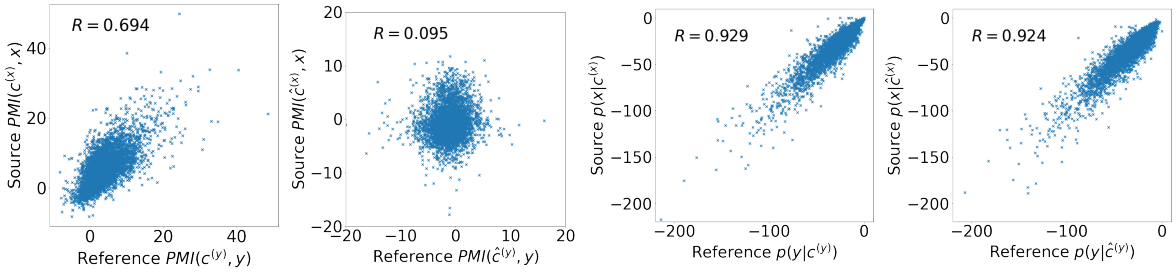
4 Results and Analysis

4.1 General translation performance measured by BLEU scores

Table 2 lists the performance of the models in terms of BLEU scores. Bayes DocReranker and our C-AWARE Rerank consistently outperformed the baseline SentTransformer, even when it used data augmentation by back-translation, while the other methods are just comparable to the baseline. Althogh Bayes DocReranker performed the best among all the models, the comparison to Bayes DocReranker without context information (using $p_{\text{S-LM}}(\boldsymbol{y}_i)$ instead of $p_{\text{D-LM}}(\boldsymbol{y}_i|\boldsymbol{y}_{<i})$) reveals that most of the improvement is not obtained by the use of contexts. Back-translation did not contribute to BLEU possibly because the original parallel data is already large and there was little room for improvement with additional pseudo data.

4.2 Results on contrastive test sets

Tables 3 lists evaluation results (accuracy) of the contrastive tests with models using 30M monolingual data. The highest scores on each column



(a) PMI (b) PMI with rand. context (c) Cond. prob. (d) Cond. prob. (rand. context)

Figure 1: Source-target correlation of contextual PMI (a, b) and conditional probability (c, d), calculated based on the correct context (a, c) and wrong context that is randomly chosen from the dataset (b, d). The dataset is a subset of the training data from the English-Russian parallel corpus. Plots are for 4166 sentence pairs in the dataset.

are in bold, and additionally, the higher one of the two D-LM-based scores is shown in bold. The contrastive test include four test sets: *deixis* is for person deixis, *lex.c* is for lexical cohesion, *ell.infl* is for inflection of Russian nouns caused by ellipsis in the source sentence, and *ell.vp* is for verb ellipsis in English text which is not allowed in Russian. Although the contrastive test is targeted at context-aware NMT models, it is possible to answer the contrastive questions by $\arg\max_y \text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, y)$ or $\arg\max_y p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$. Scores obtained by these two objectives are also reported in the table in addition to the scores obtained by SentTransformer.

Our C-SCORE outperforms all the context-aware models other than DocRepair. The performance of C-SCORE is slightly worse than DocRepair for *deixis* (2.2 points) and *ell.infl* (4.0 points), while achieving large improvements for *lex.c* (19.1 points) and *ell.vp* (9.8 points) over DocRepair.

D-LM *only* objectives achieve higher scores than C-SCORE, except for *ell.infl*. This is not surprising because the choices in the tests are guaranteed to be valid as translation for the source sentences if given some appropriate context, so the questions can be solved without translation. This result still indicates that the D-LM scores give good hints for tackling contextual ambiguities. The advantage of C-SCORE over the SentTransformer is demonstrated by the excellent performance of D-LM in capturing contexts in translation.

4.3 On translation efficiency

The inference speed depends mainly on the model size and beam size. In our experiments on a single TITAN Xp GPU, SentTransformer decoded the fastest at 66 sents/sec, followed by DocTransformer that ran in 40 sents/sec. DocRepair ran in about 28 sents/sec, slightly slower because it decodes in two passes. C-AWARE Rerank and Bayes DocReranker were about 4.3 sents/sec and 7.7 sents/sec respectively. We expect that these models would be accelerated by using a language model with a better cache mechanism (e.g. TransformerXL (Dai et al., 2019)). C-AWARE Beam ran in about 13 sents/sec.[6] We leave thorough analysis on speed/performance trade-offs to future work.

4.4 PMI correlation analysis

In § 4.2 we have confirmed the effectiveness of PMI as a measure of a valid translation given context using contrastive tests. To gain a deeper insight into how well PMI conveys semantic connections between the current sentence and its context, we analyze the correlation of PMI between source and target sentences.

PMI correlation between source and target

The main result we show in this section is that the PMI of the source and target correlate well. This is important because this supports the idea that PMI is a language-independent measure of the connection between the current sentence and its context.

Although we have discussed only target-side $\text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})$ defined by Eq. 4, we can compute the source-side $\text{PMI}(\boldsymbol{c}^{(\boldsymbol{x})}, \boldsymbol{x})$ in the same way. Given a document-level parallel corpus, we measure a correlation between $\text{PMI}(\boldsymbol{c}^{(\boldsymbol{x})}, \boldsymbol{x})$ and $\text{PMI}(\boldsymbol{c}^{(\boldsymbol{y})}, \boldsymbol{y})$ for each sentence pair $(\boldsymbol{x}, \boldsymbol{y})$ in the corpus.

Figure 1a shows the PMI correlation for about

[6] Note that the running time of NMT decoding also depends on the degree of parallelism, and for C-AWARE Beam, decoding multiple sentences in parallel is less trivial since it demands that all the previous sentences in the document are translated by the time it starts to translate the current one. In our experiments, assuming a practical scenario where a large number of users input their documents for translation, we translate multiple documents in parallel so that multiple sentences from different documents can be translated in parallel.

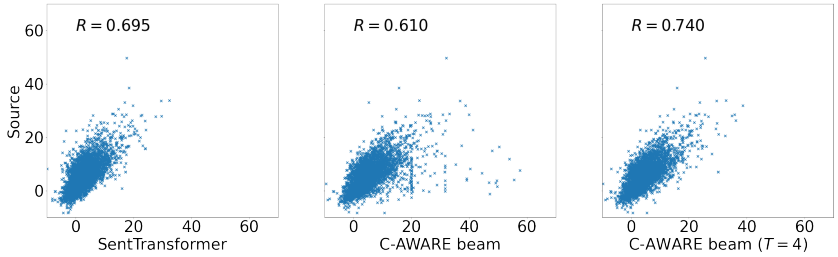


Figure 2: Correlation of contextual PMI between the source sentences (from the training data) and the outputs of some models (SentTransformer, C-AWARE beam without T-scaling, and C-AWARE beam with T-scaling of $T = 4$).

4000 sentence pairs taken from the dev data. The pairs of PMI values are computed using English and Russian language models trained on the training data. We observe a clear correlation between source and target, which agrees with the intuition that if the target sentence matches well in the context, so does the source sentence. What is also obvious in Figure 1a is that most of the points lay in the first quadrant where both the source and target contextual PMI is greater than 0, which is explained by the simple intuition that most sentences should have positive co-occurrence relation with their contexts. This behavior is lost when computing the contextual PMI using an incorrect context $\tilde{c}$ randomly chosen in the dataset as shown in Figure 1b.

The effectiveness of PMI as a measure of the valid translation of the current sentence given context is further emphasized when compared to the conditional probability $p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})})$, which could be an alternative measure of how suitable $\boldsymbol{y}$ is in the context as described in § 2.2.4. Figure 1c and 1d are the conditional probability version of Figure 1a and 1b: $(p(\boldsymbol{x}|\boldsymbol{c}^{(\boldsymbol{x})}), p(\boldsymbol{y}|\boldsymbol{c}^{(\boldsymbol{y})}))$ for each sentence pair (x, y) in the same dataset are plotted in Figure 1c and the same tuples but with random contexts are plotted in Figure 1d. Unlike the contextual PMI correlation, conditional probability correlation remains high even when we give wrong contexts. This is because the conditional probability of a sentence is highly affected by how frequently the sentence is observed regardless of context; if the source sentence is written with common expressions, then so is the target sentence and they are likely to be observed regardless of the context.

Analysis of the model outputs

PMI correlation gives us a good explanation of how C-AWARE beam without T-scaling fails. We plot the PMI correlation between the source sentences and their translations obtained with NMT models (Figure 2). We can find some outliers in the bottom right area of the plot for C-AWARE beam without T-scaling, which is the cause of the low correlation coefficient $R = 0.610 < R_{\text{src}-\text{ref}} = 0.695$. This result suggests that C-AWARE beam without T-scaling chooses some tokens based on excessively high token-wise PMI, which breaks some translations resulting in the low BLEU. Translation of the SentTransformer shows a higher correlation with the source texts than the reference translation (Figure 1a). One possible explanation for this is alignment errors in the corpus: although worse than the reference translations in quality, outputs of SentTransformer are considered to be perfectly aligned to the source sentences. C-AWARE beam with T-scaling ($T = 4$) seems to solve this issue and achieves the highest PMI correlation $R = 0.740$.

5 Related Work

The effectiveness of incorporating context into translation was shown in earlier literature on document-level NMT (Tiedemann and Scherrer, 2017; Bawden et al., 2018) using the single encoder architecture. Multi-encoder architectures were explored to better capture contextual information (Wang et al., 2017; Tu et al., 2018; Jean et al., 2017; Miculicich et al., 2018; Voita et al., 2018; Bawden et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019; Kang et al., 2020; Zhang et al., 2020). However, since parallel data is often constructed by picking up reliable sentential alignments from comparable documents, document-level sentence-aligned parallel data for training these document-level NMT models are expensive to obtain and available in only a few domains and language pairs (Sugiyama and Yoshinaga, 2019).

Recent studies have therefore started to focus on modeling contexts using document-level monolingual data. The current approaches are grouped into three categories: data augmentation via back-

translation (Sugiyama and Yoshinaga, 2019), a post-editing model (Voita et al., 2019a), and modeling document-level fluency via document-level LMs (Stahlberg et al., 2019; Yu et al., 2020; Jean and Cho, 2020). In what follows, we review these approaches in detail.

Sugiyama and Yoshinaga (2019) reported that the data augmentation by back-translation (Sennrich et al., 2016) enhances a document-level NMT model with a single encoder architecture in low-resource settings. However, we have obtained limited improvements in our settings (Table 2 and Table 3). Moreover, this approach is expensive since it learns a document-level NMT model from a massive amount of pseudo parallel data.

Voita et al. (2019a) proposed DocRepair, a context-aware post-editing model that corrects outputs of a sentence-level NMT model. Because DocRepair ignores the confidence of the first-stage sentence-level translation and possible alternative translations, it can miscorrect outputs of the sentence-level NMT model when they are irregular but correct. Moreover, when we change the target sentence-level NMT model, the accompanying post-editing model must be trained from its outputs. Our approaches, on the other hand, attempt a more "soft" revision, taking into account the output probabilities, i.e., confidence of the sentence-level NMT, and can perform context-aware decoding with any sentence-level NMT model, reusing a pre-trained document-level LM.

Stahlberg et al. (2019) and Yu et al. (2020) utilize a document-level LM to model document-level fluency of outputs; these approaches are similar to shallow fusion (Gulcehre et al., 2015)[7] with document-level LM (§ 2.2.4), although they perform a document-level reranking of translation hypotheses generated for individual source sentences by using sentence-level NMT. In particular, Yu's formulation has a probabilistic foundation like our approaches, and additionally utilizes a backward translation model. Although their formulation brings a significant improvement in BLEU (Table 2), the score is not obtained by better document-level translation; the comparable BLEU score of the no-context version of the method (Table 2) and the results of the contrastive tests (Table 3) reveal that the improvement is mostly due to the context-agnostic language model prior and the backward translation model. As we have discussed in § 2.2.4, document-level LM scores prefer tokens which frequently appear regardless of context and are unlikely to lead to better document-level translation. Moreover, their method requires training a back-translation model corresponding to the target sentence-level NMT model.

Finally, we noticed that Jean and Cho (2020) (which appeared after the preprint version of this paper (Sugiyama and Yoshinaga, 2020)[8] had been submitted) have reached a formulation that is very similar to the one presented in this paper by reformulating a noisy channel model of Bayes DocReranker (Yu et al., 2020). Concrete differences between our work and theirs include the fact that we conducted thorough analysis on the performance of different decoding strategies (not only beam search but also reranking). We also interpreted the subtraction of LM scores as point-wise mutual information and analyzed it by observing PMI correlation between source and target PMI to deepen the understanding of the formulation.

6 Conclusions

We present an approach to context-aware NMT based on PMI between the context and the current sentence. We first provide the formulation of the objective, C-SCORE, and the computation process of the C-SCORE using a sentence-level translation model and a document-level language model. We investigate two search methods, reranking and beam search, and evaluate the methods for English-Russian translation. We also provide some analysis and visualization to better understand the nature of PMI between the context and the current sentence.

We plan to design context-aware BLEU using PMI for evaluating context-aware NMT models. We will evaluate our method on non-autoregressive NMT (Gu et al., 2017). We will release all code and data to promote the reproducibility of results.[9]

[7] Our work is also related to shallow fusion (Gulcehre et al., 2015), in which token-wise probabilities output by an NMT model and a sentence-level LM are combined to be used as translation scores in decoding. The theoretical background of shallow fusion and our C-SCORE are different: in shallow fusion, the LM is intended to promote fluency of translations, whereas in our C-SCORE, we use the probability ratio of two LM probabilities which only provides contextual difference and fluency is still left to the translation model.

[8] This preprint is submitted to and rejected from EMNLP 2020; the interested reader may refer to this paper for experiments on other language pairs such as English to French and English to Japanese translation.

[9] http://www.tkl.iis.u-tokyo.ac.jp/~sugi/NAACL2021/

Acknowledgements

We thank anonymous reviewers for their valuable comments. We also thank Joshua Tanner for proofreading this paper. We also thank Masato Neishi for technical advice on implementations of neural machine translation. The research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the third International Conference on Learning Representations (ICLR)*.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. In *Proceedings of the the fifth International Conference for Learning Representations (ICLR)*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *Computing Research Repository*, arXiv:1503.03535.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Sébastien Jean and Kyunghyun Cho. 2020. Log-linear reformulation of the noisy channel model for document-level neural machine translation. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 95–101, Online.

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

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online.

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.

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).

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.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. 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.

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.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. CUED@WMT19:EWC&LMs. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 364–373, Florence, Italy.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China.

Amane Sugiyama and Naoki Yoshinaga. 2020. Context-aware decoder for neural machine translation using a target-side document-level language model. *Computing Research Repository*, arXiv:2010.12827.

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.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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, page 6000–6010. Curran Associates, Inc.

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

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

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.

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.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

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.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online.