Capturing document context inside sentence-level neural machine translation models with self-training

Elman Mansimov
New York University
mansimov@cs.nyu.edu

Gábor Melis
DeepMind
melisgl@google.com

Lei Yu
DeepMind
leiyu@google.com

Abstract

Neural machine translation (NMT) has arguably achieved human level parity when trained and evaluated at the sentence-level. Document-level neural machine translation has received less attention and lags behind its sentence-level counterpart. The majority of the proposed document-level approaches investigate ways of conditioning the model on several source or target sentences to capture document context. These approaches require training a specialized NMT model from scratch on parallel document-level corpora. We propose an approach that doesn't require training a specialized model on parallel document-level corpora and is applied to a trained sentence-level NMT model at decoding time. We process the document from left to right multiple times and self-train the sentence-level model on pairs of source sentences and generated translations. Our approach reinforces the choices made by the model, thus making it more likely that the same choices will be made in other sentences in the document. We evaluate our approach on three document-level datasets: NIST Chinese-English, WMT19 Chinese-English and Open-Subtitles English-Russian. We demonstrate that our approach has higher BLEU score and higher human preference than the baseline. Qualitative analysis of our approach shows that choices made by model are consistent across the document.

1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014) has achieved great success, arguably reaching the levels of human parity (Hassan et al., 2018) on Chinese to English news translation that led to its popularity and adoption in academia and industry. These models are predominantly trained and evaluated on sentence-level parallel corpora. Document-level machine translation that requires capturing the context to accurately translate sentences has been recently gaining more popularity and was selected as one of the main tasks in the premier machine translation conference WMT19 (Barrault et al., 2019) and WMT20 (Barrault et al., 2020).

A straightforward solution to translate documents by translating sentences in isolation leads to inconsistent but syntactically valid text. The inconsistency is the result of the model not being able to resolve ambiguity with consistent choices across the document. For example, the recent NMT system that achieved human parity (Hassan et al., 2018) inconsistently used three different names "Twitter Move Car", "WeChat mobile", "WeChat move" when referring to the same entity (Sennrich, 2018).

To tackle this issue, the majority of the previous approaches (Jean et al., 2017; Wang et al., 2017; Kuang et al., 2017; Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Agrawal et al., 2018; Zhang et al., 2018; Xiong et al., 2018; Miculicich et al., 2018; Voita et al., 2019a,b; Jean et al., 2019; Junczys-Dowmunt, 2019) proposed context-conditional NMT models trained on document-level data. However, none of the previous approaches are able to exploit trained NMT models on sentence-level parallel corpora and require training specialized context-conditional NMT models for document-level machine translation.

We propose a way of incorporating context into a trained sentence-level neural machine translation model at decoding time. We process each document monotonically from left to right one sentence at a time and self-train the sentence-level NMT model on its own generated translation. This procedure reinforces choices made by the model and hence increases the chance of making the same choices in the remaining sentences in the document. Our approach does not require training a separate context-conditional model on parallel document-

Algorithm 1: Document-level NMT with self-training at decoding time

> **Input:** Document $\boldsymbol{D} = (X_1, ..., X_n)$, pretrained sentence-level NMT model $f(\theta)$, learning rate α, decay prior λ and number of passes over document P
> **Output:** Translated sentences $(Y_1, ..., Y_n)$
> Backup original values of parameters $\tilde{\theta} \leftarrow \theta$
> **for** $p = 1$ **to** P {*multi-pass document*} **do**
>> **for** $i = 1$ **to** n **do**
>>> Translate sentence X_i using sentence-level model $f(\theta)$ into target sentence Y_i.
>>> Calculate cross-entropy loss $L(X_i, Y_i)$ using Y_i as target.
>>> **for** $j = 1$ **to** m **do**
>>>> $\theta \leftarrow \theta - \alpha \nabla_\theta L(X_i, Y_i) + \lambda(\tilde{\theta} - \theta)$
>>> **end for**
>> **end for**
> **end for**

level data and allows us to capture context in documents using a trained sentence-level model.

We make the key contribution in the paper by introducing the document-level neural machine translation approach that does not require training a context-conditional model on document data and does not require separate document-level language model to rank the outputs of the NMT model according to consistency of translated document. We show how to adapt a trained sentence-level neural machine translation model to capture context in the document during decoding. We evaluate and demonstrate improvements of our proposed approach measured by BLEU score and preferences of human annotators on several document-level machine translation tasks including NIST Chinese-English, WMT19 Chinese-English and OpenSubtitles English-Russian datasets. We qualitatively analyze the decoded sentences produced using our approach and show that they indeed capture the context.

2 Proposed Approach

We translate a document $\boldsymbol{D}$ consisting of n source sentences $X_1, X_2, ..., X_n$ into the target language, given a well-trained sentence-level neural machine translation model f_θ. The sentence-level model parametrizes a conditional distribution $p(Y|X) = \prod_{i=1}^{T} p(y_t|Y_{<t}, X)$ of each target word y_t given the preceding words $Y_{<t}$ and the source

sentence X. Decoding is done by approximately finding $\arg\max_Y p(Y|X)$ using greedy decoding or beam-search. f is typically a recurrent neural network with attention (Bahdanau et al., 2014) or a Transformer model (Vaswani et al., 2017) with parameters θ.

2.1 Self-training

We start by translating a first source sentence X_1 in the document $\boldsymbol{D}$ into the target sentence Y_1. We then self-train the model on the sentence pair (X_1, Y_1), which maximizes the log probabilities of each word in the generated sentence Y_1 given source sentence X_1. The self-training procedure runs gradient descent steps for a fixed number of steps with a weight decay. Weight decay keeps the updated values of weights closer to original values. We repeat the same update process for the remaining sentences in the document. The detailed implementation of self-training procedure during decoding is shown in Algorithm 1.

2.2 Multi-pass self-training

Since the document is processed in the left-to-right, monotonic order, our self-training procedure does not incorporate the choices of the model yet to be made on unprocessed sentences. In order to leverage global information from the full document and to further reinforce the choices made by the model across all generated sentences, we propose multi-pass document decoding with self-training. Specifically, we process the document multiple times monotonically from left to right while continuing self-training of the model. Multi-pass self-training only requires adding additional parameter P to self-training Algorithm 1. This parameter specifies the number of passes over the entire document.

2.3 Oracle self-training to upper bound performance

Since generated sentences are likely to contain some errors, our self-training procedure can reinforce those errors and thus potentially hurt the performance of the model on unprocessed sentences in the document. In order to isolate the effect of imperfect translations and estimate the upper bound of performance, we evaluate our self-training procedure with ground-truth translations as targets, which we call *oracle self-training*. Running oracle self-training makes it similar to the dynamic evaluation approach introduced in language modeling (Mikolov, 2012; Graves, 2013; Krause et al.,

2018), where input text to the language model is the target used to train the neural language model during evaluation. Oracle self-training is also related to domain adaptation in machine translation (Axelrod et al., 2011; Freitag and Al-Onaizan, 2016; Chu and Wang, 2018). Unlike domain adaptation in MT, oracle self-training only runs adaptation within a single document and does not rely on the entire in-domain document-level test data. We do not use the oracle in multi-pass self-training since this would make it equivalent to memorizing the correct translation for each sentence in the document and regenerating it again.

3   Related Work

Although there have been some attempts at tackling document-level neural machine translation (for example see proceedings of discourse in machine translation workshop (Popescu-Belis et al., 2019)), it has largely received less attention compared to sentence-level neural machine translation. Prior document-level NMT approaches (Jean et al., 2017; Wang et al., 2017; Kuang et al., 2017; Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Agrawal et al., 2018; Zhang et al., 2018; Miculicich et al., 2018) proposed different ways of conditioning NMT models on several source sentences in the document. Perhaps closest of those document NMT approaches to our work is the approach by Kuang et al. (2017), where they train a NMT model with a separate non-parametric cache (Kuhn and Mori, 1990) that incorporates topic information about the document. Recent approaches (Jean et al., 2019; Junczys-Dowmunt, 2019; Voita et al., 2019a) use only partially available parallel document data or monolingual document data. These approaches proposed to fill in missing context in the documents with random or generated sentences. Another line of document-level NMT work (Xiong et al., 2018; Voita et al., 2019b) proposed a two-pass document decoding model inspired by the deliberation network (Xia et al., 2017) in order to incorporate target side document context. A parallel line of work (Garcia et al., 2017, 2019; Yu et al., 2019) introduced document-level approaches that do not require training the context-conditional NMT model by introducing a separate language model to enforce the consistency in the outputs of sentence-level NMT model. Garcia et al. (2019) used a simple n-gram based semantic space language model (Hardmeier et al., 2012) to re-rank

the outputs of the sentence-level NMT model inside the beam-search algorithm to enforce document-level consistency. Yu et al. (2019) proposed a novel beam search method that incorporates document context inside noisy channel model (Shannon, 1948; Yu et al., 2017; Yee et al., 2019) that uses a powerful GPT2 language model (Radford et al., 2019). Similar to our work, their approach doesn't require training context-conditional models on parallel document corpora, but relies on separate target-to-source NMT model and unconditional language model to re-rank hypotheses of the source-to-target NMT model.

Closest to our work is the *dynamic evaluation* approach proposed by Mikolov (2012) and further extended by Graves (2013); Krause et al. (2018), where a neural language model is trained at evaluation time. However unlike language modeling where inputs are ground-truth targets used both during training and evaluation, in machine translation ground-truth translation are not available at decoding time in practical settings. The general idea of storing memories in the weights of the neural network rather than storing memories as copies of neural network activations, that is behind our approach and dynamic evaluation, goes back to 1970s and 1980s work on associative memory models (Willshaw et al., 1969; Kohonen, 1972; Anderson and Hinton, 1981; Hopfield, 1982) and to more recent work on fast weights (Ba et al., 2016).

Our work belongs to the broad category of self-training or pseudo-labelling approaches (Scudder, 1965; Lee, 2013) proposed to annotate the unlabeled data to train supervised classifiers. Self-training has been successfully applied to NLP tasks such as word-sense disambiguation (Yarowsky, 1995) and parsing (McClosky et al., 2006; Reichart and Rappoport, 2007; Huang and Harper, 2009). Self-training has also been used to label monolingual data to improve the performance of sentence-level statistical and neural machine translation models (Ueffing, 2006; Zhang and Zong, 2016). Recently, He et al. (2019) proposed noisy version of self-training and showed improvement over classical self-training on machine translation and text summarization tasks. Backtranslation (Sennrich et al., 2016a) is another popular pseudo-labelling technique that utilizes target-side monolingual data to improve performance of NMT models.

Method	Model	MT06	MT03	MT04	MT05	MT08
(Wang et al., 2017)	RNNSearch	37.76	-	-	36.89	27.57
(Kuang et al., 2017)	RNNSearch	-	-	38.40	32.90	31.86
(Kuang et al., 2017)	Transformer	48.14	48.05	47.91	48.53	38.38
(Zhang et al., 2018)	Doc Transformer	49.69	50.21	49.73	49.46	39.69
Ours (Non-BT)	Transformer	48.90	48.55	49.55	48.04	40.89
	Transformer + self-train	49.17	49.46	50.12	48.67	41.18
Ours (BT)	Transformer	51.29	51.69	52.48	53.00	42.72
	Transformer + self-train	52.30	53.36	52.83	53.67	43.68

Table 1: Results on NIST evaluation sets. The first four rows show the performance of the previous document-level NMT models from (Wang et al., 2017; Kuang et al., 2017; Zhang et al., 2018). The last four rows show performance of our baseline sentence-level Transformer models with and without self-training. BT: backtranslation.

	Model	MT06	MT03	MT04	MT05	MT08	Δ_{BLEU}
Non-BT	Baseline	48.90	48.55	49.55	48.04	40.89	-
	ST	49.26	48.91	49.81	48.27	41.10	+0.28
	Multi-pass ST	49.17	49.46	50.12	48.67	41.18	+0.53
	Oracle ST	49.82	49.59	49.93	48.58	42.30	+0.86
BT	Baseline	51.29	51.69	52.48	53.00	42.72	-
	ST	52.19	52.45	52.63	53.01	43.25	+0.47
	Multi-pass ST	52.30	53.36	52.83	53.67	43.68	+0.93
	Oracle ST	53.11	53.90	53.05	53.81	45.45	+1.63

Table 2: Ablation study on NIST evaluation sets measuring the effect on multiple passes of decoding and the oracle on self-training procedure. BT: backtranslation. ST: self-training.

4 Experiments

4.1 Datasets

We use the NIST Chinese-English (Zh-En), the WMT19 Chinese-English (Zh-En) and the Open-Subtitles English-Russian (En-Ru) datasets in our experiments.

The NIST training set consists of 1.5M sentence pairs from LDC-distributed news. We use MT06 set as validation set. We use MT03, MT04, MT05 and MT08 sets as held out test sets. The MT06 validation set consists of 1649 sentences with 21 sentences per document. MT03, MT04, MT05 and MT08 consist of 919, 1788, 1082 and 1357 sentences with 9, 9, 11 and 13 sentences on average per document respectively. We follow previous work (Zhang et al., 2018) when preprocessing NIST dataset. We preprocess the NIST dataset with punctuation normalization, tokenization, and lowercasing. Sentences are encoded using byte-pair encoding (Sennrich et al., 2016b) with source and target vocabularies of roughly 32K tokens. We use the case-insensitive `multi-bleu.perl` script with 4 reference files to evaluate the model.

The WMT19 dataset includes the UN corpus, CWMT, and news commentary. We filter the training data by removing duplicate sentences and sen-

tences longer than 250 words. The training dataset consits of 18M sentence pairs. We use news-dev2017 as a validation set and use newstest2017, newstest2018 and newstest2019 as held out test sets. newsdev2017, newstest2017, newstest2018 and newstest2019 consist of total of 2002, 2001, 3981 and 2000 sentences with average of 14, 12, 15 and 12 sentences per document respectively. We similarly follow previous work (Xia et al., 2019) when preprocessing the dataset. Chinese sentences are preprocessed by segmenting and normalizing punctuation. English sentences are preprocessed by tokenizing and true casing. We learn a byte-pair encoding (Sennrich et al., 2016b) with source and target vocabularies of roughly 32K tokens. We use sacreBLEU (Post, 2018) for evaluation.

The OpenSubtitles English-Russian dataset, consisting of movie and TV subtitles, was prepared by (Voita et al., 2019b).[1] The training dataset consists of 6M parallel sentence pairs. We use the context aware sets provided by the authors consisting of 10000 documents both in validation and test sets. Due to the way the dataset is processed, each document only contains 4 sentences. The dataset is preprocessed by tokenizing and lower casing. We

---

[1] https://github.com/lena-voita/good-translation-wrong-in-context

use byte-pair encoding (Sennrich et al., 2016b) to prepare source and target vocabularies of roughly 32K tokens. We use `multi-bleu.perl` script for evaluation.

4.2 Hyperparameters

We train a Transformer (Vaswani et al., 2017) on all datasets. Following previous (Zhang et al., 2018; Voita et al., 2019b; Xia et al., 2019) work we use the Transformer base configuration (`transformer_base`) on the NIST Zh-En and the OpenSubtitles En-Ru datasets and use the Transformer big configuration (`transformer_big`) on the WMT19 Zh-En dataset. Transformer base consists of 6 layers, 512 hidden units and 8 attention heads. Transformer big consists of 6 layers, 1024 hidden units and 16 attention heads. We use a dropout rate (Srivastava et al., 2014) of 0.1 and label smoothing to regularize our models. We train our models with the Adam optimizer (Kingma and Ba, 2014) using the same warm-up learning rate schedule as in (Vaswani et al., 2017). During decoding we use beam search with beam size 4 and length penalty 0.6. We additionally train backtranslated models (Sennrich et al., 2016a) on the NIST Zh-En and the OpenSubtitles En-Ru datasets. We use the publicly available English gigaword dataset (Graff et al., 2003) to create synthetic parallel data for the NIST Zh-En dataset and use synthetic parallel data provided by (Voita et al., 2019a) for the OpenSubtitles En-Ru dataset. When training backtranslated models, we oversample the original parallel data to make the ratio of synthetic data to original data equal to 1 (Edunov et al., 2018). We tune the number of update steps, learning rate, decay rate, and number of passes over the document of our self-training approach with a random search on a validation set. We use the range of $(5 \times 10^{-5}, 5 \times 10^{-1})$ for learning rate, range of $(0.001, 0.999)$ for decay rate, number of update steps $(2, 4, 8)$ and number of passes over the document $(2, 4)$ for random search. We found that best performing models required a small number of update steps (either 2 or 4) with a relatively large learning rate ($\sim 0.005 - 0.01$) and small decay rate ($\sim 0.2 - 0.5$). We use 3 random seeds to train each model in our experiments and report the average results. We use the Tensor2Tensor library (Vaswani et al., 2018) to train baseline models and to implement our method.

5 Results

We present translation quality results measured by BLEU on NIST dataset on Table 1. The self-training procedure improves the results of our sentence-level baseline by the average of 0.53 BLEU for non-backtranslated model and by 0.93 BLEU for backtranslated model for all evaluation sets. Our baseline sentence-level Transformer model trained without backtranslation outperforms previous document-level models by Wang et al. (2017) and Kuang et al. (2017) and is comparable to the document-level model proposed by Zhang et al. (2018). Backtranslation further improves the results of our sentence-level model leading to higher BLEU score compared to the Document Transformer (Zhang et al., 2018).

In Table 2, we show a detailed study of effects of multi-pass self-training and oracle self-training on BLEU scores on NIST evaluation sets. First, multiple decoding passes over the document give an additional average improvement of $0.25 - 0.45$ BLEU points compared to the single decoding pass over the document. Using oracle self-training procedure gives an average of 0.86 and 1.63 BLEU improvement over our non-backtranslated and backtranslated sentence-level baseline models respectively. Compared to using generated translations by the model, oracle self-training gives an improvement of 0.3 and 0.7 BLEU points for non-backtranslated and backtranslated models respectively.

The results on the WMT19 evaluation sets are presented on Table 3. Compared to the NIST dataset our self-training procedure shows an improvement of 0.1 BLEU over a sentence-level baseline model. Oracle self-training outperforms sentence-level baselines by a significant margin of 2.5 BLEU. We hypothesize that such a large gap between performance of oracle and non-oracle self-training is due to the more challenging nature of the WMT dataset which is reflected in the worse performance of sentence-level baseline on WMT compared to NIST. We investigate this claim by measuring the relationship between BLEU achieved by self-training and the relative quality of the sentence-level model on the NIST dataset. Figure 1 shows that the BLEU difference between self-training and sentence-level models monotonically increases as the quality of the sentence-level model gets better on the NIST dataset. This implies that we can expect a larger improvement from applying self-training as we improve the sentence-level model

Method	Architecture	Model	Valid17	Test17	Test18	Test19	Δ_{BLEU}
(Xia et al., 2019)	Transformer Big	-	-	24.2	24.5	-	-
Ours	Transformer Big	Baseline	22.5	23.4	24.1	25.4	-
		ST	22.6	23.4	24.1	25.5	+0.0
		Multi-pass ST	22.6	23.5	24.2	25.6	+0.1
		Oracle ST	24.3	25.7	25.6	29.8	+2.5

Table 3: Results on WMT'19 Chinese-English evaluation sets. The first row shows the performance of the Transformer Big model by (Xia et al., 2019). All models were trained without additional monolingual data and without pretraining. ST: self-training.

	Model	Valid	Test	Δ_{BLEU}
Non-BT	Baseline	31.38	31.38	-
	ST	31.41	31.39	+0.02
	Multi-pass ST	31.44	31.43	+0.06
	Oracle ST	31.65	31.66	+0.28
BT	Baseline	32.22	32.21	-
	ST	32.24	32.23	+0.02
	Multi-pass ST	32.25	32.26	+0.04
	Oracle ST	32.56	32.54	+0.34

Table 4: Results on OpenSubtitles English-Russian evaluation sets. ST: self-training.

on the WMT dataset. Preliminary experiments on training back-translated models didn't improve results on the WMT dataset. We leave further investigation of ways to improve the sentence-level model on the WMT dataset for future work.

The results on OpenSubtitles evaluation sets are in Table 4. Our self-training and oracle self-training approaches give the performance improvement of 0.1 and 0.3 BLEU respectively. We hypothesize that the small improvement of self-training is due to relatively small number of sentences in the documents in the OpenSubtitles dataset. We validate this claim by varying the number of sentences in the document used for self-training on NIST dataset. Figure 2 shows that the self-training approach achieves higher BLEU improvement as we increase the number of sentences in documents used for self-training.

6 Human Evaluation

We conduct a human evaluation study on the NIST Zh-En and the OpenSubtitles En-Ru datasets. For both datasets we sample 50 documents from the test set where translated documents generated by the self-training approach are not exact copies of the translated documents generated by the sentence-level baseline model. For the NIST Zh-En dataset we present reference documents, translated documents generated by the sentence-level baseline, and

translated documents generated by self-training approach to 4 native English speakers. For the Open-Subtitles En-Ru dataset we follow a similar setup, where we present reference documents, translated documents generated by sentence-level baseline, and translated documents generated by self-training approach to 4 native Russian speakers. All translated documents are presented in random order with no indication of which approach was used to generate them. We highlight the differences between translated documents when presenting them to human evaluators. The human evaluators are asked to pick one of two translations as their preferred option for each document. We ask the human evaluators to consider fluency, idiomaticity and correctness of the translation relative to the reference when entering their preferred choices.

We follow the setup of Yu et al. (2019) when performing human evaluation. We collect a total of 200 annotations for 50 documents from all 4 human evaluators and show results in Table 5. For both datasets, human evaluators prefer translated documents generated by the self-training approach to translated documents generated by the sentence-level model. For NIST Zh-En, 122 out of 200 annotations indicate a preference towards translations generated by self-training approach. For OpenSubtitles En-Ru, 118 out of 200 annotations similarly show a preference towards translations generated by our self-training approach. This is a statistically significant preference $p < 0.05$ according to two-sided Binomial test. When aggregated for each document by majority vote, for NIST Zh-En, translations generated by the self-training approach are considered better in 25 documents, worse in 12 documents, and the same in 13 documents. For OpenSubtitles En-Ru, translations generated by self-training approach are considered better in 23 documents, worse in 15 documents, and the same in 12 documents. The agreement between annotators for NIST Zh-En and OpenSub-
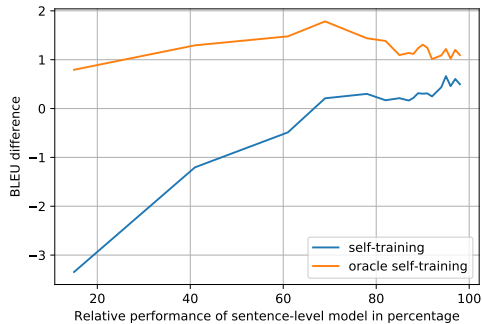
Figure 1: Relationship between relative performance of the sentence-level model and BLEU difference of self-training on the NIST dataset.

Figure 2: Relationship between number of sentences and BLEU improvement of self-training on the NIST dataset.

Dataset	Total	Self-train	Baseline
NIST Zh-En	200	122	78
OpenSubtitles En-Ru	200	118	82

Table 5: Human evaluation results on the NIST Zh-En and the OpenSubtitles En-Ru datasets. "Total" denotes total number of annotations collected from humans. "Self-train" denotes number of times evaluators preferred documents by the self-training approach. "Baseline" denotes number of times evaluators preferred documents by sentence-level baseline.

titles En-Ru is $\kappa = 0.293$ and $\kappa = 0.320$ according to Fleiss' kappa (Fleiss, 1971). For both datasets, the inter-annotator agreement rate is considered fair. The agreement rate is similar to the inter-annotator agreement found in the previous WMT human evaluation studies (Bojar et al., 2014).

7 Qualitative Results

In Table 6, we show four reference document pairs together with translated documents generated by the baseline sentence-level model and by our self-training approach. We emphasize the underlined words in all documents.

In the first two examples we emphasize the gender of the person marked on verbs and adjectives in translated Russian sentences. In the first example, the baseline sentence-level model inconsistenly produces different gender markings on the underlined verb сказал (masculine told) and underlined adjective сильной (feminine strong). The self-training approach correctly generates a translation with consistent male gender markings on both the underlined verb сказал and the underlined adjective сильным. Similarly, in the second example, the baseline model inconsistenly produces different
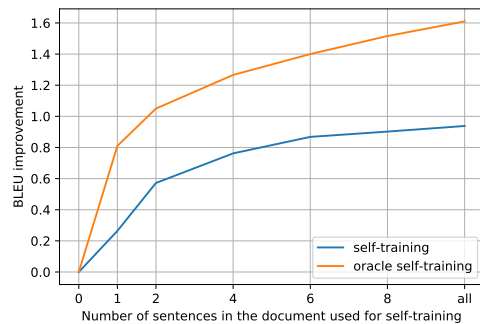
gender markings on the underlined verbs приглашена (feminine invited) and поругался (masculine fought). Self-training consistently generates female gender markings on both the underlined verbs приглашена (feminine invited) and поссорилась (feminine fought).

In the third example, we emphasize the underlined named entity in reference and generated translations. The baseline sentence-level model inconsistently generates the names "doyle" and "du" when referring to the same entity across two sentences in the same document. The self-training approach consistently uses the name "doyle" across two sentences when referring to the same entity. In the fourth example, we emphasize the plurality of the underlined words. The baseline model inconsistenly generates both singular and plural forms when referring to same noun in consecutive sentences. Self-training generates the noun "pilots" in correct plural form in both sentences.

8 Conclusion

In this paper, we propose a way of incorporating the document context inside a trained sentence-level neural machine translation model using self-training. We process documents from left to right multiple times and self-train the sentence-level NMT model on the pair of source sentence and generated target sentence. This reinforces the choices made by the NMT model thus making it more likely that the choices will be repeated in the rest of the document.

We demonstrate the feasibility of our approach on three machine translation datasets: NIST Zh-En, WMT19 Zh-En and OpenSubtitles En-Ru. We show that self-training improves sentence-level

149

Ref	мы с эйприл развелись . как я и <u>сказал</u> ... игра в ожидание . будь <u>сильным</u> . и всё получится .
Baseline	мы с эйприл развелись . ну , как я уже <u>сказал</u> ... игра ожидания . будь <u>сильной</u> . ты справишься .
Ours	мы с эйприл развелись . ну , как я уже <u>сказал</u> ... игра ожидания . будь <u>сильным</u> . ты справишься .

Ref	сёрен устраивает вечеринку по поводу своего дня рождения в субботу , а я не знаю , <u>пойду</u> ли я . почему бы тебе не пойти ? просто всё пошло не так . - и я <u>поссорился</u> с кнудом .
Baseline	в субботу день рождения сёрена и я не знаю , <u>приглашена</u> ли я . почему тебя не пригласили ? все просто пошло не так . - и я <u>поругался</u> с кнудом .
Ours	в субботу день рождения сёрена и я не знаю , <u>приглашена</u> ли я . почему тебя не пригласили ? все просто пошло не так . - и я <u>поссорилась</u> с кнудом .

Ref	we are actively seeking a local partner to set up a joint fund company , " <u>duchateau</u> said . <u>duchateau</u> said that the chinese market still has ample potentials .
Baseline	we are actively looking for a local partner to establish a joint venture fund company , " <u>doyle</u> said . <u>du</u> said that there is still a lot of room for the chinese market .
Ours	we are actively looking for a local partner to establish a joint venture fund company , " <u>doyle</u> said . <u>doyle</u> said that there is still great room for the chinese market .

Ref	in may this year , <u>13 pilots</u> with china eastern airlines wuhan company in succession handed in their resignations , which were rejected by the company . soon afterwards , <u>the pilots</u> applied one after another at the beginning of june to the labor dispute arbitration commission of hubei province for labor arbitration , requesting for a ruling that their labor relationship with china eastern airlines wuhan company be terminated .
Baseline	in may this year , <u>13 pilots</u> of china eastern 's wuhan company submitted their resignations one after another , but the company refused . <u>the pilot</u> then applied for labor arbitration with the hubei province labor dispute arbitration committee in early june , requesting the ruling to terminate the labor relationship with the wuhan company of china eastern airlines .
Ours	in may this year , <u>13 pilots</u> of china eastern 's wuhan company submitted their resignations one after another , but the company refused . subsequently , in early june , <u>the pilots</u> successively applied for labor arbitration with the hubei province labor dispute arbitration committee , requesting that the labor relationship with china eastern airlines be terminated .

Table 6: Four reference documents together with translations generated by the baseline sentence-level model and by our self-training approach. First two documents are taken from the OpenSubtitles English-Russian and second two documents are taken from the NIST Chinese-English dataset.

baselines by up to 0.93 BLEU. We also conduct a human evaluation study and show a strong preference of the annotators to the translated documents generated by our self-training approach. Our analysis demonstrates that self-training achieves higher improvement on longer documents and using better sentence-level models.

In this work, we only use self-training on source-to-target NMT models in order to capture the target side document context. One extension could investigate the application of self-training on both target-to-source and source-to-target sentence-level models to incorporate both source and target document context into generated translations. Overall, we hope that our work would motivate novel approaches of making trained sentence-level models better suited for document translation at decoding time.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides.

James A Anderson and Geoffrey E Hinton. 1981. Models of information processing in the brain. *Parallel models of associative memory*.

A. Axelrod, X. He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*.

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih,

Joel Z. Leibo, and Catalin Ionescu. 2016. Using fast weights to attend to the recent past. In *NIPS*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. 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 *ACL*.

Ondrej Bojar, C. Buck, C. Federmann, B. Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and A. Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *WMT@ACL*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*.

Markus Freitag and Y. Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *ArXiv*, abs/1612.06897.

Eva Martínez Garcia, C. Creus, and C. España-Bonet. 2019. Context-aware neural machine translation decoding. In *DiscoMT@EMNLP*.

Eva Martínez Garcia, C. Creus, C. España-Bonet, and Lluís Màrquez i Villodre. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108:85 – 96.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Christian Hardmeier, Joakim Nivre, and J. Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *EMNLP-CoNLL*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

J J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*.

Zhongqiang Huang and Mary Harper. 2009. Self-training pcfg grammars with latent annotations across languages. In *EMNLP*.

Sebastien Jean, Ankur Bapna, and Orhan Firat. 2019. Fill in the blanks: Imputing missing sentences for larger-context neural machine translation. *arXiv preprint arXiv:1910.14075*.

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

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *WMT*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Teuvo Kohonen. 1972. Correlation matrix memories. *IEEE Transactions on Computers*.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2018. Dynamic evaluation of neural sequence models. In *ICML*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Modeling coherence for neural machine translation with dynamic and topic caches. *arXiv preprint arXiv:1711.11221*.

Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. In *PAMI*.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *ACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *ACL*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *EMNLP*.

Tomas Mikolov. 2012. *Statistical language models based on neural networks*. Ph.D. thesis, Brno University of Technology.

Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong. 2019. *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. https://www.aclweb.org/anthology/D19-65.pdf.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Roi Reichart and Rai Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *ACL*.

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theor.*

Rico Sennrich. 2018. *Why the Time Is Ripe for Discourse in Machine Translation*. http://homepages.inf.ed.ac.uk/rsennric/wnmt2018.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Nicola Ueffing. 2006. Using monolingual source-language data to improve mt performance. In *IWSLT*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

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 *NIPS*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *EMNLP*.

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 *ACL*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. 1969. Non-holographic associative memory. *Nature*.

Yingce Xia, Xu Tan, Fei Tian, Fei Gao, Di He, Weicong Chen, Yang Fan, Linyuan Gong, Yichong Leng, Renqian Luo, Yiren Wang, Lijun Wu, Jinhua Zhu, Tao Qin, and Tie-Yan Liu. 2019. Microsoft research asia's systems for WMT19. In *WMT*.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *NIPS*.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2018. Modeling coherence for discourse neural machine translation. *arXiv preprint arXiv:1811.05683*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*.

Kyra Yee, Nathan Ng, Yann N. Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *EMNLP*.

L. Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, P. Blunsom, and Chris Dyer. 2019. Better document-level machine translation with bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. 2017. The neural noisy channel. In *ICLR*.

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 *EMNLP*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.