When and Why is Document-level Context Useful in Neural Machine Translation?

Yunsu Kim   Duc Thanh Tran   Hermann Ney
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
`{surname}@cs.rwth-aachen.de`

Abstract

Document-level context has received lots of attention for compensating neural machine translation (NMT) of isolated sentences. However, recent advances in document-level NMT focus on sophisticated integration of the context, explaining its improvement with only a few selected examples or targeted test sets. We extensively quantify the causes of improvements by a document-level model in general test sets, clarifying the limit of the usefulness of document-level context in NMT. We show that most of the improvements are not interpretable as utilizing the context. We also show that a minimal encoding is sufficient for the context modeling and very long context is not helpful for NMT.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017) has been originally developed to work sentence by sentence. Recently, it has been claimed that sentence-level NMT generates document-level errors, e.g. wrong coreference of pronouns/articles or inconsistent translations throughout a document (Guillou et al., 2018; Läubli et al., 2018).

A lot of research addresses these problems by feeding surrounding context sentences as additional inputs to an NMT model. Modeling of the context is usually done with fully-fledged NMT encoders with extensions to consider complex relations between sentences (Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019). Despite the high overhead in modeling, translation metric scores (e.g. Bleu) are often only marginally improved, leaving the evaluation to artificial tests targeted for pronoun resolution (Jean et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita

et al., 2018, 2019). Even if the metric score gets significantly better, the improvement is limited to specific datasets or explained with only a few examples (Tu et al., 2018; Maruf and Haffari, 2018; Kuang and Xiong, 2018; Cao and Xiong, 2018; Zhang et al., 2018; Maruf et al., 2019).

This paper systematically investigates when and why document-level context improves NMT, asking the following research questions:

- In general, how often is the context utilized in an interpretable way, e.g. coreference?

- Is there any other (non-linguistic) cause of improvements by document-level models?

- Which part of a context sentence is actually meaningful for the improvement?

- Is a long-range context, e.g. in ten consecutive sentences, still useful?

- How much modeling power is necessary for the improvements?

To answer these questions, we conduct an extensive qualitative analysis on non-targeted test sets. According to the analysis, we use only the important parts of the surrounding sentences to facilitate the integration of long-range contexts. We also compare different architectures for the context modeling and check sufficient model complexity for a significant improvement.

Our results show that the improvement in Bleu is mostly from a non-linguistic factor: regularization by reserving parameters for context inputs. We also verify that very long context is indeed not helpful for NMT, and a full encoder stack is not necessary for the improved performance.

2 Document-level NMT

In this section, we review the existing document-level approaches for NMT and describe our strategies to filter out uninteresting words in the context

input. We illustrate with an example of including one previous source sentence as the document-level context, which can be easily generalized also to other context inputs such as target hypotheses (Agrawal et al., 2018; Bawden et al., 2018; Voita et al., 2019) or decoder states (Tu et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018).

For the notations, we denote a source sentence by $\mathbf{f}$ and its encoded representations by H. A subscript distinguishes the previous (pre) and current (cur) sentences. e_i indicates a target token to be predicted at position i, and e_1^{i-1} are already predicted tokens in previous positions. Z denotes encoded representations of a partial target sequence.

2.1 Single-Encoder Approach

The simplest method to include context in NMT is to just modify the input, i.e. concatenate surrounding sentences to the current one and put the extended sentence in a normal sentence-to-sentence model (Tiedemann and Scherrer, 2017; Agrawal et al., 2018). A special token is inserted between context and current sentences to mark sentence boundaries (e.g. _BREAK_).

Figure 1 depicts this approach. Here, a single encoder processes the context and current sentences together as one long input. This requires no change in the model architecture but worsens a fundamental problem of NMT: translating long inputs (Koehn and Knowles, 2017). Apart from the data scarcity of a higher-dimensional input space, it is difficult to optimize the attention component to the long spans (Sukhbaatar et al., 2019).
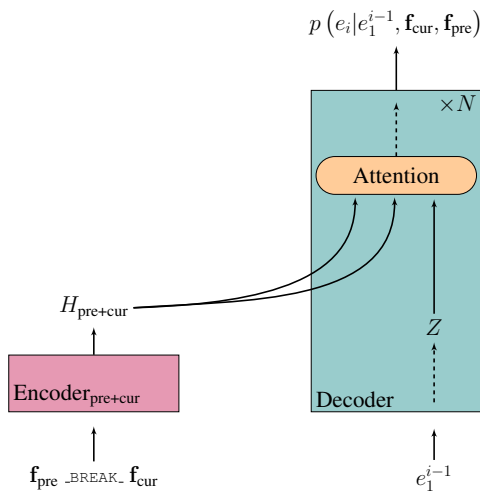
2.2 Multi-Encoder Approach

Alternatively, multi-encoder approaches encode each additional sentence separately. The model learns representations solely of the context sentences which are then integrated into the baseline model architecture. This tackles the integration of additional sentences on the architecture level, in contrast to the single-encoder approach. In the following, we describe two methods of integrating the encoded context sentences. The descriptions below do not depend on specific types of context encoding; one can use recurrent or self-attentive encoders with a variable number of layers, or just word embeddings without any hidden layers on top of them (Section 3.1).

2.2.1 Integration Outside the Decoder

The first method combines encoder representation of all input sentences before being fed to the decoder (Maruf and Haffari, 2018; Voita et al., 2018; Miculicich et al., 2018; Zhang et al., 2018; Maruf et al., 2019). It attends from the representations of the current sentence (H_{cur}) to those of the previous sentence (H_{pre}), yielding $\bar{H}$. Afterwards, a linear interpolation with gating is applied:

$$g\bar{H} + (1 - g)H_{\mathrm{cur}} \qquad (1)$$

where $g = \sigma\left(W_g\left[\bar{H}; H_{\mathrm{cur}}\right] + b_g\right)$ is gating activation and W_g, b_g are learnable parameters. This type of integration is depicted in Figure 2. By using such a gating mechanism, the model is capable of learning how much additional context information shall be included.
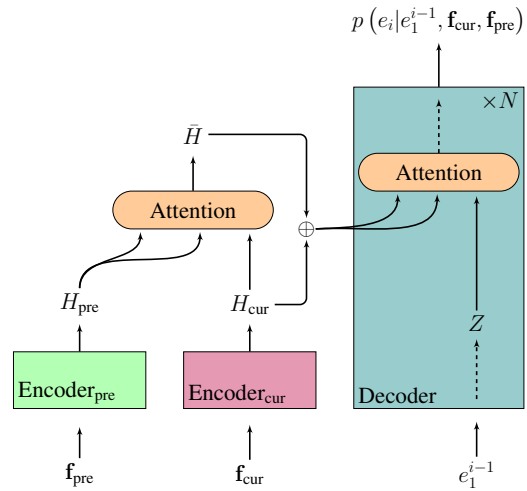

Figure 1: Single-encoder approach.

Figure 2: Multi-encoder approach integrating context outside the decoder.

25

2.2.2 Integration Inside the Decoder

Another method integrates the context inside the decoder; the partial target history e_1^{i-1} is available during the integration. Here, using the (encoded) target history as a query, the decoder attends directly to the context representations. It also has the original attention to the current sentence. Depending on the order of these two attention components, this type of integration has two variants.

Sequential Attentions The first variant is stacking the two attention components, with the output of one component being the query of another (Tu et al., 2018; Zhang et al., 2018).

Figure 3 shows the case when the current sentence is attended by the decoder first, which is then used to attend to the context sentence. This refines the regular attention to the current source sentence with additional context information. The order of the attention components may be switched. To block signals of potentially unimportant context information, a gating mechanism can be employed between the regular and context attention outputs like Section 2.2.1.
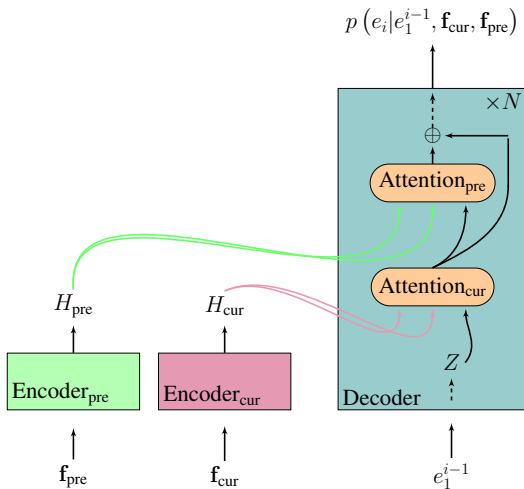

Figure 3: Multi-encoder approach integrating context inside the decoder with sequential attentions.

Parallel Attentions Figure 4 shows the case when performing the two attention operations in parallel and combining them with a gating afterwards (Jean et al., 2017; Cao and Xiong, 2018; Kuang and Xiong, 2018; Bawden et al., 2018; Stojanovski and Fraser, 2018). This method relates document-level context to the target history independently of the current source sentence, and lets the decoding computation faster.
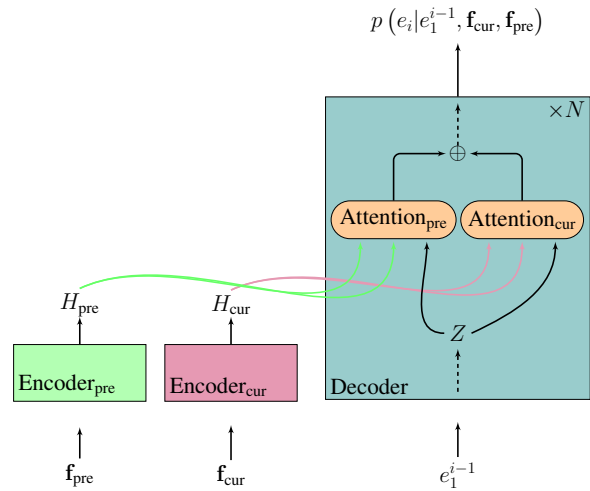

Figure 4: Multi-encoder approach integrating context inside the decoder with parallel attentions.

For each category above, we have described a common architecture shared by previous works in that category. There are slight variations but they do not diverge much from our descriptions.

2.3 Filtering of Words in the Context

Document-level NMT inherently has heavy computations due to longer inputs and additional processing of context. However, intuitively, not all of the words in the context are actually useful in translating the current sentence. For instance, in most literature, the improvements from using document-level context are explained with coreference, which can be resolved with just nouns, articles, and the conjugated words affected by them.

Under the assumption that we do not need the whole context sentence in document-level NMT, we suggest to retain only the context words that are likely to be useful. This makes the training easier with a smaller input space and less memory requirement. Concretely, we filter out words in the context sentences according to pre-defined word lists or predicted linguistic tags:

- Remove stopwords using a pre-defined list[1]
- Remove $n \in \mathbb{N}$ most frequent words
- Retain only named entities
- Retain only the words with specific parts-of-speech (POS) tags

The first method has the same motivation as Kuang et al. (2018) to ignore function words. The second method aims to keep infrequent words that

---

[1]https://github.com/explosion/spaCy

Original source	in recent years, I correctly foresaw that, in the absence of stronger fiscal stimulus (which was not forthcoming in either Europe or the United States), recovery from the Great Recession of 2008 would be slow.
Remove stopwords	recent years, I correctly foresaw absence stronger fiscal stimulus (forthcoming Europe United States), recovery Great Recession 2008 slow.
Remove most frequent words	recent correctly foresaw absence stronger fiscal stimulus forthcoming either States recovery Great Recession 2008 slow
Retain named entities	recent years Europe the United States the Great Recession 2008
Retain specific POS	years I foresaw the absence stimulus was forthcoming either Europe or the United States recovery the Great Recession 2008 would be

Table 1: Examples for filtering of words in the context (News Commentary v14 English→German).

are domain-specific or containing gender information. We empirically found that $n = 150$ works reasonably well. For the last two methods, we use the FLAIR[2] (Akbik et al., 2018) toolkit. We exclude the tags that are irrelevant to syntax/semantics of the current sentence. The detailed lists of retained tags can be found in the appendix.

The filtering is performed on word level in the preprocessing. When a sentence is completely pruned, we use a special token to denote an empty sentence (e.g. _EMPTY_). Table 1 gives examples of the filtering. We can observe that the original sentence is shortened greatly by removing redundant tokens, but the topic information and the important subjects still remain.

3 Experiments

We evaluate the document-level approaches in IWSLT 2017 English→Italian[3] and WMT 2018 English→German[4] translation tasks. We used TED talk or News Commentary v14 dataset as the training data respectively, preprocessed with the Moses tokenizer[5] and byte pair encoding (Sennrich et al., 2016) trained with 32k merge operations jointly for source and target languages. In all our experiments, one previous source sentence was given as the document-level context. A special token was inserted at each document boundary, which was also fed as context input when translating sentences around the boundaries. Detailed corpus statistics are given in Table 2.

All experiments were carried out with

SOCKEYE (Hieber et al., 2018). We used Adam optimizer (Kingma and Ba, 2015) with the default parameters. The learning rate was reduced by 30% when the perplexity on a validation set was not improving for four checkpoints. When it did not improve for ten checkpoints, we stopped the training. Batch size was 3k tokens, where the bucketing was done for a tuple of current/context sentence lengths. All other settings follow a 6-layer base Transformer model (Vaswani et al., 2017).

In all our experiments, a sentence-level model was pre-trained and used to initialize document-level models, which was crucial for the performance. We also shared the source word embeddings over the original and context encoders.

	en-it	en-de
Running Words	4.3M	8.1M
Sentences	227k	329k
Documents	2,045	8,891
Document Length (avg. #sent)	111	37

Table 2: Training data statistics.

3.1 Model Comparison

Model Architecture   Firstly, we compare the performance of existing single-encoder and multi-encoder approaches (Table 3). For each category of document-level methods (Section 2), we test one representative architecture (Figures 2, 3, 4) which encompasses all existing work in that category except slight variations. The tested methods are equal or closest to:

- Single-Encoder: Agrawal et al. (2018)

| | Context Encoder | | en-it | | en-de | |
Approach	Architecture	#layers	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	·	·	31.4	56.1	28.9	61.8
Single-Encoder	Transformer	6	31.5	57.2	28.9	61.4
Multi-Encoder (Out.)	Transformer	6	31.3	56.1	29.1	61.4
Multi-Encoder (Seq.)	Transformer	6	32.6	55.2	29.9	60.7
Multi-Encoder (Para.)	Transformer	6	**32.7**	**54.7**	30.1	60.3
		2	32.6	55.2	30.2	60.5
		1	32.2	55.8	30.0	60.4
	Word Embedding	·	32.5	54.8	**30.3**	**59.9**

Table 3: Comparison of document-level model architectures and complexity.

- Integration outside the decoder: Voita et al. (2018) without sharing the encoder hidden layers over current/context sentences
- Integration inside the decoder
  - Sequential attention: Decoder integration of Zhang et al. (2018) with the order of attentions (current/context) switched
  - Parallel attention: Gating version of Bawden et al. (2018)

The training of the single-encoder method was quite unstable. It took about twice as long as other document-level models, yet yielding no improvements, which is consistent with Kuang and Xiong (2018). Longer inputs make the encoder-decoder attention widely scattered and harder to optimize. We might need larger training data, massive pre-training, and much larger batches to train the single-encoder approach effectively (Junczys-Dowmunt, 2019); however, these conditions are often not realistic.

For the multi-encoder models, if the context is integrated outside the decoder ("Out."), it barely improves upon the baseline. By letting the decoder directly access context sentences with a separate attention component, they all outperform the single-encoder method, improving the sentence-level baseline up to +1.4% BLEU and -1.9% TER. Particularly, when attending to current and context sentences in parallel ("Para."), it provides more flexible and selective information flow from multiple source sentences to the decoder, thus producing better results than the sequential attentions ("Seq.").

Model Complexity In the linguistic sense, surrounding sentences are useful in translating the current sentence mostly by providing case distinctions of nouns or topic information (Section 4). The sequential relation of tokens in the surrounding sentences is important for neither of them. Therefore we investigate how many levels of sequential encoding is actually needed for the improvement by the context. From a 6-layer Transformer encoder, we gradually reduce the model complexity of the context encoder: 2-layer, 1-layer, and only using word embeddings without any sequential encoding. We remove positional encoding (Vaswani et al., 2017) when we encode only with word embeddings.

The results are shown in the lower part of Table 3. Context encoding without any sequential modeling (the last row) shows indeed comparable performance to using a full 6-layer encoder. This simplified encoding eases the memory-intensive document-level training by having 22% fewer model parameters, which allows us to adopt a larger batch size without accumulating gradients. For the remainder of this paper, we stick to using the multi-encoder approach with parallel attention components in the decoder and restricting the context encoding to only word embeddings.

3.2 Filtering Words in the Context

To make the context modeling even lighter, we analyze the effectiveness of the filtered context (Section 2.3) in Table 4. All filtering methods shrink the context input drastically without a significant loss of performance. Each method has its own motivation to retain only useful tokens in the

28

| | en-it | | en-de | | |
Context sentence	BLEU [%]	TER [%]	BLEU [%]	TER [%]	#tokens
None	31.4	56.1	28.9	61.8	-
Full sentence	32.5	54.8	30.3	59.9	100%
Remove stopwords	32.2	55.2	30.3	59.9	63%
Remove most frequent words	32.1	55.6	30.2	60.2	51%
Retain only named entities	32.3	55.4	30.3	60.3	**13%**
Retain specific POS	**32.5**	**55.2**	**30.4**	**60.0**	59%

Table 4: Comparison of context word filtering methods.
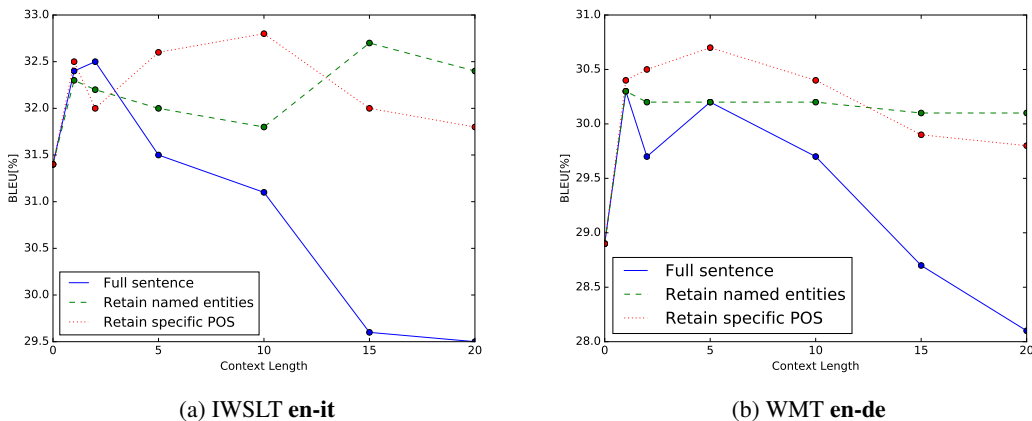

(a) IWSLT **en-it**

(b) WMT **en-de**

Figure 5: Translation performance as a function of document-level context length (in the number of sentences).

context; the results show that they are all reasonable in practice. In particular, using only named entities as context input, we achieve the same level of improvement with only 13% of tokens in the full context sentences. By filtering words in the context sentences, we can use more examples in each batch for a robust training.

3.3 Context Length

Filtered context inputs (Section 3.2) with a minimal encoding (Section 3.1) make it also feasible to include much longer context without much difficulty. Most of previous works on document-level NMT have not examined context inputs longer than three sentences.

Figure 5 shows the translation performance with an increasing number of context sentences. If we concatenate full context sentences (plain curves), the performance deteriorates severely. We found that it is hard to fit such long sequences in memory as the training becomes very erratic.

The training is much more stable with filtered context; the dashed/dotted curves do not drop significantly even when using 20 context sen-

tences. In the English→Italian task, the performance slightly improves up to 15 context sentences. In the English→German task, there is no improvement by extending the context length over 5 sentences. This discrepancy can be explained with document lengths in each dataset (Table 2). The TED talk corpus for English→Italian has much longer documents, thus it is probable to benefit from larger context windows. However, in general we observe only marginal improvements by enlarging the context length to more than one sentence, as seen also in Bawden et al. (2018), Miculicich et al. (2018), or Zhang et al. (2018).

4 Analysis

Simplifying the context encoder (Section 3.1) and filtering the context input (Section 3.2) are both inspired by the intuition that only a small part of the context is useful for NMT. In order to verify this intuition rigorously, we conduct an extensive analysis on how document-level context helps the translation process, manually checking every output of sentence-level/document-level NMT

models; automatic metrics are inherently not suitable for distinguishing document-level behavior. Our analysis is not constrained to certain discourse phenomena which are favored in evaluating document-level models. We quantify various causes of the improvements 1) regardless of its linguistic interpretability and 2) in a realistic scenario where not all the test examples require document-level context. Here are the steps we take:

1. Translate a test set with a sentence-level baseline and a document-level model.

2. Compute per-sentence TER scores of outputs from both models.

3. Select those cases where the document-level model improves the per-sentence TER over the sentence-level baseline.

4. Examine each case of 3 by looking at:
   - Source, context, and translation outputs
   - Attention distribution over the context tokens for each target token: averaged over all decoder layers/heads
   - Gating activation (Equation 1)

5. Classify each case into "coreference", "topic-aware lexical choice", or "not interpretable".

Statistics of each category on the test sets are reported in Table 5. The manual inspection of translation outputs is done by a native-level speaker of Italian or German, respectively.

Only a couple of cases belong to coreference, which is ironically the most advocated improvement in the literature on document-level NMT. One of them is shown in Table 6a. In the document-level NMT, the English word "said" is translated to a correct conjugation of "sagen" (= say) for the third person noun "der Präsident" (= the President). This can be explained by the high attention energy on "Trump" (Figure 7a) in the context sentence.

Another interpretable cause is topic-aware lexical choice (Table 6b). The document-level model actively attends to "seized" and "cocaine" in the context sentence (Figure 7b), and does not miss the source word "raids" in the translation ("Razzien"). When it corrects the translation of polysemous words, it is related to word sense disambiguation (Gonzales et al., 2017; Marvin and Koehn, 2018; Pu et al., 2018). This category includes also a coherence of text style in the translation outputs, depending on the context topic.

Category	#cases	
	en-it	en-de
Coreference	21	2
Topic-aware lexical choice	66	33
Not interpretable	292	1,211
Total TER improved	379	1,246
Total	1,147	2,998

Table 5: Causes of improvements by document-level context.

We found that only 7.5% of the TER-improved cases can be interpreted as utilizing document-level context. The other cases are mostly general improvements in adequacy or fluency which are not related to the given context. Table 6c shows such an example. It improves the translation by a long-range reordering and rephrasing some nouns, whose clues do not exist in the previous source sentence. Its attention distribution over the context words is totally random and blurry (Figure 7c).

A possible reason for the non-interpretable improvements is regularization of the model, since the training data of our experiments are relatively small. Figure 6 shows that, for most of the improved cases, the model has non-negligible gating activation towards document-level context, even if the output seems not to benefit from the context. It means that, when combining the encoded representations of context/current sentences, the model can reserve some of its capacity to the information from context inputs. This might effectively mitigate overfitting to the given training data.
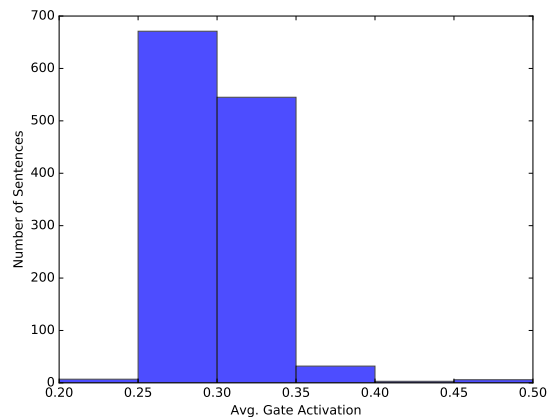

Figure 6: Gating activation for all TER-improved cases of the English→German task, averaged over all layers and target positions.

30

Previous src	inside the White House, <u>Trump</u> addressed Sikorsky representatives, joking with the media about his own fleet of company products.
Current src	"I know Sikorsky very well," the President said, "I have three of them."
Reference	„ich kenne Sikorsky sehr gut", sagte der Präsident, „ich habe drei davon."
Sent-level hyp	„ich kenne Sikorsky sehr gut", **so** der Präsident, „habe drei davon."
Doc-level hyp	„ich kenne Sikorsky sehr gut," **sagte** der Präsident, „ich habe drei davon".

(a) Coreference

Previous src	in addition, officials <u>seized</u> large quantities of marijuana and <u>cocaine</u>, firearms and several hundred thousand euros.
Current src	at simultaneous raids in Italy, two people were detained.
Reference	bei zeitgleichen Razzien in Italien wurden zwei Personen festgenommen.
Sent-level hyp	gleichzeitig wurden in Italien zwei Personen verhaftet.
Doc-level hyp	bei gleichzeitigen **Razzien** in Italien wurden zwei Menschen inhaftiert.

(b) Topic-aware lexical choice

Previous src	other cities poach good officials and staff members and offer attractive conditions.
Current src	the talk is of a downright "contest between public employers".
Reference	die Rede ist von einem regelrechten „Wettbewerb der öffentlichen Arbeitgeber".
Sent-level hyp	das Gerede ber einen „Wettkampf zwischen öffentlichen Arbeitgebern" ist von einem Gerechtigkeitstreit.
Doc-level hyp	die Rede ist von einem herben „Wettbewerb zwischen öffentlichen Arbeitgebern".

(c) Not interpretable

Table 6: Example translation outputs for each analysis category (WMT English→German newstest2018).
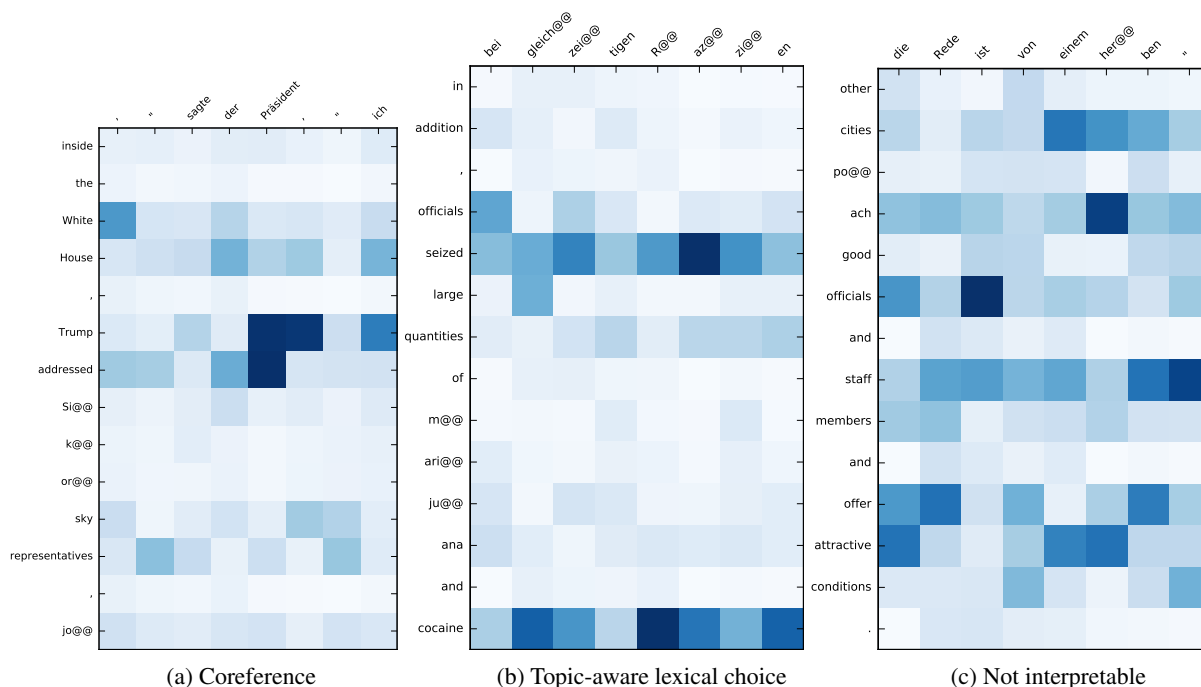

(a) Coreference	(b) Topic-aware lexical choice	(c) Not interpretable

Figure 7: Attention distribution over context words from target hypothesis.

| Condition | System | BLEU [%] | |
		en-it	en-de
Dropout 0.1	Sentence-level	31.4	28.9
	Document-level	**32.5**	**30.3**
Dropout 0.3	Sentence-level	**33.7**	**32.3**
	Document-level	33.5	32.0
Large training data	Sentence-level	-	**40.2**
	Document-level	-	39.9

Table 7: Sentence-level vs. document-level translation performance in different data/training conditions.

We argue that the linguistic improvements with document-level NMT have been sometimes oversold, and the document-level components should be tested on top of a well-regularized NMT system. In our experiments, we obtain a much stronger sentence-level baseline by applying a simple regularization (dropout), which the document-level model cannot outperform (Table 7).

On a larger scale, we also built a sentence-level model with all parallel training data available for the WMT 2019 task and fine-tuned only with document-level data (Europarl, News Commentary, newstest2008-2014/2016). The document-level training does not give any improvement in BLEU (last two rows of Table 7). There may exist document-level improvements which are not highlighted by the automatic metrics, but the amount of such improvements must be very small without a clear gain in BLEU or TER.

5 Conclusion

In this work, we critically investigate the advantages of document-level NMT with a thorough qualitative analysis and expose the limit of its improvements in terms of context length and model complexity. Regarding the questions asked in Section 1, our answers are:

- In general, document-level context is utilized rarely in an interpretable way.

- We conjecture that a dominant cause of the improvements by document-level NMT is actually the regularization of the model.

- Not all of the words in the context are used in the model; we leave out redundant tokens without loss of performance.

- A long-range context gives only marginal additional improvements.

- Word embeddings are sufficient to model document-level context.

For a fair evaluation of document-level NMT methods, we argue that one should make a sentence-level NMT baseline as strong as possible first, i.e. by using more data or applying proper regularization. This will get rid of by-product improvements from additional information flows and help to focus only on document-level errors in translation. In this condition, we show that document-level NMT can barely improve translation metric scores against such strong baselines. Targeted test sets (Bawden et al., 2018; Voita et al., 2019) might be helpful here to emphasize the document-level improvements. However, one should bear in mind that a big improvement in such test sets may not carry over to practical scenarios with general test sets, where the number of document-level errors in translation is inherently small.

Given these conclusions, a future research direction would be building a lightweight post-editing model to correct only document-level errors, not complicating the sentence-level model too much for a very limited amount of document-level improvements. To strengthen our arguments, we also plan to conduct the same qualitative analysis on other types of context inputs (e.g. translation history) and different domains.

Our implementation of document-level NMT methods is publicly available on the web.[6]

---

[6] https://github.com/ducthanhtran/sockeye_document_context

References

Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the english–german mt systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 200–207.

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 *Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, pages 424–432, Florence, Italy.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Shaohui Kuang and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 607–617.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606.

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.

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.

Sameen Maruf, André FT 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.

Rebecca Marvin and Philipp Koehn. 2018. Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 125–131.

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.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

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.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

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

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.