On Search Strategies for Document-Level Neural Machine Translation

Christian Herold **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{herold|ney}@cs.rwth-aachen.de

Abstract

Compared to sentence-level systems, document-level neural machine translation (NMT) models produce a more consistent output across a document and are able to better resolve ambiguities within the input. There are many works on document-level NMT, mostly focusing on modifying the model architecture or training strategy to better accommodate the additional context-input. On the other hand, in most works, the question on how to perform search with the trained model is scarcely discussed, sometimes not mentioned at all. In this work, we aim to answer the question how to best utilize a context-aware translation model in decoding. We start with the most popular document-level NMT approach and compare different decoding schemes, some from the literature and others proposed by us. In the comparison, we are using both, standard automatic metrics, as well as specific linguistic phenomena on three standard document-level translation benchmarks. We find that most commonly used decoding strategies perform similar to each other and that higher quality context information has the potential to further improve the translation.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2014; Vaswani et al., 2017) is widely adopted and produces excellent translations for many domains and language pairs. However, when these automatic translations are evaluated on the document level, they reveal shortcomings when it comes to consistency in style, entity-translation or correct inference of the gender, among other things (Läubli et al., 2018; Müller et al., 2018; Thai et al., 2022). Document-level NMT aims to resolve these shortcomings by taking the context of a sentence into account during translation. There exist many works on the topic of document-level NMT, proposing various changes to the standard transformer (Vaswani et al., 2017) architecture and training criteria to improve context incorporation and consequently translation quality. However, while the modeling and training aspects are covered in great detail in these works, the exact decoding strategy is often not very clearly described and sometimes not mentioned at all.

In this work, we head out to answer the question, which decoding strategy is most beneficial for document-level NMT systems. We compare all commonly used strategies, as well as some additional ones, on three standard document-level translation benchmarks. We find that most of the analyzed decoding strategies perform similar to each other. Also, higher quality context information can lead to better translations in certain scenarios.

2 Related Work

The earliest approaches to document-level NMT simply concatenate consecutive sentences without any further changes to the architecture compared to the sentence-level systems (Tiedemann and Scherrer, 2017; Agrawal et al., 2018). Later, some changes were made to the vanilla transformer architecture, like segment embeddings (Ma et al., 2020) or attention masking (Zhang et al., 2020; Petrick et al., 2022) and a move was made towards translating longer segments (Junczys-Dowmunt, 2019; Liu et al., 2020; Zheng et al., 2021; Bao et al., 2021; Sun et al., 2022). Other works employ a separate encoder to include the additional context on the source side (Jean et al., 2017; Bawden et al., 2018; Zhang et al., 2018; Voita et al., 2018) or make use of the context in a post-editing fashion (Voita et al., 2019; Xiong et al., 2019). Further approaches include the usage of a cache (Wang et al., 2017; Maruf and Haffari, 2018; Tu et al., 2018) or hierarchical attention networks (Miculicich et al., 2018; Maruf et al., 2019; Wong et al., 2020). Recently, several works have concluded that the simple concatenation approach used with the

Findings of the Association for Computational Linguistics: ACL 2023, pages 12827–12836

vanilla transformer architecture performs as good - if not better - than more complicated approaches that modify the model structure (Sun et al., 2022; Majumde et al., 2022). Since we also observed this in our internal comparisons, we decided to focus on this simple approach for our analysis in this work.

Several works made the argument that the improvements seen in automatic metric scores for document-level NMT systems are from regularization effects rather than from utilizing the additional context information (Kim et al., 2019; Li et al., 2020; Nguyen et al., 2021). In order to better asses the improvements gained by document-level NMT, several targeted test suites have been released (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019; Jwalapuram et al., 2019). However, all of these are based on just scoring contrastive examples without actually translating anything. Recently, Jiang et al. (2022) and Currey et al. (2022) have released frameworks that allow to score MT systems on their ability to generate contextually correct translations.[1]

3 Search Strategies

Training a document-level NMT system that takes the last k sentences as context is straightforward using the standard concatenation strategy (Tiedemann and Scherrer, 2017). Given some document level training data $(F_n, E_n), n = 1, ..., N$, where (F_n, E_n) denotes the n-th source-target sentence pair, during training we optimize the parameters Θ of the model towards

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \left\{ \sum_n \log p_\Theta(E_{n-k}^n | F_{n-k}^n) \right\}.$$

Here, E_{n-k}^n denotes the concatenation of the sentences $E_{n-k}, ..., E_n$.

During search, given a document F_1^M, we want to find the best translation $\hat{E}_1^M$ according to the model. Of course, exact search can not be performed and different works have used different methods to generate a translation:

full segment (Liu et al., 2020; Bao et al., 2021; Sun et al., 2022): we split the document into non-overlapping parts $F_1^k, F_{k+1}^{2k}, ..., F_{M-k}^M$ and translate each part separately using

$$\hat{E}_{i-k}^i = \underset{E_{i-k}^i}{\operatorname{argmax}} \left\{ p(E_{i-k}^i | F_{i-k}^i) \right\}, \quad (1)$$

which is approximated using standard beam search on the token level.

last sentence (Bawden et al., 2018; Agrawal et al., 2018; Zhang et al., 2020; Petrick et al., 2022; Majumde et al., 2022): we split the document into overlapping parts $..., F_{i-k}^i, F_{i-k+1}^{i+1}, ...$ and translate each part separately using Equation 1. From each translated part we choose only the last sentence to get one translation for every sentence in the document.

first sentence (Zhang et al., 2020): similar to *last sentence*, but from each translated part we choose only the first sentence to get one translation for every sentence in the document.

2-pass decoding (Maruf and Haffari, 2018; Maruf et al., 2019; Voita et al., 2019; Xiong et al., 2019): we first generate a translation $\tilde{E}_1^M$ of the document using a sentence-level NMT system. Then, the final hypothesis $\hat{E}_i$ for each sentence F_i is created using

$$\hat{E}_i = \underset{E_i}{\operatorname{argmax}} \left\{ p(E_i | F_{i-k}^i, \tilde{E}_{i-k}^{i-1}) \right\}.$$

doc-trans (Miculicich et al., 2018; Voita et al., 2019; Garcia et al., 2019; Fernandes et al., 2021): we generate the translation sentence by sentence, meaning

$$\hat{E}_1 = \underset{E_1}{\operatorname{argmax}} \left\{ p(E_1 | F_1) \right\},$$

$$\hat{E}_2 = \underset{E_2}{\operatorname{argmax}} \left\{ p(E_2 | F_1^2, \hat{E}_1) \right\},$$

$$...$$

doc-trans (beam) : similar to *doc-trans*, but instead of keeping just the best context $\hat{E}_1^{i-1}$, we keep the top-h candidates and prune them after each step i, analogous to beam search on the token level. $h = 12$ for all our experiments, the same as our token-level beam-size.

cheating : this is just used as a tool for analysis. The translation of each sentence F_i is created using the true target reference $\mathring{E}_1^M$ as context

$$\hat{E}_i = \underset{E_i}{\operatorname{argmax}} \left\{ p(E_i | F_{i-k}^i, \mathring{E}_{i-k}^{i-1}) \right\}.$$

no context : this is just used as a tool for analysis. The translation of each sentence F_i is created using no context information at all

$$\hat{E}_i = \underset{E_i}{\operatorname{argmax}} \left\{ p(E_i | F_i) \right\}.$$

[1]The framework by Currey et al. (2022) was not yet made available when we conducted our experiments.

	Cost
sentence-level	$\mathcal{O}(NL)$
document-level	
full segment	$\mathcal{O}(NL)$
last sentence	$\mathcal{O}(NLk)$
first sentence	$\mathcal{O}(NLk)$
2-pass decoding	$\mathcal{O}(2NL)$
doc trans	$\mathcal{O}(NL)$
doc trans (beam)	$\mathcal{O}(NLh)$

Table 1: Computational cost of decoding (=number of forward passes through the decoder) for each of the search strategies described above. h denotes the sentence-level beam size.

The different search strategies also have a different computational cost associated with them. The biggest factor regarding the decoding cost is the number of forward passes through the model, specifically the decoder, that we have to do. We list the computational costs for the different decoding approaches in Table 1 under the assumption that the document consists of N sentences with average sentence length L and the model uses $k-1$ sentences as context. Please note that the decoding time might follow a different dependence than the cost in the above table, since it heavily depends on the available hardware. For example, *doc trans* and *doc trans (beam)* might have the same decoding time, if we have enough computational resources available, since the additional computations in *doc trans (beam)* can all be done in parallel.

4 Experiments

We perform experiments on three document-level translation benchmarks, called **NEWS** (En→De), **TED** (En→It) and **OS** (En→De). For the details regarding data conditions and preparation, as well as model training, we refer to Appendix A. For the context-aware systems, we concatenate 3 adjacent sentences (i.e. $k=3$) using a special token <sep>. For the two En→De tasks, we also evaluate the systems on the ContraPro test set (Müller et al., 2018). Instead of scoring and ranking the contrastive examples in ContraPro, as the authors have originally envisioned, we translate the source side to calculate BLEU and TER as well as to score the pronoun translations according to Section 4.1. We can not evaluate the *full segment* search strategy on ContraPro, because the sentences are not adjacent since they come from different documents.

	NEWS	TED	OS
sentence-level			
ref	4.61	4.15	2.97
hyp	1.63	1.56	1.48
document-level			
ref	4.46	3.96	2.70
hyp *no context*	1.64	1.53	1.47
hyp *full segment*	1.62	1.50	1.41
hyp *last sentence*	1.61	1.48	1.42
hyp *first sentence*	1.63	1.52	1.48
hyp *2-pass decoding*	1.68	1.49	1.48
hyp *doc trans*	1.67	1.48	1.42
hyp *doc trans (beam)*	1.67	1.48	1.41
hyp *cheating*	1.69	1.53	1.56

Table 2: Perplexity values on the test set for different search strategies.

4.1 Evaluating Pronoun Translation

As further analysis, we measure how well ambiguous pronouns are handled when translating from English to German. Regarding gender, the English third-person pronoun 'it' (and its other forms), can be translated to the German words 'er', 'sie' or 'es', depending on which noun it refers to. On the other hand, ambiguities in the formality come from second-person pronouns. For example, the English word 'you'can be translated to 'sie' or 'du' depending if we are in a formal setting or not. To report accuracies for pronoun (3 classes: male/female/neuter) and formality (2 classes: formal/informal) translation, we extend the BlonDe metric created by Jiang et al. (2022)[2]. First, we expand the framework to work for German references, by including German NER and POS taggers[3] as well as including German pronoun mappings. For the gender category, we mostly follow Jiang et al. (2022), but additionally require that a corresponding pronoun must also be present in the source sentence.[4] For the style category, we take into account examples where a second person pronoun appears in the source sentence, and a corresponding formal

[2]Our extension can be found in this fork: https://github.com/christian3141/BlonDe

[3]https://spacy.io/models/de

[4]In ContraPro, we find 5011/4085/4817 examples for male/female/neuter respectively. The difference to the 4000/4000/4000 reported by (Müller et al., 2018) means that in some cases we count multiple occurrences in a single example.

	NEWS				TED		OS			
	test		ConPro		test		test		ConPro	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
sentence-level										
external	†32.3	-	-	-	‡33.4	-	*37.3	-	*30.5	-
ours	32.8	49.0	18.4	65.5	34.2	46.3	37.1	43.8	29.7	52.8
document-level										
no context	33.4	48.5	18.6	65.6	34.0	46.7	36.9	44.5	29.4	53.2
full segment	33.4	48.6	-	-	34.3	46.3	38.2	43.9	-	-
last sentence	33.4	48.3	19.7	63.4	34.7	45.9	37.8	43.9	31.4	51.5
first sentence	33.4	48.6	18.8	65.6	34.1	46.3	37.8	44.1	29.5	53.1
2-pass decoding	32.8	48.6	19.5	63.9	34.5	46.2	37.4	44.4	31.1	51.8
doc trans	33.0	48.3	19.8	63.3	34.6	46.0	37.7	44.0	31.4	51.3
doc trans (beam)	33.0	48.3	19.7	63.4	34.5	46.0	38.3	43.8	31.3	51.5
cheating	32.2	49.4	19.3	65.1	34.1	46.6	39.6	42.9	33.3	49.1

Table 3: BLEU and TER scores (in percent) for the different tasks and decoding strategies. External baselines are from † Kim et al. (2019), ‡ Yang et al. (2022) and *Huo et al. (2020).

	NEWS	OS	
	ConPro gender	test style	ConPro gender
sentence-level	45.3	59.4	41.4
document-level			
no context	45.9	59.7	42.3
full segment	-	60.8	-
last sentence	56.1	60.3	66.5
first sentence	44.9	59.2	43.0
2-pass decoding	55.6	59.9	65.1
doc trans	56.1	58.7	66.4
doc trans (beam)	56.1	60.6	66.3
cheating	63.3	73.2	73.7

Table 4: F1 scores (in percent) for pronoun translation on different test sets.

or informal pronoun appears in the reference.[5]

4.2 Perplexities

First, we compare the perplexities of the hypotheses from the different search strategies, which are listed in Table 2. The first thing to note is, that the reference has a much higher perplexity than all hypotheses, which is commonly seen for NMT systems. All document-level search strategies result in different hypotheses, which however have a similar perplexity score. Surprisingly, the *cheating* setting generates the worst translation perplexity-wise, even worse than using *no context*. This might be related to the observation, that the reference has a worse perplexity than any hypothesis, which is rather a modelling error than a search error.

4.3 Automatic Metrics

Next, we evaluate the hypotheses based on the common automatic metrics BLEU and TER. The results are shown in Table 3. The hypotheses created with *no context* seem to have the same quality as the sentence-level baseline. Surprisingly, the true reference as context does not improve performance on the NEWS and TED test sets. This indicates that the improvements seen on these test sets for the document-level system might not be related to better context incorporation. On the contrary, the OS system creates the best hypothesis with the true reference as context. All the actual decoding strategies give similar performance in terms of BLEU and TER with *2-pass decoding* being a little bit behind. A special case is the *first sentence* strategy, which performs quite well on the standard test sets but poorly on ContraPro. This is, because ContraPro is designed in a way that the left side context is more important for translation than the right side.

Finally, we analyze the quality of the pronoun translation as discussed in Section 4.1. In principle, we could calculate the F1 score for both, gender and formality, on all En→De test sets. However, we discard the cases where one or more classes have less than 100 examples. This leaves us with the three test sets depicted in Table 4. As a sanity check,

[5]In the OS test set, we count 416 and 605 examples for formal and informal examples respectively.

we also report the ContraPro accuracies calculated from scoring the contrastive references as described in (Müller et al., 2018). They are 48.2/45.8 for sentence-level and 68.2/82.2 for document-level for NEWS/OS respectively. That means, with just scoring, we overestimate the capabilities of the system, but the trend is still consistent.[6] Using the true reference leads to the best results in all cases. *no context* and *first sentence* leaves us with sentence-level performance on the gender tasks, while all other decoding strategies perform similarly. For the formality, none of the methods can significantly outperform the sentence level system, although the *cheating* experiment shows that the system could do better if a better context information is provided. This might be, because segments of 3 sentences are too short to reliably detect if a setting is formal or informal, without access to the true reference.

5 Conclusion

In this work, we analyze decoding strategies for document-level NMT systems. Using the most popular document-level translation approach, we compare different search strategies found in the literature against methods developed by us. We find that most of the commonly used decoding strategies result in similar performance, both in terms of common automatic metrics, as well as on specific pronoun evaluation tasks. Therefore, we conclude that it is important to include the context information during decoding, but the exact way in which to do this is not as important. Also, we find that the document-level systems could actually profit from higher quality context information, in situations where this context is most relevant for translation.

Acknowledgements

This work was partially supported by the project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A, and by NeuroSys which, as part of the initiative "Clusters4Future", is funded by the Federal Ministry of Education and Research BMBF (03ZU1106DA).

Limitations

In this work, we limit our experiments to the most commonly used document-level system architecture and training criterion. Other approaches exist, which might exhibit a different behavior in decoding. Two out of the three document-level translation tasks we use in this work are low resource with less than 500k sentence-pairs as training data. We chose these tasks due to computational limitations and to be better comparable to other works, but higher resource scenarios are more realistic for actual applications. We limit the analysis of pronoun translation to the English-German language pair. Also, there are other aspects of document-level NMT, like consistent translation of entities, which we did not consider in our analysis.

[6] The precision and recall are roughly the same for the F1 scores reported in Table 4.

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.

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

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455.

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.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Anna Currey, Maria Nădejde, Raghavendra Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. *arXiv preprint arXiv:2211.01355*.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al.

2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André FT Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478.

Eva Martínez Garcia, Carles Creus, and Cristina España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 13–23.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616.

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

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

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

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

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.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. *arXiv preprint arXiv:2005.03393*.

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)*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511.

Suvodeep Majumde, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation. *arXiv preprint arXiv:2210.10906*.

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.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.

Mathias Müller, Annette Rios Gonzales, 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.

Toan Q Nguyen, Kenton Murray, and David Chiang. 2021. Data augmentation by concatenation for low-resource translation: A mystery and a solution. *IWSLT 2021*, page 287.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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.

Frithjof Petrick, Jan Rosendahl, Christian Herold, and Hermann Ney. 2022. Locality-sensitive hashing for long context neural machine translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 32–42.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

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)*, pages 1715–1725.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. *arXiv preprint arXiv:2210.14250*.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *arXiv preprint arXiv:1704.04347*.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. *arXiv preprint arXiv:2004.09894*.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7338–7345.

Jian Yang, Yuwei Yin, Liqun Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Furu Wei, and Zhoujun Li. 2022. Gtrans: Grouping and fusing transformer layers for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

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. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3983–3989.

A Appendix

For the **NEWS En→De** task, the parallel training data (around 300k sentence pairs, news-domain) comes from the `NewsCommentaryV14` corpus[7]. As validation/test set we use the WMT `newstest2015/newstest2018` test sets from the WMT news translation tasks (Farhad et al., 2021). For the **TED En→It** task, the parallel training data (around 200k sentence pairs, scientific-talks-domain) comes from the IWSLT17 Multilingual Task (Cettolo et al., 2017). As validation set we use the concatenation of `IWSLT17.TED.dev2010` and `IWSLT17.TED.tst2010` and as test set we use `IWSLT17.TED.tst2017.mltlng`. For the **OS En→De** task, the parallel training data (around 22.5M sentence pairs, subtitle-domain) comes from the `OpenSubtitlesV2018` corpus (Lison et al., 2018). We use the same train/validation/test splits as Huo et al. (2020) and additionally remove all segments that are used in the ContraPro test suite (Müller et al., 2018) from the training data. The data statistics for all tasks can be found in Table 5.

task	dataset	# sent.	# doc.
NEWS	train	330k	8.5k
	valid	2.2k	81
	test	3k	122
	ContraPro	12k	12k
TED	train	232k	1.9k
	valid	2.5k	19
	test	1.1k	10
OS	train	22.5M	29.9k
	valid	3.5k	5
	test	3.8k	5
	ContraPro	12k	12k

Table 5: Data statistics for the different document-level translation tasks.

Since in the original release of ContraPro only left side context is provided, we extract the right side context ourselves from `OpenSubtitlesV2018` based on the meta-information of the segments.

We tokenize the data using byte-pair-encoding (Sennrich et al., 2016; Kudo, 2018) with 15k joint merge operations (32k for OS En→De). The models are implemented using the fairseq toolkit (Ott et al., 2019) following the transformer base architecture (Vaswani et al., 2017) with dropout 0.3 and label-smoothing 0.2 for **NEWS En→De** and **TED En→It** and dropout 0.1 and label-smoothing 0.1 for **OS En→De**. This resulted in models with ca. 51M parameters for NEWS and TED and ca. 60M parameters for OS for both the sentence-level and the document-level systems. All systems are trained until the validation perplexity does no longer improve and the best checkpoint is selected using validation perplexity as well. Training took around 24h for NEWS and TED and around 96h for OS on a single NVIDIA GeForce RTX 2080 Ti graphics card. Due to computational limitations, we report results only for a single run. For the generation of segments (see Section 3), we use beam-search on the token level with beam-size 12 and length normalization. To calculate BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) we use SacreBLEU (Post, 2018).

[7] https://data.statmt.org/news-commentary/v14/

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☒ A2. Did you discuss any potential risks of your work?
The authors do not foresee potential risks of this work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 Experiments

☑ B1. Did you cite the creators of artifacts you used?
Section 4 Experiments

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All artifacts that were used allow such usage for research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All artifacts that were used allow such usage for research purposes.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We only use standard datasets which allow usage for research purposes.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 Experiments

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section Appendix

C ☑ Did you run computational experiments?

Section 4 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 Experiments

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 Experiments

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.