Exploring Paracrawl for Document-level Neural Machine Translation

Yusser Al Ghussin[1,2], **Jingyi Zhang**[3] **and Josef van Genabith**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus, Saarbrucken, Germany
[2]Department of Language Science and Technology, Saarland University, Germany
[3]Hasso-Plattner-Institut (HPI), Potsdam, Germany
yusser.al_ghussin/Josef.Van_Genabith@dfki.de,Jingyi.Zhang@hpi.de

Abstract

Document-level neural machine translation (NMT) has outperformed sentence-level NMT on a number of datasets. However, document-level NMT is still not widely adopted in real-world translation systems mainly due to the lack of large-scale general-domain training data for document-level NMT. We examine the effectiveness of using Paracrawl for learning document-level translation. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models. We also perform a targeted pronoun evaluation and show that document-level models trained with Paracrawl data can help context-aware pronoun translation. We release our data and code here[1].

1 Introduction

The Transformer translation model (Vaswani et al., 2017), which performs sentence-level translation based on attention networks, has achieved great success and significantly improved the state-of-the-art in machine translation. Compared to sentence-level translation, document-level translation (Xu et al., 2021; Bao et al., 2021; Jauregi Unanue et al., 2020; Ma et al., 2020; Maruf et al., 2019; Tu et al., 2018; Maruf and Haffari, 2018) performs translation at document-level and can potentially further improve translation quality, e.g., document-level context can help word disambiguation for translating words with multiple senses, document-level translation can help pronoun translation which requires context outside of the current sentence (Müller et al., 2018), document-level translation can improve document-level lexical cohesion in the translation (Voita et al., 2019).

Document-level neural machine translation (NMT) has received much attention in recent years (Bao et al., 2021; Donato et al., 2021; Fernandes et al., 2021; Kang et al., 2020; Saunders et al., 2020; Yu et al., 2020; Zheng et al., 2020; Yang et al., 2019; Kuang et al., 2018; Bawden et al., 2018; Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018). Existing works showed that document-level translation can outperform sentence-level translation for a number of datasets, such as TED, News, Europarl (Bao et al., 2021; Donato et al., 2021; Xu et al., 2021). Although document-level NMT has shown promising results on a number of benchmarks, document-level NMT is still not widely adopted in real-world translation systems mainly due to the lack of large-scale general domain training data for document-level NMT.

We examine the effectiveness of using Paracrawl (Bañón et al., 2020) for learning document-level NMT. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus[2] was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can

[1]https://github.com/Yusser96/Exploring-Paracrawl-for-Document-level-Neural-Machine-Translation

[2]https://paracrawl.eu/

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1304–1310

be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models and also improving context-aware pronoun translation (Müller et al., 2018).

2 Extracting Parallel Paragraphs from Paracrawl

Paracrawl (Bañón et al., 2020) is a large-scale parallel corpus crawled from the Internet and contains texts from various domains. The official Paracrawl corpus was released as aligned sentence pairs. For high-resource language pairs like German-English, the Paracrawl corpus provides 278M aligned sentences for the translation task. As Paracrawl was released as aligned sentence pairs, previous work only used Paracrawl for learning sentence-level translation. However, as described in the Paracrawl documentation, Paracrawl was constructed by first identifying aligned webpages (URLs) and then aligning sentences within aligned webpages. Although texts in parallel webpages tend to be noisy and are only loosely aligned, we demonstrate that extracting parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments can provide effective training data for learning document-level translation. Below we describe how we extract parallel paragraphs from Paracrawl.

Extracting We extract parallel paragraphs from Paracrawl and examine the effectiveness of using these aligned paragraphs for learning document-level NMT. We used German-English, a high-resource language pair, in our experiments. As the main purpose of the original Paracrawl project is to collect aligned sentences, Paracrawl did not officially release all the webpage-level aligned texts. Paracrawl only released a subset of all the aligned webpage texts, for the German-English language pair[3]. We extract parallel paragraphs from the released parallel webpages: we first download aligned webpages[4] (one German webpage is aligned with one English webpage) and then download the automatic sentence alignments (vecalign[5]) provided by Paracrawl; we then split a webpage into paragraphs according to the newline symbol; then extract aligned paragraphs from aligned webpages according to the automatic sentence alignments. We consider two paragraphs to be aligned if sentences in these two paragraphs are aligned to each other and not aligned to any other paragraphs. We discarded sentences that are not one-to-one aligned (e.g. one English sentence aligned to two German sentences) and discarded sentences that are not aligned to any other sentences. We discarded repeated paragraphs and discarded paragraphs that only contain a single sentence. Paragraphs with non-monotonic sentence alignments were also discarded.

	Paragraphs	Sentences	Words	
			En	De
Train	1.5M	5.5M	118M	109M
Dev	402	1504	32K	29K
Test	411	1510	33K	30K

Table 1: Statistics of parallel paragraphs extracted from Paracrawl.

Original vecalign	147M
After parallel paragraph extraction	11.7M
After cleaning	5.5M

Table 2: Numbers of remaining sentence pairs after different processing steps.

Cleaning To improve the quality of the extracted parallel paragraphs, we removed sentences which do not belong to the correct language[6], removed paragraphs that are too short (contain less than 30 words) and removed paragraphs with more than 50% overlap.

We randomly split the extracted parallel paragraphs into training, development and test sets as shown in Table 1. Note that, compared to the officially released Paracrawl corpus, the size of parallel paragraphs that we extracted from Paracrawl is still relatively small. This is because (i) the aligned webpages provided by Paracrawl we used for parallel paragraph extraction is only a subset of all Paracrawl data (ii) many parallel sentences were discarded due to our strict extraction rules as shown in Table 2. For future work, we will collect more webpage-level aligned Paracrawl data and test more flexible extraction rules.

We show the length statistics of the parallel paragraphs that we extracted from Paracrawl in Table 3. Compared to normal documents, the parallel paragraphs extracted from Paracrawl are generally much shorter. However, sentences in the same paragraph are usually closely related and can provide

[3]https://www.statmt.org/paracrawl-benchmarks/

[4]https://www.statmt.org/paracrawl-benchmarks/paracrawl-benchmark.en-de.aligned-docs.xz

[5]https://www.statmt.org/paracrawl-benchmarks/paracrawl-benchmark.en-de.vecalign.xz

[6]https://pypi.org/project/langid/

Sentences	Distribution
2	34.63%
3	29.31%
4	15.99%
5∼10	18.29%
>10	1.77%

Table 3: Distribution of paragraph length. For example, the first line of numbers mean 34.63% of the extracted paragraphs contain 2 sentences.

useful context information to help the translation of each other. In our experiments, we use the extracted parallel paragraphs as parallel documents to train document-level NMT models and we show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models.

3 Document-level Translation with Paracrawl

3.1 Modeling

Previous works have shown that document-level NMT models can outperform sentence-level models on several benchmarks. We adopted the recent document-level translation model, G-Transformer (Bao et al., 2021)[7], in our experiments to test the effectiveness of using Paracrawl data for learning document-level translation. The G-Transformer model is based on the standard sentence-level Transformer (Vaswani et al., 2017), but uses a whole document together as the input of the model and then generates translation for the whole document. G-Transformer improves the standard Transformer model for document-level translation with extra group tags and group attention. Each word (both source and target words) in the document is assigned with a group tag to indicate which sentence this word belongs to. Compared to the standard Transformer attention, G-Transformer computes group attention using group tags to encourage local attention and reducing the hypothesis space of the attention, especially from target to source, for long documents. G-Transformer outperformed sentence-level Transformer and obtained new state-of-the-art results on three document-level translation benchmarks.

[7]https://github.com/baoguangsheng/g-transformer

3.2 Experimental Setting

We tokenize and truecase all data with MOSES (Koehn et al., 2007) scripts, and then perform subword segmentation with byte pair encoding (BPE) (Sennrich et al., 2016) using 30k merging operations. For both the sentence-level Transformer and the document-level G-Transformer, we used the base model setting (Vaswani et al., 2017) with 6-layer encoder/decoder, 512-dimension word embedding and 2048 hidden units for the feed forward networks. Following the G-Transformer experimental settings (Bao et al., 2021), we set the max length of a document to 512 BPE tokens (if a document is longer than 512 tokens, we split it into multiple instances). For model training, we first pretrained a standard sentence-level Transformer for 100k training steps and then finetuned the sentence-level Transformer to learn document-level G-Transformer for another 100k training steps. For a fair comparison between sentence-level and document-level translation, we also applied pretraining for sentence-level translation, i.e. we train a sentence-level Transformer with 100k pretraining and 100k finetuning steps in order to compare with G-Transformer.

3.3 Evaluating with BLEU, COMET and Targeted Pronoun Evaluation

We used BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) for translation quality evaluation[8] and we performed significance testing with bootstrap resampling (Koehn, 2004). In addition to BLEU and COMET which evaluate the general translation quality, we also performed a targeted evaluation for pronoun translation following Müller et al. (2018)'s work. The correct translation of a pronoun often requires context outside of the current sentence. Therefore, evaluation of pronoun translation can demonstrate the advantage of document-level NMT models. Following Müller et al. (2018)'s work, we computed the accuracy of our models choosing the correct translation for the English pronoun "it" from the three possible German words "es", "er" and "sie". We used a context of 5 sentences for the test data[9] provided by Müller et al. (2018) for pronoun evaluation.

[8]Both BLEU and COMET scores were computed on sentence-level using
https://github.com/mjpost/sacrebleu
https://unbabel.github.io/COMET/html/index.html

[9]https://github.com/ZurichNLP/ContraPro

		Train-Paracrawl-Only				Train-Combined			
		EnDe		DeEn		EnDe		DeEn	
		sent	doc	sent	doc	sent	doc	sent	doc
BLEU	Paracrawl	26.57	**26.96**	30.94	**31.94**†	26.44	**27.40**†	31.36	**31.92**†
	Europarl	22.78	**23.04**	27.76	**28.69**†	28.60	**29.30**†	35.53	**35.90**†
	TED	23.56	**24.39**†	27.89	**28.24**	25.70	**26.13**	30.15	**31.49**†
	News	31.55	**32.07**†	34.59	**35.39**†	33.08	**33.89**†	35.99	**36.39**
COMET	Paracrawl	16.16	**17.99**	18.90	**20.99**†	17.66	**20.44**†	20.60	**22.02**
	Europarl	38.34	**39.84**†	42.39	**45.10**†	54.10	**54.80**†	55.78	**56.31**†
	TED	23.14	**23.77**	41.34	**42.62**	34.21	**35.25**	46.02	**46.92**
	News	33.84	**35.51**†	44.45	**48.26**†	42.49	**44.20**†	49.66	**51.35**†

Table 4: Translation results. Train-Paracrawl-Only: only Paracrawl as training data. Train-Combined: Paracrawl, Europarl, TED and News combined as training data. † represents a significant difference at the $p < 0.01$ level.

	Train-Paracrawl-only				Train-Combined			
	total	es	er	sie	total	es	er	sie
sent	0.43	0.91	0.15	0.24	0.43	**0.94**	0.17	0.20
doc	**0.55**	**0.93**	**0.34**	**0.37**	**0.60**	0.92	**0.44**	**0.44**

Table 5: Accuracy on contrastive pronoun test set with regard to reference pronoun.

	Train-Paracrawl-Only		Train-Combined	
	inside current sentence	outside current sentence	inside current sentence	outside current sentence
sent	**0.68**	0.37	0.71	0.37
doc	0.67	**0.52**	**0.76**	**0.56**

Table 6: Accuracy on contrastive pronoun test set with regard to antecedent location.

	T-P-O		T-C	
	EnDe	DeEn	EnDe	DeEn
Paracrawl	26.88	31.37	27.35	31.65
Europarl	23.22	28.47	29.03	35.61
TED	24.24	28.04	25.62	29.91
News	31.86	34.92	33.41	35.99

Table 7: Translation results (BLEU) of using the document-level G-Transformer for translating single sentences. T-P-O: Train-Paracrawl-Only. T-C: Train-Combined.

T-P-O		T-C	
inside	outside	inside	outside
0.67	0.36	0.78	0.35

Table 8: Accuracy on contrastive pronoun test set with regard to antecedent location when using the document-level G-Transformer for translating single sentences without context.

3.4 Training with Only Paracrawl

We trained both sentence-level Transformers and document-level G-Transformers with only Paracrawl (see Table 1) as training data. Parallel paragraphs were used as parallel documents to train document-level G-Transformers and parallel sentences contained in parallel paragraphs were used to train sentence-level Transformers.[10] For evaluation, we used test data from 4 datasets, Paracrawl, Europarl, TED and News. For Europarl, TED and News, we used the same test data following the original G-Transformer (Bao et al., 2021) work, i.e., the Europarl, TED and News test sets are parallel documents in contrast to the Paracrawl test set which is parallel paragraphs. BLEU and COMET scores are given in Table 4 as Train-Paracrawl-Only. The document-level G-Transformers achieved higher translation quality than sentence-level Transformers for all 4 test sets and only used Paracrawl as training data, which demonstrates that Paracrawl can provide useful document-level information for effective training of document-level NMT models. Table 4 also shows that the document-level information contained in Paracrawl is robust across domains as document-level G-Transformers trained with only Paracrawl data can help to translate real documents from TED, News and Europarl test sets.

Table 5 and Table 6 give results of targeted pronoun evaluation. Results show that the document-level model, trained with only Paracrawl data, significantly outperformed the sentence-level model for pronoun translation especially when the an-

[10]Therefore, training data for the document-level model and the sentence-level model contain the same number of sentence pairs.

tecedent location of a pronoun is outside of the current sentence (see Table 6), which again demonstrates that Paracrawl can provide useful information outside of the current sentence for effective learning of document-level NMT.

3.5 Training with Paracrawl, TED, News and Europarl Combined

We also trained translation models with training data from Paracrawl, Europarl, TED and News combined. The sentence-level Transformers were trained with sentence pairs from all the 4 training datasets. The document-level G-Transformers were trained with parallel paragraphs from Paracrawl and parallel documents from Europarl, TED and News. We then computed BLEU and COMET scores for the 4 test sets as shown in Table 4 as Train-Combined. The targeted pronoun evaluation results are also given in Table 5 and Table 6 as Train-Combined. Results show that using additional parallel documents together with parallel (Paracrawl) paragraphs as training data further improved the general translation quality (BLEU and COMET) and also helped the document-level model to obtain a higher accuracy for targeted pronoun translation evaluation.

3.6 Document-level G-Transformer for Translating Single Sentences

We also evaluated how the document-level G-Transformer (trained with parallel documents) performs for translating single sentences (i.e. each sentence is considered as a single document at test time). Table 7 shows that, when considering the general translation quality (e.g. BLEU), document-level G-Transformers can perform well for translating single sentences, even outperforming the sentence-level Transformers in Table 4 for most of the test sets. However, the targeted pronoun evaluation results in Table 8 show that the accuracy of translating pronouns with antecedent location outside of the current sentence dropped more than 10% compared to the document-level model in Table 6, which demonstrates that document-level G-Transformers indeed require context outside of the current sentence for improving pronoun translation.

4 Conclusion

As document-level translation lacks large-scale general-domain document-level training data, we examine the effectiveness of using Paracrawl data for learning document-level translation. Paracrawl was officially released as parallel sentences extracted from parallel webpages. In this work, we extract parallel paragraphs from Paracrawl aligned webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level NMT models. We show that document-level models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level Transformers and also improving context-aware pronoun translation.

Limitations

Compared to the officially released Paracrawl corpus (278M sentence pairs for German-English), the size of parallel paragraphs that we extracted from Paracrawl is still relatively small. This is because (i) Paracrawl only released a subset of all webpage-level aligned texts and we only extracted parallel paragraphs from these released webpage texts (ii) we used very strict rules for extracting parallel paragraphs from Paracrawl and many parallel sentences were discarded by our extraction rules. For future work, we will collect more Paracrawl webpage-level aligned data for parallel paragraph extraction and we will test more flexible extraction rules.

Acknowledgements

The authors acknowledge the financial support by the German Federal Ministry for Education and Research (BMBF) within the projects “CORA4NLP" 01IW20010 and “KI-Servicezentrum Berlin Brandenburg" 01IS22092.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Domenic Donato, Lei Yu, and Chris Dyer. 2021. Diverse pretrained context encodings improve document 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 1299–1311, Online. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. 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, Online. Association for Computational Linguistics.

Inigo Jauregi Unanue, Nazanin Esmaili, Gholamreza Haffari, and Massimo Piccardi. 2020. Leveraging discourse rewards for document-level neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4467–4482, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation

training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

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

Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8435–8448, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

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

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization. Main track.