Combining Local and Document-Level Context: The LMU Munich Neural Machine Translation System at WMT19

Dario Stojanovski and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
{stojanovski,fraser}@cis.lmu.de

Abstract

We describe LMU Munich's machine translation system for English→German translation which was used to participate in the WMT19 shared task on supervised news translation. We specifically participated in the document-level MT track. The system used as a primary submission is a context-aware Transformer capable of both rich modeling of limited contextual information and integration of large-scale document-level context with a less rich representation. We train this model by fine-tuning a big Transformer baseline. Our experimental results show that document-level context provides for large improvements in translation quality, and adding a rich representation of the previous sentence provides a small additional gain.

1 Introduction

In this paper we describe the system we developed at the LMU Munich Center for Information and Language Processing, which we used to participate in the news translation task at WMT19. We submitted system runs for the English→German translation direction and specifically focus on the document-level translation track. The goal of the document-level track is to train machine translation models capable of taking into account larger context or even entire documents when translating sentences.

Supervised NMT has achieved state-of-the-art results (Bahdanau et al., 2015; Vaswani et al., 2017). Several works have claimed translation quality on a level similar to human translation. Wu et al. (2016) report translation quality on par with average bilingual human translators and Hassan et al. (2018) argue for parity to professional human translators on news translation from Chinese to English. However, these claims have been challenged in several ways with recent work (Läubli et al., 2018; Toral et al., 2018). One challenge is that these evaluations were done without giving evaluators access to the whole document-level context. They further show that human translations are preferred over automatic ones if evaluators are given document-level context. This is precisely the motivation for the document-level MT track in this year's WMT19.

One of the reasons for the failure of NMT in these context-dependent cases is not being able to model discourse-level phenomena. The straightforward reason for this is that traditional NMT does not have access to the context. As a result, it fails to account for several discourse-level phenomena, prominent ones being coreference resolution and coherence.

Coreference resolution has a particular impact on English→German translation, specifically for pronoun translation. English has only one third person singular pronoun that is routinely used for non-human references ("it"), while German has three, each representing a specific gender: masculine, feminine and neuter. Consider the following sentence: *We know it won't change students' behaviour instantly.* The translation of *it* into German can be, *er*, *sie* or *es* depending on the gender of the noun the English *it* is referencing. Since traditional NMT is working on the sentence-level, it has no way of ascertaining the appropriate gender and usually falls back to the data-driven prior, which is the neuter *es*.

Coherence is important in order to provide coherent translations across the whole given document. It is usually undesirable to produce translations with different meanings within a single document for the same ambiguous word.

Taking into account the whole document when generating translations will address some of the relevant discourse-level phenomena. An implicit effect that one could expect by modeling the whole

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 400–406
Florence, Italy, August 1-2, 2019.

document is also modeling the underlying domain. On an abstract level, one can presume that this is happening in sentence-level models as well, however access to larger context is likely to improve the ability to implicitly identify the domain. Domain adaptation and multi-domain NMT have been extensively studied (Kobus et al., 2017; Freitag and Al-Onaizan, 2016; Farajian et al., 2017; Sajjad et al., 2017; Zhang and Xiong, 2018; Chen et al., 2017; Tars and Fishel, 2018). However, most previous works assume that the domain of each sentence is known at training time, which is often not the case.

Taking into consideration different discourse-level phenomena, we develop a Transformer (Vaswani et al., 2017) which can richly model the previous sentence, but also takes advantage of larger context. We borrow on previous work on context-aware NMT (Stojanovski and Fraser, 2018; Voita et al., 2018; Miculicich et al., 2018; Zhang et al., 2018) and add additional parameters in the encoder and decoder to account for the previous sentence. We limit the context since we want this part of the model to be able to do coreference resolution which very often can be addressed by looking at the first previous sentence. We additionally take the 10 previous sentences and create a simple document representation by averaging their embeddings. This embedding is subsequently added to each source token in the sentence to be translated in the same fashion as positional embeddings are added to the token-level embeddings in the Transformer. We assume that this representation can help provide a clear domain signal.

The remainder of the paper outlines the model in detail, and presents the experimental setup and obtained results.

2 Related Work

There are large number of works in NMT focusing on integrating document-level information into otherwise sentence-level models (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Tu et al., 2018; Maruf and Haffari, 2018). These works have shown that improvements in pronoun translation are achieved by better handling coreference resolution. Smaller improvements are observed for coherence and cohesion. The main intuition behind the models in these works is that they employ an additional encoder for contextual sentences and integrate the information in the encoder or decoder using a gating mechanism. Our model is similar to the context-aware Transformer models proposed in these works with some specifics which we discuss in Section 3.

We also extend the Transformer model with a simple document representation which we assume provides for a domain signal. This could be useful for domain disambiguation and improved coherence and cohesion. This model is similar to previous work on domain adaptation for NMT (Kobus et al., 2017; Tars and Fishel, 2018) where special domain tokens are either added to the beginning of the sentence or concatenated as additional features to the token-level embeddings. However, they assume a set of known domains in advance which is not the case in our work. We model the domain implicitly.

3 Model

In this work we develop two models: a previous-sentence and document-level context-aware Transformer. For our primary submission, we use a joint model combining both approaches into a single model. We use source side context only, both at training and testing time.

3.1 Previous-sentence context-aware Transformer

This context-aware model is in line with previous works on context-aware NMT (Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Zhang et al., 2018). The standard Transformer is extended to be able to receive an additional sentence as input. In this work we only use the first previous sentence. We feed this context sentence through the Transformer encoder. As suggested in Voita et al. (2018), we share the encoder for the main and context sentence. In order to provide information as to what is being encoded, we add a special token at the beginning of the context sentence. We share the encoder layers up to and including the penultimate layer. Unlike Voita et al. (2018), we do not integrate the context in the encoder, but rather in the decoder. As a result, the last encoder layer is the standard Transformer encoder, but it is not shared across the main and context sentence.

We modify the decoder by adding an additional

multi-head attention (MHA) sublayer on the context representation. As in the standard Transformer decoder layer, at training time, we first compute self-attention over the target sentence and use this to compute the MHA representation c_i over the main sentence. The output of this step is used to condition the MHA c_i^c over the context. Subsequently, the outputs of the MHA over the main and context representations, c_i and c_i^c, are merged using a gated sum. The use of the gate is similar to previous work (Wang et al., 2017; Voita et al., 2018). It is conditioned on c_i and c_i^c. The output is computed as follows:

$$s_i = g_i \otimes c_i + (1 - g_i) \otimes c_i^c$$

and the gate is computed as:

$$g_i = \sigma(W_e c_i + W_c c_i^c)$$

where σ represents sigmoid activation and $\otimes$ element-wise multiplication. The gate enables the model to control how much information should be used from the main sentence and from the context sentence. Finally, the output of the gated sum is passed through a feed-forward neural network.

3.2 Document-level context-aware Transformer

We also extend the model with the ability to consume larger context. Miculicich et al. (2018) proposed a model capable of using large context using hierarchical attention. They tackle the memory requirements of such models by reusing already computed sentence representations. This introduces limitations as to how the random batching usually used to train NMT works, since it is necessary to have the previous sentences of a given sentence in a document already processed. Furthermore, Miculicich et al. (2018) report that they fail to obtain significant improvements as the context increases. They do not improve results beyond context sizes of 2 or 3 sentences.

As a result, we make a simple modification to the Transformer which enables it to handle large context sizes. In this work we use up to 10 sentences of context, all of which are previous sentences (but it would also be possible to use the following sentences as well). We take the embeddings of all tokens within the context and simply average them. This averaged document representation is then passed through a feed-forward network. The final document-level representation is then added to all token-level source embeddings in the sentence to be translated in the same manner as the positional embeddings are added in the Transformer. A similar approach was proposed by Kobus et al. (2017) for domain adaptation in RNN-based NMT. The work differs since they have special tokens which indicate the domain and they concatenate them instead of adding them to the token-level embeddings. Our approach is more flexible since it only relies on having access to contextual information and does not require explicit domain knowledge. Our intuition with this approach is that the document representation should be informative of the type or domain of the document being translated.

We share all source, target, output and context embeddings. We freeze them in the continued training phase with the context-aware model in order for the model to be more memory efficient.

4 Experimental Setup

4.1 Preprocessing

The data is preprocessed by normalizing punctuation, tokenizing and truecasing with the scripts from Moses. We apply BPE splitting (Sennrich et al., 2016b) with 32K merge operations. BPE is computed jointly on both languages.

Corpus	sentences
CommonCrawl	2.1M x2
Europarl	1.5M x2
NewsCommentary	0.3M x2
Rapid	1.4M x2
WikiTitles	1.3M x2
ParaCrawl	13.5M
NewsCrawl	9.3M
NewsCrawl v2	16.9M

Table 1: Training data sizes after filtering. x2 - oversampling factor.

4.2 Data filtering

Samples where the length of the source, target or first previous sentence before BPE-splitting is over 50 tokens are removed. For the purposes of our document-level model, we also use larger context. In our experiments, we restrict the model to access only the 10 previous sentences at most. Samples where the total length of these sentences exceeds 500 are also removed. After applying BPE splitting, an additional length filtering step

is applied with a maximum length allowed of 100 for the source, target and first previous sentence. Document-level context is limited to 800.

WMT provides the large ParaCrawl corpus which is very noisy. In previous years at WMT, high scoring systems showed that it is necessary to perform aggressive filtering. We reuse some of the data selection steps proposed in Stahlberg et al. (2018). We run language identification and remove non-English and non-German sentences. Furthermore, all sentences are removed where one of the following conditions is met: a word is over 40 characters long, HTML tags in text, sentence length less than 4 words, character ratio between source and target sentence is over 1:3 or 3:1, source or target sentence is not identical after removing non-numerical characters and sentence does not end in a punctuation mark. As a result, the size of the ParaCrawl corpus was reduced from 30M to 13.5M sentences. Unfortunately, due to time constraints, we were not able to reproduce the data filtering and data selection suggested by Junczys-Dowmunt (2018) which obtained the top BLEU scores at WMT18. They showed that the optimal number of sentences is 8M. We assume that the higher number of presumably noisy sentences is affecting our initial baseline.

4.3 Backtranslation

As shown in previous years, using backtranslations (Sennrich et al., 2016a) is essential for strong translation quality. We train a German→English small Transformer and use it to backtranslate NewsCrawl data. Due to time constraints, we were not able to use the backtranslated data in the initial training of the English→German model. As a result, we fine-tune the already trained baseline with the backtranslated data mixed in with the parallel WMT data.

4.4 Hyperparameters

We train a big Transformer as a baseline. Embedding and hidden dimension size in the encoder and decoder is 1024. All attention sublayers use dot product attention and have 16 attention heads. The size of the feed-forward neural networks is 4096. The hidden dimension size of the context-aware encoder and context attention sublayer in the decoder is 512. All context-related attention sublayers have 8 attention heads. All models have 6 encoder and decoder layers. We use sinusoidal positional embeddings which are added to the token-level embeddings. In the case of the document-level model, we further add the average of all large-context embeddings. We apply residual dropout of 0.1 as in (Vaswani et al., 2017). Additionally, dropout of 0.1 is applied to the multi-head attention and feed-forward network. We also use label smoothing of value 0.1.

4.5 Training

We train the Transformer baseline with a warmup period and a learning rate of 10^{-4}. In all cases of continued training in the paper, we set the learning rate to 10^{-5}. We train the models with early-stopping based on the perplexity on the development set. We checkpoint the model every 4000 updates. The learning rate is reduced by a factor of 0.7 if no improvements are observed for 8 checkpoints. Training converges if no improvements are observed after 32 checkpoints. We train our context-aware models by continued training on the converged baseline. All parameters relating only to the context-aware parts of the architecture are randomly initialized. The batch size is set to 4096 tokens.

Model	parameters
baseline	217M
previous-sentence context	253M
document-level context	225M
joint model	261M

Table 2: Number of model parameters. All models are big Transformer models.

The number of parameters for all models are presented in Table 2. We train the models on 4 GTX 1080 Ti GPUs with 12GB RAM. We use Sockeye[1] (Hieber et al., 2018) to train the baseline and our context-aware models.

5 Empirical Evaluation

We present the results we obtain with our models in Table 3. We report results on the English→German newstest2017, newstest2018 and newstest2019. We report BLEU scores using sacreBLEU[2] (Post, 2018) on detokenized text. For the final submission, we processed quotation marks to match the German style.

We train our baseline on the data presented in Table 1. We initially train on the ParaCrawl

[1] https://github.com/awslabs/sockeye
[2] https://github.com/mjpost/sacreBLEU

dataset and an oversampled version of the other datasets. We train this baseline until convergence with early-stopping based on the perplexity on the development set. As a development set, we use newstest2018. After convergence, we fine-tune with 9.3M NewsCrawl backtranslations in addition to the dataset we used for the initial baseline. This baseline is used to initialize all the other context-aware models. It is interesting to observe that fine-tuning with NewsCrawl backtranslations and WMT data improves on newstest2017 and newstest2018, but significantly decreases the BLEU score on newstest2019.

	en→de		
Model	nt17	nt18	nt19
baseline	29.8	45.3	39.5
baseline*	30.3	45.6	38.5
previous-sentence*	30.5	46.0	38.6
document-level*	30.5	45.7	39.3
document-level	**31.1**	47.0	40.0
joint	**31.1**	**47.1**	**40.3**

Table 3: BLEU scores on newstest2017, newstest2018 and newstest2019. * - model trained with NewsCrawl backtranslations. All context-aware models fine-tuned on baseline*.

For training the context-aware models, we ignore the ParaCrawl data and use the remaining datasets. Depending on the setup, we either use the 16.9M NewsCrawl backtranslations with document boundaries or completely ignore them. Our previous sentence context-aware Transformer trained with NewsCrawl backtranslations do not provide for significant improvements. It increases the BLEU score from 38.5 to 38.6. However, the document-level model with averaging context embeddings obtains a BLEU score of 39.3.

We also remove the NewsCrawl backtranslations when fine-tuning our average context embedding Transformer. This proves to be very helpful and we manage to obtain 40.0 BLEU. It is interesting that this model also substantially improves the BLEU score on newstest2017 and newstest2018. One possible explanation of the adverse effect of using backtranslations is that our document-level model is more sensitive to noisy input. We leave a further examination of the issue for future work.

Finally, we train a joint model where we combine the average context embedding approach with the previous-sentence context-aware Transformer where we employ a separate encoder and modify the decoder. This further pushes the BLEU score to 40.3 on newstest2019 and slightly improves results on the other test sets. This is the system we used for the primary submission.

We also tried ensembling context-aware joint models. However, due to time constraints we only managed to train a single baseline. Therefore, all context-aware models were trained by fine-tuning on top of the single baseline. As a result, these models were not diverse enough and ensembling did not help. After the evaluation period, we also tried averaging the last 5 checkpoints of a single run of the joint model. This improved the score on newstest2019 to 40.8 BLEU.

6 Conclusion

In this work, we presented our system which we used to participate in the English→German news translation task at WMT19. We proposed two modifications to the standard Transformer architecture. We propose a context-aware Transformer which has a separate encoder and a modified decoder in order to provide for a fine-grained access to a limited context. We further extend this model by proposing to average the context token-level embeddings and add them to the main sentence embeddings. This enables access to large scale context. We show that the latter modification provides for large improvements with regards to a baseline and that combining both approaches leads to a further performance increase.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable input. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15.

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.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137. Association for Computational Linguistics.

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

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

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 (Volume 1: Research Papers)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

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. 2018. Microsoft's submission to the WMT2018 news translation task: How i learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 429–434, Brussels, Belgium. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378. INCOMA Ltd.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. 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.

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. Association for Computational Linguistics.

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.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. *CoRR*, abs/1708.08712.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Felix Stahlberg, Adri de Gispert, and Bill Byrne. 2018. The University of Cambridges machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 508–516, Brussels, Belgium. Association for Computational Linguistics.

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, Volume 1: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

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.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 113–123, Brussels, Belgium. 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. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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. Association for Computational Linguistics.

Shiqi Zhang and Deyi Xiong. 2018. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190. Association for Computational Linguistics.