Pretrained Language Models for Document-Level Neural Machine Translation

Liangyou Li and **Xin Jiang** and **Qun Liu**
Huawei Noah's Ark Lab
liliangyou@huawei.com

Abstract

Previous work on document-level NMT usually focuses on limited contexts because of degraded performance on larger contexts. In this paper, we investigate on using large contexts with three main contributions: (1) Different from previous work which pertrained models on large-scale sentence-level parallel corpora, we use pretrained language models, specifically BERT (Devlin et al., 2018), which are trained on monolingual documents; (2) We propose context manipulation methods to control the influence of large contexts, which lead to comparable results on systems using small and large contexts; (3) We introduce a multi-task training for regularization to avoid models overfitting our training corpora, which further improves our systems together with a deeper encoder. Experiments are conducted on the widely used IWSLT data sets with three language pairs, i.e., Chinese–English, French–English and Spanish–English. Results show that our systems are significantly better than three previously reported document-level systems.

1 Introduction

Recently, document-level Neural Machine Translation (NMT) is drawing more attention from researchers studying on incorporating contexts into translation models (Wang et al., 2017; Jean et al., 2017; Tiedemann and Scherrer, 2017; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Kuang et al., 2018; Tu et al., 2018). It has been shown that nmt can be improved by taking document-level context information into consideration. However, one of the common practices in previous work is to only consider very limited contexts (e.g., two or three sentences) and therefore long dependencies in documents are usually absent during modeling the translation process. Although previous work has shown that when increasing the length of contexts, system performance would be degraded, to the best of our knowledge, none of them addresses the problem this work considers.

Given the importance of long-range dependencies (Dai et al., 2019), in this paper we investigate approaches to take large contexts (up to 512 words in our experiments) into consideration. Our model is based on the Transformer architecture (Vaswani et al., 2017) and we propose methods to narrow the performance gap between systems using different lengths of contexts. In summary, we make three main contributions:

- We use pretrained language models (PLMs) to initialize parameters of encoders. Different from pretrained models on large-scale sentence-level parallel corpora (Tu et al., 2018; Zhang et al., 2018), PLMs are trained on monolingual documents which are easier to obtain than bilingual corpora.
- We propose methods to manipulate the integration of context information to control the influence of large contexts. In our experiments, these methods lead to comparable results on systems using small and large contexts.
- We introduce a multi-task training which adds an extra task on the encoder side regularizing our model and further improving our systems together with a deeper encoder.

Experimental results on the widely used IWSLT data sets (Cettolo et al., 2012) show that our final systems significantly outperform systems in previous work on three language pairs, i.e., Chinese–English (Zh-En), French–English (Fr-En) and Spanish–English (Es-En). Our results also demonstrate the necessity of PLMs and usefulness of the multi-task training and the context manipulation methods.

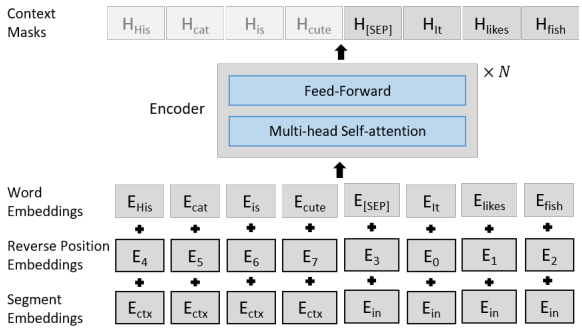

Figure 1: The proposed encoder structure. An input sentence "*It likes fish*" is concatenated with its contexts "*His cat is cute*". A separation mark "*[SEP]*" is inserted between them. Compared with the original Transformer architecture, we make the following changes: (1) *segment embeddings* are added to distinguish contexts from the input; (2) *reversed position embeddings* are introduced as an alternative of the original sequential position embeddings; (3) *context masks* are used during decoding to avoid attention weights on the contexts. In this figure, each "E_*" denotes an embedding vector, while each "H_*" denotes an output vector of the encoder.

2 Document-Level NMT

In this paper, we consider source contexts, i.e., sentences before the current input to be translated. Following Tiedemann and Scherrer (2017), the current input and its contexts are concatenated, as shown in Figure 1. Instead of defining an additional hyperparameter on the length of contexts (e.g., k sentences), we set the maximum total length of an input and its contexts to 512 words, which is defined by the model capacity (i.e., the maximum input length).

However, incorporating such large contexts could result in unstable training and introduces much irrelevant information. To alleviate these problems, we (1) use PLMs trained on large-scale monolingual documents to initialize parameters of encoders; (2) propose a few changes of the encoder architecture to control impacts of contexts; and (3) introduce a multi-task training mechanism to regularize our model. Figure 1 shows the input format and encoder architecture we use in this paper.

2.1 Pretrained Language Models

Because large-scale parallel corpora with document boundaries are usually unavailable, researchers have tried to make use of sentence-level parallel corpora to help training a document-level NMT model (Zhang et al., 2018; Tu et al., 2018). Different from them, we use large-scale monolingual data which is much easier to obtain, e.g., from Wikipedia or other public websites. Instead of training a language model from scratch on monolingual documents, we directly use pretrained BERT models to initialize our encoder and then fine-tune our NMT model on document-level parallel corpora. BERT is chosen based on the following reasons: (1) BERT is based on the Transformer architecture considering bidirectional contexts which makes it compatible with our encoder; (2) BERT is trained over long sequences and learns relationships between them, and thus is suitable to model document-level contexts; (3) BERT codes and multilingual models are publicly available, which makes it easier to replicate our results.

2.2 Context Manipulation

Unlike previous work which uses additional components, such as context encoders or attention layers, to encode and integrate contexts, in this paper we use a single encoder on the concatenation of an input and its contexts. Although a model can be directly trained without any modification (Tiedemann and Scherrer, 2017), we found it does not work well when large contexts are used especially without PLMs. We presume this is because large contexts introduce much more irrelevant information which would overwhelm the source sentence to be translated. To alleviate this problem, we introduce three techniques into the encoder to explicitly make a distinction between contexts and the input.

2.2.1 Segment Embeddings

The concept of *segment embeddings* is introduced in BERT. The basic idea is that each sequence has a unique embedding so as to distinguish from other sequences. In this paper, we directly adapt the idea into the encoder and add different segment embeddings to contexts and the source sentence, respectively.

2.2.2 Reversed Position Embeddings

By default, position embeddings in the Transformer are assigned words by words. However, when contexts are concatenated to an input sentence, position embeddings of the source input will depend on length of the contexts which precede the source. To alleviate this, we propose to first assign position embeddings to the source input and then to its contexts, called *reversed position embeddings*, as illustrated in Figure 1, which keeps the positional

representations of source sentences stable.

To alleviate this problem, we propose to first assign position embeddings to the source input and then to its contexts, called *reversed position embeddings*, as illustrated in Figure 1, which keeps the positional representations of source sentences stable.

2.2.3 Context Masks

Tiedemann and Scherrer (2017) has shown that directly augmenting an input with its contexts improves translation quality in RNN-based models. However, they only use the immediately previous sentence as contexts in experiments. When larger contexts are used, this kind of method does not work well because large contexts could result in unstable training and it would be more challenging for the model to learn appropriate attention weights to distinguish contexts and inputs. Therefore, in this paper we add *context masks* to avoid the decoder attending to the contexts as we presume representations of the source part are already context-aware through the underlying self-attention in the encoder.

2.3 Multi-task Training

Inspired by Ramachandran et al. (2017), we introduce an extra task on the encoder side to avoid the model overfitting our training corpus. The task we use is called masked language model (MLM) prediction which is also used to train BERT. When MLM is considered, the training objective becomes:

$$\hat{\theta} = \arg\max_{\theta} \sum_{(X,C,Y)\in D} \log P(Y|S) + \sum_{k\in M} \log P(s_k|S),$$

where S is the extended input by combining the input X and context C but with some words randomly masked (about 16% with a maximum number of 20 words), M is the set of masked positions, s_k is the real word form at position k.

3 Experiments

3.1 Data sets and Settings

We conduct experiments on the widely used IWSLT data sets with three language pairs: Zh-En, Fr-En and Es-En, each of which contains around 0.2M sentence pairs. We use *dev2010* for development. *tst2010-2013* (Zh-En), *tst2010* (Fr-En) and *tst2010-2012* (Es-En) are used for testing.

The size of our baseline NMT model follows that of BERT-base models. We train models up to 300K steps with each batch around 3072 source or target tokens. Adam (Kingma and Ba, 2015) is used to optimize parameters with the same learning rate as the original Transformer. We directly use pretrained Chinese and multilingual BERT models[1] to initialize encoders for Zh, Fr and Es, respectively. Beam search is used with a beam width of 4 and a length penalty (Wu et al., 2016) of 1.

3.2 Overall Results

In this section, we compare our document-level NMT systems with three existing approaches (Tu et al., 2018; Miculicich et al., 2018; Zhang et al., 2018). To ensure a fair comparison, all systems are based on the Transformer architecture, and our translations are processed and evaluated following these papers as well.

Table 1 shows the overall evaluation results on test sets in all three language pairs. Our systems with BERT and context manipulation methods achieve significantly better BLEU scores than previous work. Specifically, gains on Zh-En, Fr-En and Es-En are 2.80 BLEU, 2.06 BLEU and 2.84 BLEU, respectively. We also found using a deeper (12 layers) encoder improves systems compared to a shallower (6 layers) encoder (by up to 0.49 BLEU on Zh-En, 0.64 BLEU on Fr-En, and 0.55 BLEU on Es-En, respectively). When we introduce the MLM task into the deep model, the systems is further improved.

3.3 Ablation Study

Despite the overall improvements shown in Table 1, it would be interesting to know the contribution of each method we applied. Therefore, in this section, we conduct ablation study with results shown in Table 2. We found that simply using the BERT to initialize parameters of the encoder improves the baseline system by 1.04 BLEU. When contexts are concatenated with source sentences which are then directly taken as inputs of the model without any changes in the network structure, the system (i.e., the one with "+*Large Context*") is further improved by 1.31 BLEU. Finally, when the three context manipulation methods are integrated, the system achieves the best BLEU score.

[1] https://github.com/google-research/bert

Table 1: BLEU scores and increment over the best previous approach on three language pairs. "*+Large Context Manipulation*" denotes the three context manipulation methods on large contexts. "*+Encoder-12*" means to increase the number of encoder layers to 12. "*+MLM*" means adding the MLM prediction task into the training objective. The best BLEU scores are in bold.

	Systems	**Zh–En**	**Fr–En**	**Es–En**
previous	(Tu et al., 2018)	17.32	-	36.46
	(Miculicich et al., 2018)	17.79	-	37.24
	(Zhang et al., 2018)	-	36.04	-
our work	Baseline	17.31	35.33	37.01
	+BERT +Large Context Manipulation	20.10 (+2.31)	37.46 (+1.42)	39.53 (+2.29)
	+Encoder-12L	20.59 (+2.80)	38.10 (+2.06)	40.08 (+2.84)
	+MLM	**20.72 (+2.93)**	**38.76 (+2.72)**	**40.31 (+3.07)**

Table 2: BLEU scores and increment over the baseline on dev2010 in ablation study. "*+Large Context*" means large contexts are simply concatenated with inputs following Tiedemann and Scherrer (2017). *CtxMask*, *SegEmb* and *RevPos* denote Context Masks, Segment Embeddings and Reverse Position Embeddings, respectively.

Systems	**Zh–En**
Baseline	12.19
+BERT	13.23 (+1.04)
+Large Context	14.54 (+2.35)
+CtxMask	14.87 (+2.68)
+SegEmb	15.00 (+2.81)
+RevPos	15.30 (+3.11)

Table 3: BLEU scores and increment over the baseline on dev2010 when context length is varied. *Small context* denotes the immediately previous sentence. *+Manipulation* means the three context manipulation techniques.

Systems	**Zh–En**
Baseline	12.19
+Small Context	12.29 (+0.1)
+Large Context	Diverge
+BERT	13.23 (+1.04)
+Small Context	15.54 (+3.35)
+Large Context	14.54 (+2.35)
+BERT +Manipulation	-
+Small Context	15.50 (+3.31)
+Large Context	15.30 (+3.11)

3.4 Context Length

Table 3 shows results of varying context length. We found that NMT models with (especially large) contexts considered do not work well without pretraining. When parameters are initialized with BERT, both small and large contexts bring significant improvements even without using the three manipulation methods. This suggests the importance of pretraining when document-level parallel corpora are in small-scale and is consistent with findings in previous work (Tu et al., 2018; Zhang et al., 2018). However, a difference is that they pretrained models on sentence-level parallel corpora, which we think could help to further improve our systems.

Another finding is that using smaller contexts achieves a significantly better BLEU score (+1.0) than larger contexts, similar to Zhang et al. (2018); Miculicich et al. (2018). However, when our manipulation methods are applied, the system with large contexts is further improved resulting a narrowed gap (0.2 BLEU difference) between it and the system using small contexts. We also found that our manipulation methods does not improve the system with small contexts. This is expected since they are designed for controlling the influence of large contexts. Our results suggest that sophisticated manipulation on the integration of large contexts is necessary and promising to achieve a better performance.

4 Conclusion

In this paper, we investigate document-level NMT using large contexts. We (1) use pretrained language models, i.e. BERT, to initialize the encoder; (2) propose manipulation methods to control the influence of large contexts; and (3) introduce a multi-task training mechanism for model regularization. Experiments on IWSLT data sets showed that our systems achieved the best BLEU scores compared with previous work on Chinese–English, French–English and Spanish–English.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261—268, Trento, Italy.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

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

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.

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, Brussels, Belgium. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

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, Copenhagen, Denmark. 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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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, Łukasz 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, Brussels, Belgium. Association for Computational Linguistics.