Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation

Bei Li[1], Hui Liu[1], Ziyang Wang[1], Yufan Jiang[1],
Tong Xiao[1,2]*, Jingbo Zhu[1,2], Tongran Liu[3], Changliang Li[4]

[1]NLP Lab, Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
[3]CAS Key Laboratory of Behavioral Science, Institute of Psychology, CAS, Beijing, China
[4]Kingsoft AI Lab, Beijing, China

```
{libei_neu,jiangyufan2018}@outlook.com,
{huiliu,wangziyang}@stumail.neu.edu.cn,
{xiaotong,zhujingbo}@mail.neu.edu.com,
liutr@psych.ac.cn,lichangliang@kingsoft.com
```

Abstract

In encoder-decoder neural models, multiple encoders are in general used to represent the contextual information in addition to the individual sentence. In this paper, we investigate multi-encoder approaches in document-level neural machine translation (NMT). Surprisingly, we find that the context encoder does not only encode the surrounding sentences but also behaves as a noise generator. This makes us rethink the real benefits of multi-encoder in context-aware translation - some of the improvements come from robust training. We compare several methods that introduce noise and/or well-tuned dropout setup into the training of these encoders. Experimental results show that noisy training plays an important role in multi-encoder-based NMT, especially when the training data is small. Also, we establish a new state-of-the-art on IWSLT Fr-En task by careful use of noise generation and dropout methods.

1 Introduction

Sentence-level neural machine translation (NMT) systems ignore the discourse phenomena and encode the individual source sentences with no use of contexts. In recent years, the context-aware models which learn contextual information from surrounding sentences have shown promising results in generating consistent and coherent translations (Zhang et al., 2018; Voita et al., 2018; Kim et al., 2019; Voita et al., 2019; Bawden et al., 2018; Miculicich et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019).

There are two common approaches to incorporating contexts into NMT: the simple way is to concatenate the context and the current sentence

to form a context-aware input sequence (Agrawal et al., 2018; Tiedemann and Scherrer, 2017), whereas a more widely-used approach utilizes additional neural networks to encode context sentences (Jean et al., 2017; Voita et al., 2018; Zhang et al., 2018). Here we name the former as the single-encoder approach and name the latter as the multi-encoder approach. However, large-scale document corpora are not easily available. Most context-aware NMT systems are evaluated on small datasets and significant BLEU improvements are reported (Wang et al., 2017; Zhang et al., 2018; Tu et al., 2018). In our experiments, we find that the improvement persists if we feed pseudo sentences into the context encoder, especially when we train the system on small-scale data. A natural question here is: *How much does the improvement come from the leverage of contextual information in multi-encoder*?

In this work, we aim to investigate what kinds of information that the context-aware model captures. We re-implement several widely used context-aware architectures based on the multi-encoder paradigm, and do an in-depth analysis to study whether the context encoder captures the contextual information. By conducting extensive experiments on several document-level translation benchmarks, we observe that:

- The BLEU gaps between sentence-level and context-aware models decrease when the sentence baselines are carefully tuned, e.g., proper use of dropout.

- The multi-encoder systems are insensitive to the context input. Even randomly sampled sentences can bring substantial improvements.

- The model trained with the correct context can achieve better performance during inference without the context input.
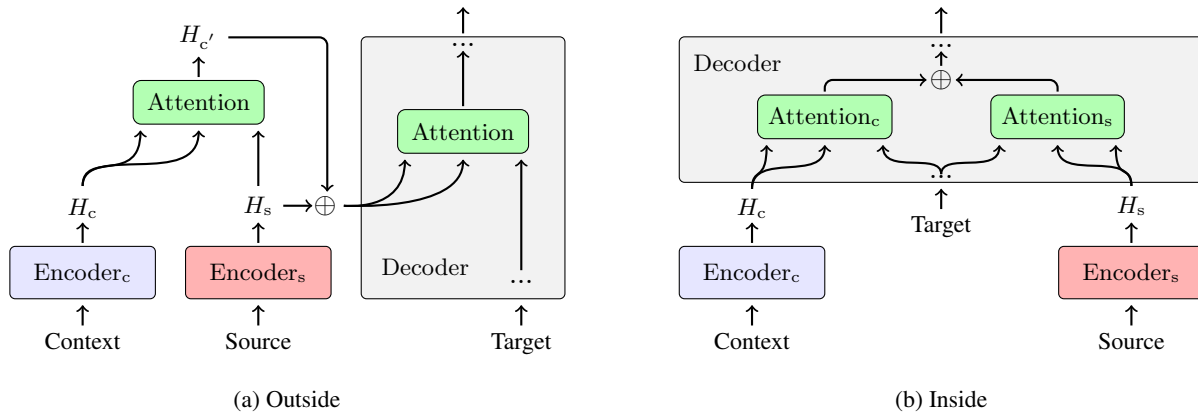
---

*Corresponding author.

Figure 1: An overview of two multi-encoder systems. In the Outside approach, H_s is the query and H_c is the key/value. In the Inside approach, Target is the query, H_s and H_c represent key/value.

Our contribution is two folds: (i) We find that the benefit of the multi-encoder context-aware approach is not from the leverage of contextual information. Instead, the context encoder acts more like a noise generator to provide richer training signals. (ii) The finding here inspires us to develop a simple yet effective training strategy: we add a Gaussian-noise to the encoder output, which can effectively alleviate the overfitting, especially on small datasets.

2 Approaches to Incorporating Contexts into NMT

Here we describe two ways of introducing contextual information into NMT systems.

2.1 The Single-Encoder Approach

The input of the single-encoder system is the concatenation of the context sentences and the current sentence, with a special symbol inserted to distinguish them (Tiedemann and Scherrer, 2017; Agrawal et al., 2018). Then the extended sentence is fed into the standard Transformer. These systems may face the challenge of encoding extremely long inputs, resulting in inefficient computation.

2.2 The Multi-Encoder Approach

The multi-encoder models take the surrounding sentences as the context and employ an additional neural network to encode the context, that is, we have a source-sentence encoder and a context encoder. Figure 1 shows two methods of integrating the context into NMT in the multi-encoder paradigm. Next we show that most of the multi-encoder approaches (Voita et al., 2018; Zhang et al., 2018)

are instances of the models described below.

- **Outside integration**. As shown in Figure 1(a), the representations of the context and the current sentence are firstly transformed into a new representation by an attention network. Then the attention output and the source sentence representation are fused by a gated sum.

- **Inside integration**. Alternatively, the decoder can attend to two encoders respectively (Figure 1(b)). Then, the gating mechanism inside the decoder is employed to obtain the fusion vector.

3 Experimental Setup

3.1 Data and Settings

We evaluated the document-level approaches on several publicly available datasets. For Chinese-English (Zh-En) and French-English (Fr-En), we used Ted talks from IWSLT15 and IWSLT16 (Cettolo et al., 2012) evaluation campaigns as the training data. We validated on *dev2010*, and tested on *tst2010-2013* (Zh-En), *tst2010* (Fr-En) respectively. For English-German (En-De), we evaluated on WMT18 task [1]. For more convincing results, we also randomly sampled 500k/1M/2M/5M sentence pairs from the Chinese-English corpus provided by WMT[2] and test on *newstest2017*. We preprocessed the sentences with Moses tokenizer[3] except Chinese sentences and used byte pair encoding (Sennrich et al., 2016) with 32K merged operations to

---

[1] We used the News-Commentary v14 as the train set
[2] http://www.statmt.org/wmt19/translation-task.html
[3] http://www.statmt.org/moses

Lang.	Train		Valid		Test	
	doc.	sent.	doc.	sent.	doc.	sent.
Zh-En	1708	209K	8	887	56	5473
Fr-En	1803	220K	8	887	11	1664
En-De	8462	329K	130	3004	122	2998
En-Ru	-	2M	-	10k	-	10k

Table 1: Details of datasets on different language pairs.

System	Layers	WS	TS	BLEU
Sentence-level	-	-	-	28.9
Outside *Context*	6	×	×	28.5
	6	✓	×	29.3
	6	×	✓	29.6
	1	×	✓	29.4

Table 2: Comparison of context-aware model with two training strategies on En-De task. WS represents weight-sharing and TS represents two-stage training.

segment words into sub-word units. The Chinese sentences were word segmented by the tool provided within NiuTrans (Xiao et al., 2012). For Fr-En and Zh-En tasks, we lowercased all sentences to obtain comparable results with previous work. We also conducted experiments on a larger English-Russian (En-Ru) dataset provided by Voita et al. (2018), consisting of 2M sentence pairs selected from publicly available OpenSubtitles2018 corpus. The data statistics of each language pair can be seen in Table 1. We chose the Transformer-base model as the sentence-level baseline. The context encoder also used the same setting as the sentence-level baseline.

We used Adam (Kingma and Ba, 2014) for optimization, and trained the systems on a single TiTan V GPU[4]. The learning rate strategy was the same as that used in Vaswani et al. (2017). Our implementation was based on Fairseq (Ott et al., 2019). More details can be found in our repository[5].

4 Results and Discussion

To study whether the context-encoder network captures contextual information in training, we present three types of context as the input of the context-encoder:

- *Context*: the previous sentence of the current sentence.

- *Random*: a sentence consisting of words randomly sampled from the source vocabulary.

- *Fixed*: a fixed sentence input for context-encoder.

4.1 Baseline Selection

Weight sharing (Voita et al., 2018) and two-stage training (Zhang et al., 2018) strategies have been proven essential to build strong context-aware systems. The former shared the first N-1 blocks of

context encoder with the source encoder, and the latter first trained a standard sentence-level Transformer and finetuned the document-level Transformer with an extra context-encoder. We first evaluated the importance of two training strategies for multi-encoder systems. We selected the multi-encoder with Outside integration (see Section 2) as the context-aware model and trained systems with two training strategies on the En-De task respectively. As shown in Table 2, we find that both two strategies outperform the sentence-level baseline by a large margin. The model with two-stage training performs slightly better than the weight-sharing system in terms of BLEU. To our surprise, the context-encoder with a single-layer can compete with a six-layers model. We suspect that this is because the training data is limited and we do not need a sophisticated model to fit it. Therefore, we choose the two-stage training and single-layer context-encoder for all experiments in the remainder of this paper.

4.2 Results

Table 3 shows the results of several context-aware models on different datasets. We see, first of all, that all multi-encoder models, including both Inside and Outside approaches outperform the sentence-level baselines by a large margin on the Zh-En and En-De datasets with a small p value of $dropout$. Also, there are modest BLEU improvements on the Fr-En and En-Ru tasks. When the models are regularized by a larger $dropout$, all systems obtain substantial improvements - but the gaps between sentence-level and multi-encoder systems decrease significantly.

We deduce that if the context-aware systems rely on the contextual information from the preceding sentence, the performance of *Random* and *Fixed* should dramatically decrease due to the incorrect context. Surprisingly, both *Random* and *Fixed* systems achieve comparable performance or even

---

System	Zh-En		Fr-En		En-De		En-Ru	
	$p=0.1$	$p=0.3$	$p=0.1$	$p=0.3$	$p=0.1$	$p=0.3$	$p=0.1$	$p=0.3$
Sentence-level	18.0	19.7	36.5	36.9	28.9	30.2	30.3	31.1
Single-encoder	18.1	19.1	36.2	37.3	28.5	30.2	30.4	31.2
Inside *Context*	19.4	20.0	36.8	37.5	29.7	31.0	30.8	31.3
Inside *Random*	19.5	20.3	37.0	37.4	29.9	30.7	30.8	31.4
Inside *Fixed*	19.5	20.3	37.0	37.2	29.3	30.8	30.8	31.4
Outside *Context*	19.4	19.8	36.8	37.4	29.4	30.7	30.9	31.1
Outside *Random*	19.4	20.1	36.8	37.3	29.6	31.1	30.7	31.1
Outside *Fixed*	19.4	20.0	36.7	37.2	29.5	31.1	30.8	31.1

Table 3: The BLEU scores [%] of different context-aware models with three context inputs. We use $dropout = 0.1$ and $dropout = 0.3$ respectively.

System	Inside		Outside	
	Aware	Agnostic	Aware	Agnostic
Context	31.0	31.0	30.7	31.1
Random	30.7	30.8	31.1	31.3
Fixed	30.8	30.8	31.1	31.1

Table 4: The BLEU scores [%] of context-aware systems with two inference schemas. Aware represents the inference process matches the training. Agnostic represents that models ignore context encoder during inference.

System	Zh-En	Fr-En	En-De	En-Ru
Baseline	19.7	36.9	30.2	31.1
Context	19.8	37.4	30.7	31.1
Noise	19.9	37.4	30.9	31.3
Context+Noise	19.9	37.3	30.9	31.3

Table 5: Comparison of Outside *Context* and Gaussian-noise methods on three tasks, with $dropout = 0.3$, $\sigma = 0.3$.

higher BLEU scores than *Context* in most cases (See Table 3). A possible explanation is that the context encoder does not only model the context. Instead, it acts more like a noise generator to provide additional supervised signals to train the sentence-level model.

4.3 Robust Training

To verify the assumption of robust training, we followed the work (Srivastava et al., 2014; Berger et al., 1996). We turned off the context-encoder during the inference process, and made the inference system perform as the sentence-level baseline. Table 4 shows that both *Context* and *Random* inference without context-encoder obtain modest BLEU improvements. This confirms that the information extracted by context-encoder just plays a role like introducing randomness into training (e.g., $dropout$), which is a popular method used in robust statistics. We argue that three types of context provide noise signals to disturb the distribution of the sentence-level encoder output. The BLEU improvements of both Outside and Inside are mainly due to the richer noise signals which can effectively alleviate the overfitting.

Inspired by Outside integration manner, we de-signed a simple yet effective method to regularize the training process: A Gaussian noise is added to the encoder output instead of the embedding (Cheng et al., 2018). We sample a vector $\epsilon \sim \mathbf{N}\left(0, \sigma^2 \mathbf{I}\right)$ from a Gaussian distribution with variance σ^2, where σ is a hyper-parameter. As seen in Table 5, the systems with Gaussian-noise significantly outperform the sentence-level baselines, and are slightly better than the Outside-context counterpart. Moreover, a natural question is whether further improvement can be achieved by combining the *Context* with the Gaussian-noise method. From the last line in Table 5, we observe no more improvement at all. The observation here convinced the assumption again that the context-encoder plays a similar role with the noise generator.

4.4 Large Scale Training

Most previous results are reported on small training datasets. Here we examine the effects of the noise-based method on different sized datasets. We trained the Inside-Random model and the Gaussian-noise model on different datasets consisting of 500K to 5M sentence pairs. Seen from Figure 2, the baseline model achieves better translation performance when we increase the data size. More interestingly, it is observed that Inside-*Random* and Gaussian-noise perform slightly better than
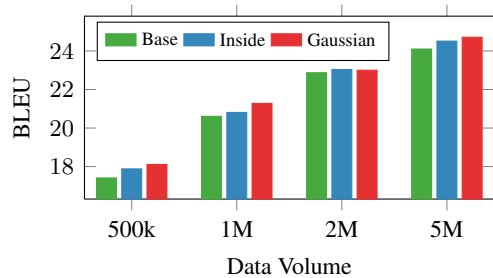
Figure 2: BLEU scores vs. different data volume on Zh-En sentence-level dataset. $dropout = 0.1$ and $\sigma = 0.3$.

the baseline, and the gaps gradually decrease with the volume increasing. This is reasonable that models trained on large-scale data may suffer less from the overfitting problem.

5   Related Work

Context-aware NMT systems incorporating the contextual information generate more consistent and coherent translations than sentence-level N-MT systems. Most of the current context-aware NMT models can be classified into two main categories, single-encoder systems (Tiedemann and Scherrer, 2017) and multi-encoder systems (Jean et al., 2017; Voita et al., 2018; Zhang et al., 2018). Voita et al. (2018) and Zhang et al. (2018) integrated an additional encoder to leverage the contextual information into Transformer-based NMT systems. Miculicich et al. (2018) employed a hierarchical attention network to model the contextual information. Maruf and Haffari (2018) built a context-aware NMT system using a memory network, and Maruf et al. (2019) encoded the whole document with selective attention network. However, most of the work mentioned above utilized more complex modules to capture the contextual information, which can be approximately regarded as multi-encoder systems.

For a fair evaluation of context-aware NMT methods, we argue that one should build a strong enough sentence-level baseline with carefully regularized methods, especially on small datasets (Kim et al., 2019; Sennrich and Zhang, 2019). Beyond this, Bawden et al. (2018) and Voita et al. (2019) acknowledged that BLEU score is insufficient to evaluate context-aware models, and they emphasized that multi-encoder architectures alone had a limited capacity to exploit discourse-level context. In this work, we take a further step to explore the main cause, showing that the context-encoder acts more like a noise generator, and the BLEU improve-

ments mainly come from the robust training instead of the leverage of contextual information. Additionally, Cheng et al. (2018) added the Gaussian noise to word embedding to simulate lexical-level perturbations for more robust training. Differently, we added the Gaussian noise to the encoder output which plays a similar role with context-encoder, which provides additional training signals.

6   Conclusions

We have shown that, in multi-encoder context-aware NMT, the BLEU improvement is not attributed to the leverage of contextual information. Even though we feed the incorrect context into training, the NMT system can still obtain substantial BLEU improvements on several small datasets. Another observation is that the NMT models can even achieve better translation quality without the context encoder. This gives us an interesting finding that the context-encoder acts more like a noise generator, which provides rich supervised training signals for robust training. Motivated by this, we significantly improve the sentence-level systems with a Gaussian noise imposed on the encoder output. Experiments on large-scale training data demonstrate the effectiveness of this method.

Acknowledgments

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides.

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.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy ap-

proach to natural language processing. *Computational Linguistics*, 22(1):39–71.

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.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

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

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, Hong Kong, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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.

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, Minneapolis, Minnesota. 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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

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.

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.

Tong Xiao, Jingbo Zhu, Hao Zhang, and Qiang Li. 2012. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation. In *Proceedings of the ACL 2012 System Demonstrations,*

pages 19–24, Jeju Island, Korea. Association for Computational Linguistics.

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.