Exploiting Cross-Sentence Context for Neural Machine Translation

Longyue Wang[†]   **Zhaopeng Tu**[‡*]   **Andy Way**[†]   **Qun Liu**[†]

[†]ADAPT Centre, School of Computing, Dublin City University, Ireland
{longyue.wang, andy.way, qun.liu}@adaptcentre.ie
[‡]Tencent AI Lab, China
tuzhaopeng@gmail.com

Abstract

In translation, considering the document as a whole can help to resolve ambiguities and inconsistencies. In this paper, we propose a cross-sentence context-aware approach and investigate the influence of historical contextual information on the performance of neural machine translation (NMT). First, this history is summarized in a hierarchical way. We then integrate the historical representation into NMT in two strategies: 1) a warm-start of encoder and decoder states, and 2) an auxiliary context source for updating decoder states. Experimental results on a large Chinese-English translation task show that our approach significantly improves upon a strong attention-based NMT system by up to +2.1 BLEU points.

1 Introduction

Neural machine translation (NMT) has been rapidly developed in recent years (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Tu et al., 2016). The encoder-decoder architecture is widely employed, in which the encoder summarizes the source sentence into a vector representation, and the decoder generates the target sentence word by word from the vector representation. Using the encoder-decoder framework as well as gating and attention techniques, it has been shown that the performance of NMT has surpassed the performance of traditional statistical machine translation (SMT) on various language pairs (Luong et al., 2015).

The continuous vector representation of a symbol encodes multiple dimensions of similarity, equivalent to encoding more than one meaning of a word. Consequently, NMT needs to spend a substantial amount of its capacity in disambiguating source and target words based on the context defined by a source sentence (Choi et al., 2016). Consistency is another critical issue in document-level translation, where a repeated term should keep the same translation throughout the whole document (Xiao et al., 2011; Carpuat and Simard, 2012). Nevertheless, current NMT models still process a documents by translating each sentence alone, suffering from inconsistency and ambiguity arising from a single source sentence. These problems are difficult to alleviate using only limited intra-sentence context.

The cross-sentence context, or global context, has proven helpful to better capture the meaning or intention in sequential tasks such as query suggestion (Sordoni et al., 2015) and dialogue modeling (Vinyals and Le, 2015; Serban et al., 2016). The leverage of global context for NMT, however, has received relatively little attention from the research community.[1] In this paper, we propose a cross-sentence context-aware NMT model, which considers the influence of *previous source sentences* in the same document.[2]

Specifically, we employ a hierarchy of Recurrent Neural Networks (RNNs) to summarize the cross-sentence context from source-side previous sentences, which deploys an additional document-level RNN on top of the sentence-level RNN encoder (Sordoni et al., 2015). After obtaining the global context, we design several strategies to integrate it into NMT to translate the current sentence:

- *Initialization*, that uses the history represen-

---

[1]To the best of our knowledge, our work and Jean et al. (2017) are two independently early attempts to model cross-sentence context for NMT.

[2]In our preliminary experiments, considering target-side history inversely harms translation performance, since it suffers from serious error propagation problems.

*Corresponding Author: Zhaopeng Tu

tation as the initial state of the encoder, decoder, or both;

- *Auxiliary Context*, that uses the history representation as static cross-sentence context, which works together with the dynamic intra-sentence context produced by an attention model, to good effect.

- *Gating Auxiliary Context*, that adds a gate to Auxiliary Context, which decides the amount of global context used in generating the next target word at each step of decoding.

Experimental results show that the proposed *initialization* and *auxiliary context* (w/ or w/o gating) mechanisms significantly improve translation performance individually, and combining them achieves further improvement.

2 Approach

Given a source sentence $\mathbf{x}_m$ to be translated, we consider its K previous sentences in the same document as cross-sentence context $C = \{\mathbf{x}_{m-K}, ..., \mathbf{x}_{m-1}\}$. In this section, we first model C, which is then integrated into NMT.
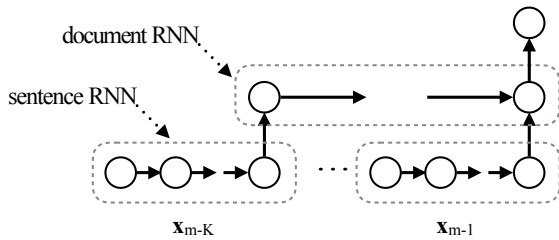

Figure 1: Summarizing global context with a hierarchical RNN ($\mathbf{x}_k$ is the k-th source sentence).

2.1 Summarizing Global Context

As shown in Figure 1, we summarize the representation of C in a hierarchical way:

Sentence RNN   For a sentence $\mathbf{x}_k$ in C, the sentence RNN reads the corresponding words $\{x_{1,k}, ..., x_{n,k}, ..., x_{N,k}\}$ sequentially and updates its hidden state:

$$h_{n,k} = f(h_{n-1,k}, x_{n,k}) \qquad (1)$$

where $f(\cdot)$ is an activation function, and $h_{n,k}$ is the hidden state at time n. The last state $h_{N,k}$ stores order-sensitive information about all the words in $\mathbf{x}_k$, which is used to represent the summary of the whole sentence, i.e. $S_k \equiv h_{N,k}$. After processing

each sentence in C, we can obtain all sentence-level representations, which will be fed into document RNN.

Document RNN   It takes as input the sequence of the above sentence-level representations $\{S_1, ..., S_k, ..., S_K\}$ and computes the hidden state as:

$$h_k = f(h_{k-1}, S_k) \qquad (2)$$

where h_k is the recurrent state at time k, which summarizes the previous sentences that have been processed to the position k. Similarly, we use the last hidden state to represent the summary of the global context, i.e. $D \equiv h_K$.

2.2 Integrating Global Context into NMT

We propose three strategies to integrate the history representation D into NMT:

Initialization   We use D to initialize either NMT encoder, NMT decoder or both. For encoder, we use D as the initialization state rather than all-zero states as in the standard NMT (Bahdanau et al., 2015). For decoder, we rewrite the calculation of the initial hidden state $s_0 = \tanh(W_s h_N)$ as $s_0 = \tanh(W_s h_N + W_D D)$ where h_N is the last hidden state in encoder and $\{W_s, W_D\}$ are the corresponding weight metrices.

Auxiliary Context   In standard NMT, as shown in Figure 2 (a), the decoder hidden state for time i is computed by

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \qquad (3)$$

where y_{i-1} is the most recently generated target word, and c_i is the intra-sentence context summarized by NMT encoder for time i. As shown in Figure 2 (b), *Auxiliary Context* method adds the representation of cross-sentence context D to jointly update the decoding state s_i:

$$s_i = f(s_{i-1}, y_{i-1}, c_i, D) \qquad (4)$$

In this strategy, D serves as an auxiliary information source to better capture the meaning of the source sentence. Now the gated NMT decoder has four inputs rather than the original three ones. The concatenation $[c_i, D]$, which embeds both intra- and cross-sentence contexts, can be fed to the decoder as a single representation. We only need to modify the size of the corresponding parameter matrix for least modification effort.
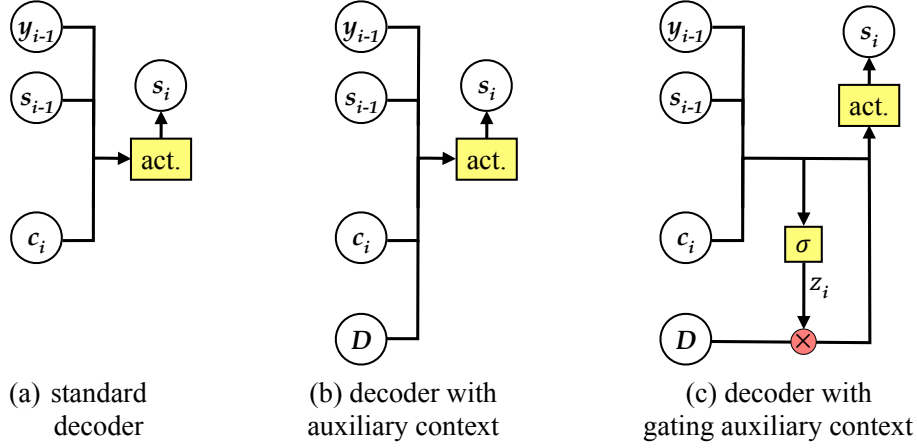
Figure 2: Architectures of NMT with auxiliary context integrations. $act.$ is the decoder activation function, and σ is a sigmoid function.

Gating Auxiliary Context The starting point for this strategy is an observation: the need for information from the global context differs from step to step during generation of the target words. For example, global context is more in demand when generating target words for ambiguous source words, while less by others. To this end, we extend auxiliary context strategy by introducing a context gate (Tu et al., 2017a) to dynamically control the amount of information flowing from the auxiliary global context at each decoding step, as shown in Figure 2 (c).

Intuitively, at each decoding step i, the context gate looks at decoding environment (i.e., s_i, y_{i-1}, and c_i), and outputs a number between 0 and 1 for each element in D, where 1 denotes "completely transferring this" while 0 denotes "completely ignoring this". The global context vector D is then processed with an element-wise multiplication before being fed to the decoder activation layer.

Formally, the context gate consists of a sigmoid neural network layer and an element-wise multiplication operation. It assigns an element-wise weight to D, computed by

$$z_i = \sigma(U_z s_{i-1} + W_z y_{i-1} + C_z c_i) \quad (5)$$

Here $\sigma(\cdot)$ is a logistic sigmoid function, and $\{W_z, U_z, C_z\}$ are the weight matrices, which are trained to learn when to exploit global context to maximize the overall translation performance. Note that z_i has the same dimensionality as D, and thus each element in the global context vector has its own weight. Accordingly, the decoder hidden state is updated by

$$s_i = f(s_{i-1}, y_{i-1}, c_i, z_i \otimes D) \quad (6)$$

3 Experiments

3.1 Setup

We carried out experiments on Chinese–English translation task. As the document information is necessary when selecting the previous sentences, we collect all LDC corpora that contain document boundary. The training corpus consists of 1M sentence pairs extracted from LDC corpora[3] with 25.4M Chinese words and 32.4M English words. We chose the NIST05 (MT05) as our development set, and NIST06 (MT06) and NIST08 (MT08) as test sets. We used case-insensitive BLEU score (Papineni et al., 2002) as our evaluation metric, and sign-test (Collins et al., 2005) for calculating statistical significance.

We implemented our approach on top of an open source attention-based NMT model, Nematus[4] (Sennrich and Haddow, 2016; Sennrich et al., 2017). We limited the source and target vocabularies to the most frequent 35K words in Chinese and English, covering approximately 97.1% and 99.4% of the data in the two languages respectively. We trained each model on sentences of length up to 80 words in the training data with early stopping. The word embedding dimension was 600, the hidden layer size was 1000, and the batch size was 80. All our models considered the previous three sentences (i.e., $K = 3$) as cross-sentence context.

---

[3] The LDC corpora indexes are: 2003E07, 2003E14, 2004T07, 2005E83, 2005T06, 2006E24, 2006E34, 2006E85, 2006E92, 2007E87, 2007E101, 2007T09, 2008E40, 2008E56, 2009E16, 2009E95.

[4] Available at https://github.com/EdinburghNLP/nematus.

#	System	MT05	MT06	MT08	Ave.	△
1	MOSES	33.08	32.69	23.78	28.24	–
2	NEMATUS	34.35	35.75	25.39	30.57	–
3	+Init$_{enc}$	36.05	36.44†	26.65†	31.55	+0.98
4	+Init$_{dec}$	36.27	36.69†	27.11†	31.90	+1.33
5	+Init$_{enc+dec}$	36.34	36.82†	27.18†	32.00	+1.43
6	+Auxi	35.26	36.47†	26.12†	31.30	+0.73
7	+Gating Auxi	36.64	37.63†	26.85†	32.24	+1.67
8	+Init$_{enc+dec}$+Gating Auxi	**36.89**	**37.76**†	**27.57**†	**32.67**	+2.10

Table 1: Evaluation of translation quality. "Init" denotes Initialization of encoder ("enc"), decoder ("dec"), or both ("enc+dec"), and "Auxi" denotes Auxiliary Context. "†" indicates statistically significant difference ($P < 0.01$) from the baseline NEMATUS.

3.2 Results

Table 1 shows the translation performance in terms of BLEU score. Clearly, the proposed approaches significantly outperforms baseline in all cases.

Baseline (Rows 1-2) NEMATUS significantly outperforms Moses – a commonly used phrase-based SMT system (Koehn et al., 2007), by 2.3 BLEU points on average, indicating that it is a strong NMT baseline system. It is consistent with the results in (Tu et al., 2017b) (i.e., 26.93 vs. 29.41) on training corpora of similar scale.

Initialization Strategy (Rows 3-5) Init$_{enc}$ and Init$_{dec}$ improve translation performance by around +1.0 and +1.3 BLEU points individually, proving the effectiveness of warm-start with cross-sentence context. Combining them achieves a further improvement.

Auxiliary Context Strategies (Rows 6-7) The gating auxiliary context strategy achieves a significant improvement of around +1.0 BLEU point over its non-gating counterpart. This shows that, by acting as a critic, the introduced context gate learns to distinguish the different needs of the global context for generating target words.

Combining (Row 8) Finally, we combine the best variants from the initialization and auxiliary context strategies, and achieve the best performance, improving upon NEMATUS by +2.1 BLEU points. This indicates the two types of strategies are complementary to each other.

3.3 Analysis

We first investigate to what extent the mis-translated errors are fixed by the proposed system.

We randomly select 15 documents (about 60 sentences) from the test sets. As shown in Table 2, we count how many related errors: i) are made by NMT (*Total*), and ii) fixed by our method (*Fixed*); as well as iii) newly generated (*New*). About *Ambiguity*, while we found that 38 words/phrases were translated into incorrect equivalents, 76% of them are corrected by our model. Similarly, we solved 75% of the *Inconsistency* errors including lexical, tense and definiteness (definite or indefinite articles) cases. However, we also observe that our system brings relative 21% new errors.

Errors	Ambiguity	Inconsistency	All
Total	38	32	70
Fixed	29	24	53
New	7	8	15

Table 2: Translation error statistics.

Hist.	这 不 等于 明着 提前 告诉 贪官 们 赶紧 转移 罪证 吗 ?
Input	能否 遏制 和 震慑 腐官 ?
Ref.	Can it inhibit and deter corrupt officials?
NMT	Can we contain and deter the *enemy*?
Our	Can it contain and deter the **corrupt officials**?

Table 3: Example translations. We italicize some *mis-translated* errors and highlight the **correct** ones in bold.

Case Study Table 3 shows an example. The word "腐官" (*corrupt officials*) is mis-translated as "*enemy*" by the baseline system. With the help

of the similar word "贪官" in the previous sentence, our approach successfully correct this mistake. This demonstrates that cross-sentence context indeed helps resolve certain ambiguities.

4 Related Work

While our approach is built on top of hierarchical recurrent encoder-decoder (HRED) (Sordoni et al., 2015), there are several key differences which reflect how we have generalized from the original model. Sordoni et al. (2015) use HRED to summarize a single representation from both the current and previous sentences, which limits itself to (1) it is only applicable to encoder-decoder framework without attention model, (2) the representation can only be used to initialize decoder. In contrast, we use HRED to summarize the previous sentences alone, which provides additional cross-sentence context for NMT. Our approach is more flexible at (1) it is applicable to any encoder-decoder frameworks (e.g., with attention), (2) the cross-sentence context can be used to initialize either encoder, decoder or both.

While both our approach and Serban et al. (2016) use Auxiliary Context mechanism for incorporating cross-sentence context, there are two main differences: 1) we have separate parameters to better control the effects of the cross- and intra-sentence contexts, while they only have one parameter matrix to manage the single representation that encodes both contexts; 2) based on the intuition that not every target word generation requires equivalent cross-sentence context, we introduce a context gate (Tu et al., 2017a) to control the amount of information from it, while they don't.

At the same time, some researchers propose to use an additional set of an encoder and attention to model more information. For example, Jean et al. (2017) use it to encode and select part of the previous source sentence for generating each target word. Calixto et al. (2017) utilize global image features extracted using a pre-trained convolutional neural network and incorporate them in NMT. As additional attention leads to more computational cost, they can only incorporate limited information such as single preceding sentence in Jean et al. (2017). However, our architecture is free to this limitation, thus we use multiple preceding sentences (e.g. $K = 3$) in our experiments.

Our work is also related to multi-source (Zoph and Knight, 2016) and multi-target NMT (Dong et al., 2015), which incorporate additional source or target languages. They investigate one-to-many or many-to-one languages translation tasks by integrating additional encoders or decoders into encoder-decoder framework, and their experiments show promising results.

5 Conclusion and Future Work

We proposed two complementary approaches to integrating cross-sentence context: 1) a warm-start of encoder and decoder with global context representation, and 2) cross-sentence context serves as an auxiliary information source for updating decoder states, in which an introduced context gate plays an important role. We quantitatively and qualitatively demonstrated that the presented model significantly outperforms a strong attention-based NMT baseline system. We release the code for these experiments at `https://www.github.com/tuzhaopeng/LC-NMT`. Our models benefit from larger contexts, and would be possibly further enhanced by other document level information, such as discourse relations. We propose to study such models for full length documents with more linguistic features in future work.

Acknowledgments

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*. San Diego, USA, pages 1–15.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521* .

Marine Carpuat and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montreal, Quebec, Canada, pages 442–449.

Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. 2016. Context-dependent word representa-

tion for neural machine translation. *arXiv preprint arXiv:1607.00578* .

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, pages 531–540.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Assocaition for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, pages 1723–1732.

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

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709.

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*. Prague, Czech Republic, pages 177–180.

Thang Luong, Hieu Pham, and D. Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 1412–1421.

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 on Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA, pages 311–318.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357* .

Rico Sennrich and Barry Haddow. 2016. *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, Berlin, Germany, chapter Linguistic Input Features Improve Neural Machine Translation, pages 83–91.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, Arizona, pages 3776–3783.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. Melbourne, Australia, pages 553–562.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 2014 Neural Information Processing Systems*. Montreal, Canada, pages 3104–3112.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017a. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics* .

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017b. Neural machine translation with reconstruction. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI-17)*. San Francisco, California, USA, pages 3097–3103.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*. Berlin, Germany, pages 76–85.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the International Conference on Machine Learning, Deep Learning Workshop*. pages 1–8.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Machine Translation Summit*. Xiamen, China, volume 13, pages 131–138.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710* .