Modeling Discourse Structure for Document-level Neural Machine Translation

Junxuan Chen[1]*, **Xiang Li**[2], **Jiarui Zhang**[1], **Chulun Zhou**[1],
Jianwei Cui[2], **Bin Wang**[2], **Jinsong Su**[1]†
[1]Xiamen University, Xiamen, China
[2]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
{chenjx,zhangjiarui,clzhou}@stu.xmu.edu.cn jssu@xmu.edu.cn
{lixiang21,cuijianwei,wangbin11}@xiaomi.com

Abstract

Recently, document-level neural machine translation (NMT) has become a hot topic in the community of machine translation. Despite its success, most of existing studies ignored the discourse structure information of the input document to be translated, which has shown effective in other tasks. In this paper, we propose to improve document-level NMT with the aid of discourse structure information. Our encoder is based on a hierarchical attention network (HAN) (Miculicich et al., 2018). Specifically, we first parse the input document to obtain its discourse structure. Then, we introduce a Transformer-based path encoder to embed the discourse structure information of each word. Finally, we combine the discourse structure information with the word embedding before it is fed into the encoder. Experimental results on the English-to-German dataset show that our model can significantly outperform both Transformer and Transformer+HAN.

1 Introduction

Neural machine translation (NMT) has made great progress in the past decade. In practical applications, the need for NMT systems has expanded from individual sentences to complete documents. Therefore, document-level NMT has gradually drawn much more attention. Contextual information is particularly important for obtaining high-quality document translation. To get better contextual information, researchers have proposed many methods (e.g., memory network and hierarchical attention network) for document-level translation (Sim Smith, 2017; Tiedemann and Scherrer, 2017; Wang et al., 2017a; Tu et al., 2017; Wang et al., 2017a; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019; Yang et al., 2019). Discourse structure, as well as raw contextual sentences, is a major component of the document. And it has been proved to be effective in many other tasks, such as automatic document summarization (Yoshida et al., 2014; Isonuma et al., 2019) and sentiment classification (Schouten and Frasincar, 2016; Nejat et al., 2017). However, to the best of our knowledge, discourse structure has not been explicitly used in document-level NMT.

To address the above problem, we propose to improve document-level NMT with the aid of discourse structure information. First, we represent each input document with a Rhetorical Structure Theory-based discourse tree (Mann and Thompson, 1988). Then, we use a Transformer-based path encoder to embed the discourse structure path of each word and combine it with the corresponding word embedding before feeding it into the sentence encoder. In this way, discourse structure information can be fully exploited to enrich word representations and guide the context encoder to capture the relevant context of the current sentence. Finally, we adopt HAN (Miculicich et al., 2018) as our context encoder to model context information in a hierarchical manner.

Our contributions are as follows: (i) We propose a novel and efficient approach to explicitly exploit discourse structure information for document-level NMT. Particularly, our approach is applicable for any other context encoder of document-level NMT; (ii) We carry out experiments on English-to-German translation task and experimental results show that our model outperforms competitive baselines.

*This work is done when Junxuan Chen was interning at Xiaomi AI Lab, Xiaomi Inc., Beijing, China.

†Corresponding author.

Proceedings of the 1st Workshop on Automatic Simultaneous Translation, pages 30–36
July 10, 2020.

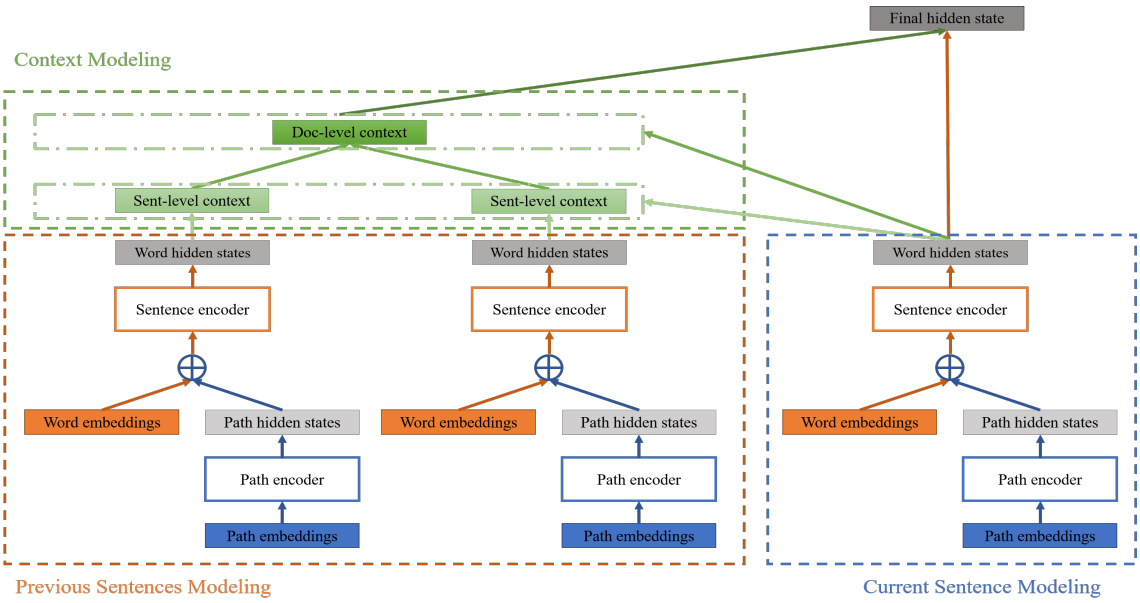


Figure 1: The architecture of our proposed encoder

2 Related Work

In the era of statistical machine translation, document-level machine translation has become one of the research focuses in the community of machine translation. (Xiao et al., 2011; Su et al., 2012; Xiao et al., 2012; Su et al., 2015). Recently, with the rapid development of NMT, document-level NMT has also gradually attracted people's attention (Voita et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Maruf et al., 2019; Yang et al., 2019). Typically, existing studies aim to improve document-level translation quality with the help of document context, which is usually extracted from neighboring sentences of the current sentence. For example,some researchers applied cache-based models to selectively remember the most relevant context information of the document (Voita et al., 2018; Maruf and Haffari, 2018; Kuang et al., 2018), while some researchers employed hierarchical context networks to catch document context information for Transformer (Miculicich et al., 2018; Maruf et al., 2019; Yang et al., 2019). Specifically, Miculicich et al. (2018) proposed a hierarchical attention network to model contextual information, Maruf et al. (2019) applied a selective attention method to select contextual information that is more relevant to the current sentence, and Yang et al. (2019) employed capsule network to model multi-angle context information.

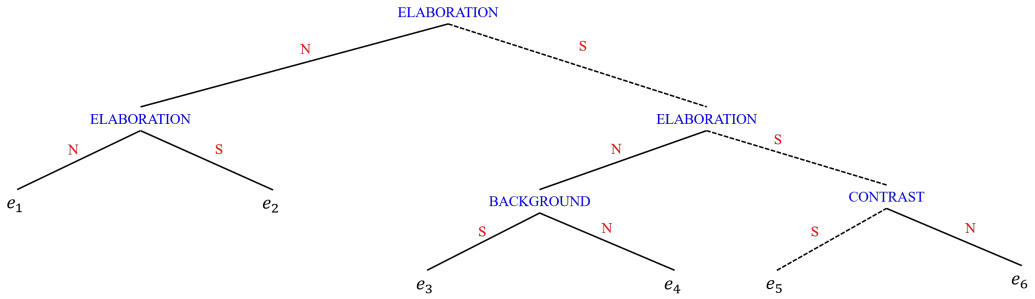
Although these methods have made some progress in document-level NMT, they all ignored the discourse structure information, which can be used to not only enrich word embedding but also guide the selection of relevant context for the current sentence.

3 Our Encoder

We propose a novel document-level NMT model based on HAN (Miculicich et al., 2018). The difference between ours and HAN lies in that we introduce the RST-based discourse structure to represent the document-level context, which is incorporated into HAN to refine translation.

Figure 1 gives the architecture of our proposed encoder. In addition to the standard encoder for the current sentence, it contains HAN (Miculicich et al., 2018) as context encoder, and a novel path encoder for the discourse structure. We first use the Transformer-based path encoder to model discourse structure information. Then, we combine the embedding of each input word with its corresponding path embedding vector and feed the combined vector into the sentence encoder. Finally, we use the hierarchical attention network to capture the global contextual embedding and update the hidden states of current sentence as the final output of the whole encoder.

In our model, the translation of a document is made by translating each of its sentences sequentially. We introduce discourse structure for both the current sentence and contextual sentences. Given a source document X, the translation probability of



1. [For example, one important achievement of former President Felipe Calderón's administration was to push through a 140-mile highway]e_1 [connecting the interior city of Durango and the Pacific port at Mazatlán.]e_2
2. [Traversing extremely rough terrain with 200 tunnels and bridges,]e_3 [it promises to cut the transit time by three or four hours.]e_4
3. **[Except for the weather,]e_5 [the highway has the feel of Switzerland.]e_6**

Figure 2: An example discourse tree of with six EDUs. N and S denote the relative importance label *NUCLEUS* and *SATELLITE*, respectively. Sentence 3 is the current sentence to be translated, with two previous context sentences 1 and 2. On the tree, the path marked with dotted lines from the root node to the leaf node e_5 is used to represent the discourse structure of e_5.

the target document Y can be defined as:

$$P(Y \mid X;\theta) = \prod_{j=1}^{J} P(Y^j \mid X^j, D^j, S;\theta), \quad (1)$$

where X^j and Y^j denote the j-th source sentence and its target translation respectively, D^j denotes the contextual sentences, S represents the discourse structure of the document to be translated, and θ is the parameter set of the model.

3.1 RST-based Discourse Structure

For each document to be translated, we parse it to obtain its discourse structure based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). RST is one of the most influential theories of document coherence. According to RST, we represent each document with a hierarchical tree. As shown in Figure 2, the discourse tree contains some leaf nodes, each of which indicates an Elementary Discourse Unit (EDU). The adjacent leaf nodes are recursively connected into units by certain coherence relations (e.g., *ELABORATION, BACKGROUND*) until the entire tree is built. Besides, *NUCLEUS* or *SATELLITE* is used to mark the relative importance of child node units in the relationship.

In this work, we represent the discourse structure information of each word using its discourse path from root node to its corresponding leaf node. Each path is a mixed label sequence composed of the discourse relationship and the importance label (e.g., *NUCLEUS_ELABORATION, SATELLITE_BACKGROUND*). Please note that all tokens in the same EDU share the same discourse structure. For example, the discourse structure of EDU e_5 is "*SATELLITE_ELABORATION SATELLITE_ELABORATION SATELLITE_CONTRAST*".

3.2 Discourse Structure Path Embedding

To integrate the structural information of words into the our HAN-based document-level NMT model, we first additionaly introduce a Transformer-based path encoder to encode discourse structure paths of words. Specifically, for each word w_i, we directly consider its discourse structure path p_i as a sequence and then employ the path encoder to learn its contextual hidden states, which can be finally averaged to produce the overall discourse embedding vector d_i. Then, we enrich each input word embedding with its corresponding discourse embedding vector before it is fed into the context encoder or the translation encoder. Concretely, for the word w_i, we define its enriched vector as the sum of its word embedding and discourse embedding: $\widetilde{x_i} = x_i + d_i$.

3.3 HAN-based Context Modeling

Following (2018), we apply hierarchical attention network (HAN) as our context encoder. Due to the advantage of accurately capturing different levels of contexts, HAN has been widely used in many tasks, such as document classification (Yang et al., 2016), stance detection (Sun et al., 2018), sentence-level NMT (Su et al., 2018b). Using this encoder, we mainly focus on two levels of context modeling:

Sentence-level Context Modeling For the i-th word of the current sentence, we employ muti-head

attention (Vaswani et al., 2017) to summarize the context from the k-th context sentence:

$$cs_{i,k} = \text{MultiHead}(f_s(h_i), \mathbf{H_k}), \quad (2)$$

where f_s is a linear transformation function, h_i denotes the hidden state representation of the i-th token of current sentence. By doing so, our context encoder can exploit different types of relation between words to better capture sentence-level context. And $\mathbf{H_k}$ is the hidden state representation of the k-th context sentence and is used as value and key for attention.

Document-level Context Modeling Unlike the above modeling, here we mainly on capturing the context information from previous K sentences for the i-th word of the current sentence.

$$cd_i = \text{FFN}(\text{MultiHead}(f_d(h_i), \mathbf{CS_i})), \quad (3)$$

where f_d is a linear transformation, and $\mathbf{CS_i} = [cs_{i,1}, cs_{i,2}, \cdots, cs_{i,K}]$ is the sentence-level context of K contextual sentences.

Integrating Document-level Context into the Translation Encoder Finally, we integrate the above-mentioned document-level context into the translation encoder via a gating operation:

$$\lambda_i = \sigma(\mathbf{W_h} h_i + \mathbf{W_{cd}} cd_i) \quad (4)$$

$$\widetilde{h_i} = \lambda_i h_i + (1 - \lambda_i) cd_i \quad (5)$$

where $\mathbf{W_h}$ and $\mathbf{W_{cd}}$ denote parameter matrices for h_i and cd_i, and $\widetilde{h_i}$ is the final output of the encoder.

4 Experiments

4.1 Settings

Datasets We conduct our experiments on English-to-German translation task. The sentence-aligned document-delimited News Comment v11 corpus [1], the WMT16 newstest2015 and the newstest2016 are used as the training set, development and test set, respectively.

We download all the above corpus from (Maruf et al., 2019), of which statistics are provided in Table 1.

[1] http://www.casmacat.eu/corpus/news-commentary.html

	#Sentences	Document length
Training	236,287	38.93
Development	2,169	26.78
Test	2,999	19.35

Table 1: The statistical of our datasets. #Sentence indicates the number of sentences, and Document length means the average number of sentences in document.

Settings We use Transformer (Vaswani et al., 2017) as our context-agnostic baseline system and Transformer+HAN (Miculicich et al., 2018) as our context-aware baseline system. We conduct experiments using the same configuration as HAN. Specifically, both sentence encoder and decoder are composed of 6 hidden layers, while path encoder is composed of 2 hidden layers. We use three previous sentences as contextual sentences for current sentence. The hidden size and pointwise FFN size are 512 and 2048 respectively. The dropout rates for all hidden states are set to 0.1. The source and target vocabulary sizes are both 30K. At the training phase, we use the Adam optimizer (Kingma and Ba, 2015) and the batch sizes of context-agnostic model and context-aware model are 4096 and 1024, respectively. Finally, we use case-sensitive BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) to measure the translation quality.

Data Preprocessing All datasets are tokenized and truecased using the scripts of Moses Toolkit (Koehn et al., 2007). We split them into subword units using a joint bye pair encoding model with 30K merge operations. To get discourse structure of the input documents, we first apply the open-source tool NeuralEDUSeg (Wang et al., 2018) obtaining non-overlapping EDUs. Then, we employ StageDP (Wang et al., 2017b) to obtain discourse structure trees of segmented documents. Afterwards, we extract the path from root node to leaf node as the discourse structure information for the corresponding EDU, where all words share the same discourse structure path.

4.2 Results and Analysis

Table 2 shows the experimental results for different models. The sentence-level Transformer, context-agnostic baseline, obtains a result of 22.78 BLEU and 59.3 TER, while the context-aware baseline Transformer+HAN (Miculicich et al., 2018) obtains 24.45 BLEU and 56.9 TER. The sentence-

Model	BLEU	TER
Transformer	22.78	59.3
Transformer+DS	23.61 (+0.83)	58.5 (-0.8)
Transformer+HAN	24.45 (+1.67)	56.9 (-2.4)
Transformer+HAN+DS	**24.84** (+2.06)	**56.4** (-2.9)

Table 2: BLEU and TER scores for different models. The best scores are marked in bold. HAN denotes Hierarchical Attention Network which is used to get context information while DS denotes Discourse Structure information.

level Transformer integrated with discourse structure achieves an improvement of 0.83 on BLEU and a decline of 0.8 on TER. By contrast, our model integrated with contextual information and discourse structure information further gains a better performance, 2.06 higher than Transformer and 0.39 higher than Transformer+HAN on BLEU, 2.9 lower than Transformer and 0.5 lower than Transformer+HAN on TER.

Our experimental results show that discourse structure features indeed provide helpful information to enhance the translation quality of both context-agnostic and context-aware document-level NMT models. Please note that our approach is also applicable to other document-level NMT models.

5 Conclusion

This paper has presented a novel discourse structure-based encoder for document-level NMT. The main idea of our encoder lies in enriching the input word embeddings with their path embeddings based on discourse structure. Experimental results on English-to-German translation verify the effectiveness of our proposed encoder.

In the future, we plan to extend our encoder to other NLP tasks, such as simultaneous translation. Simultaneous translation, as well as document-level NMT, has difficulty in modeling the long context so that it may be effective to improve the re-translation with the structure information modeled by our encoder. Finally, we will focus on refining document-level NMT with variational neural networks, which has shown effecitive in previous studies of sentence-level NMT (Zhang et al., 2016; Su et al., 2018a).

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2019QY1803, the National Natural Science Foundation of China (No. 61672440), and the Scientific Research Project of National Language Committee of China (No. YB135-49).

References

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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

Shaohui Kuang, 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.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text Talk*, 8(3):243–281.

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.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298, Saarbrücken, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kim Schouten and Flavius Frasincar. 2016. COMMIT at SemEval-2016 task 5: Sentiment analysis with rhetorical structure theory. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 356–360, San Diego, California. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 459–468, Jeju Island, Korea. Association for Computational Linguistics.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018a. Variational recurrent neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jinsong Su, Deyi Xiong, Yang Liu, Xianpei Han, Hongyu Lin, Junfeng Yao, and Min Zhang. 2015. A context-aware topic model for statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 229–238, Beijing, China. Association for Computational Linguistics.

Jinsong Su, Jiali Zeng, Deyi Xiong, Yang Liu, Mingxuan Wang, and Jun Xie. 2018b. A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):623–632.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. 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, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Advances in neural information processing systems*, pages 5998—-6008.

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. 2017a. 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.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017b. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. *Proceedings of the 13th Machine Translation Summit*, 13:131–138.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 750–758, Jeju Island, Korea. Association for Computational Linguistics.

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. 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.