Neural Machine Translation with Heterogeneous Topic Knowledge Embeddings

Weixuan Wang[1] *, **Wei Peng**[1] *, **Meng Zhang**[2], **Qun Liu**[2]
[1]Artificial Intelligence Application Research Center, Huawei Technologies
{weixuanwang2,peng.wei1}@huawei.com
[2]Noah's Ark Lab, Huawei Technologies
{zhangmeng92,qun.liu}@huawei.com

Abstract

Neural Machine Translation (NMT) has shown a strong ability to utilize local context to disambiguate the meaning of words. However, it remains a challenge for NMT to leverage broader context information like topics. In this paper, we propose heterogeneous ways of embedding topic information at the sentence level into an NMT model to improve translation performance. Specifically, the topic information can be incorporated as pre-encoder topic embedding, post-encoder topic embedding, and decoder topic embedding to increase the likelihood of selecting target words from the same topic of the source sentence. Experimental results show that NMT models with the proposed topic knowledge embedding outperform the baselines on the English → German and English → French translation tasks. [1]

1 Introduction

Neural Machine Translation (NMT) utilizes local context captured from the mapping between the bitexts to disambiguate the meaning of words. While existing NMT models can handle meaning ambiguities based on local contexts learned from explicit collocations, it remains a challenge for NMT to produce accurate results for words presented in implicit collocations. The notion of implicit collocation is referred to as the circumstance when the meaning of two or more words can not be learned from the available training data; broader context information like topics may be utilized to generate an accurate meaning. For example, in the sentence ***"he likes bank fishing"***, the word ***"bank fishing"*** is likely to produce an ill-translated Chinese word "银行钓鱼" due to a lack of collocation of ***"bank (河岸)"*** and ***"fishing (钓鱼)"***. An accurate translation "河岸钓鱼" may be approachable when the shared topic (***"recreational sport"***) is leveraged.

Incorporating topic information into NMT has been explored in Zhang et al. (2016) and Wei et al. (2019) with both studies adapting Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model topics of source and target languages. Both works utilized the traditional encoder-decoder architecture with gated recurrent units (GRU) (Cho et al., 2014). Although Wei et al. (2019) showed that topic knowledge incorporation is also applicable to the Transformer architecture (Vaswani et al., 2017), it is argued that the joint learning of topic modeling and NMT is not an ideal way. The training of topic models can leverage a large volume of easily accessible monolingual data. Once a topic model is learned, it can be reused in different translation scenarios without retraining NMT models. Therefore decoupling topic modeling and NMT is a more flexible and scalable option.

In this paper, we propose heterogeneous ways of incorporating topic information into the Transformer architecture. Specifically, the topic information can be incorporated in a heterogeneous manner, namely as pre-encoder topic embedding (ENC_{pre}), post-encoder topic embedding (ENC_{post}), and decoder topic embedding (DEC). Besides, the topic distribution learned for each word (as its topic embedding) is summarized at the sentence level and fed into the NMT model. The intuition is that aggregating topic distribution at the sentence level produces more accurate topic information than at the word level. This enables topic modeling to consider contexts conveyed in a sentence. Each target word is generated with the guidance of the topic information of both source and target sentences. The topic-enhanced NMT models are trained on WMT14 English → German translation task and tested on a range of WMT datasets. Experimental results show that our approach can significantly improve translation quality with the topic embedding by achieving up to +1.57 BLEU score improvement over the Transformer

*Co-first authors.
[1]The codes are available at https://github.com/Vicky-Wil/topic-NMT

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3197–3202
November 7–11, 2021.

baselines. The effect of the proposed method is also verified on the English → French translation task.

2 Related Work

Many studies have focused on using topic information as explicit prior knowledge to help model learn sentence representations on NLP tasks, such as Zhang et al. (2017); Kim (2014); Kobus et al. (2017). Topic modeling has shown its effectiveness in statistical machine translation (SMT) models (Xiao et al., 2012; Xiong et al., 2015; Hasler et al., 2014). Incorporating topic information into NMT has recently been explored by Chen et al. (2016); Zhang et al. (2016); Wei et al. (2019); Chen et al. (2019). Zhang et al. (2016) proposed a topic-informed NMT model leveraging source-side and target-side topics, separately learned by two independent LDA models from training data. Wei et al. (2019) designed a bilingual topic NMT model incorporating bilingual topic knowledge into NMT to improve translation performance. Both works were built upon gated recurrent units (GRU) architecture with limited coverage to the Transformer architecture.

Both studies adopted LDA to model topics of source and target languages. Dieng et al. (2020) pointed out that LDA is not an effective learner for data with an extensive vocabulary because one has to remove the most and least frequent words to fit good topic models. This pruning practice limits the scope of LDA models. The embedding topic model (ETM) (Dieng et al., 2020) was proposed to model each term as an embedding and each topic as a point in that embedding space. The per-topic conditional probability of a term has a log-linear form to preserve low-dimensional representation of the vocabulary so that ETM can discover interpretable topics with large vocabularies, including rare words and stop words. In this study, we apply ETM to handle issues associated with large vocabularies.

Chen et al. (2019) used a variant of convolutional neural networks (CNN) to learn latent topic representations implicitly from sentence-level context. An additional multi-head attention module is directly involved in learning the attentions between topics and targeting words independently from the encoding of the Transformer. Chen et al. (2019) also tried an explicit topic representation computed by TF-IDF, but did not perform better than their latent version. In this paper, we propose multiple heterogeneous ways of explicitly integrating topic information into NMT, resulting in better performance.

3 Topic-enhanced Neural Machine Translation

Figure 1 illustrates the proposed topic-enhanced NMT model with topic ENC_{pre}, ENC_{post}, and DEC, built upon the Transformer architecture. The topic knowledge in the figure is obtained from the topic embedding tables for source and target languages produced by ETM.

3.1 Pre-encoder Topic Embedding

In the encoding phase, we convert the sequence of words into a sequence of word embedding x_i and a sequence of topic embedding t_i, as shown in Figure 2(a). The word embedding is obtained by looking up the word embedding table, which is randomly initialized and updated with training. The topic embedding table is pre-calculated as the intermediate product of ETM, and it is fixed during the NMT training process. Then we add up all the topic embedding in the sequence to produce the topic information distribution of the whole sentence $topic_s$, added to each word embedding of the input source words. Finally, we take the added word embedding representation e_i as the input embedding and feed it into the encoder with positional encoding results.

$$topic_s = \sum_{i=1}^{m} t_i \tag{1}$$

$$e_i = x_i + topic_s \tag{2}$$

3.2 Post-encoder Topic Embedding

The topic information distribution can also be added to each corresponding output of the encoder. The NMT decoder can implicitly attend to the topic distributions of each source word in this way. The topic-enhanced hidden state computes the topic context vector as:

$$c_j = \sum_{i=1}^{m} \alpha_{ij}(h_i + topic_s) \tag{3}$$

3.3 Decoder Topic Embedding

The topic information can be incorporated at the decoder side as shown in Figure 2(b). At time step $j-1$, we get the topic embedding $topic_{j-1}$ by

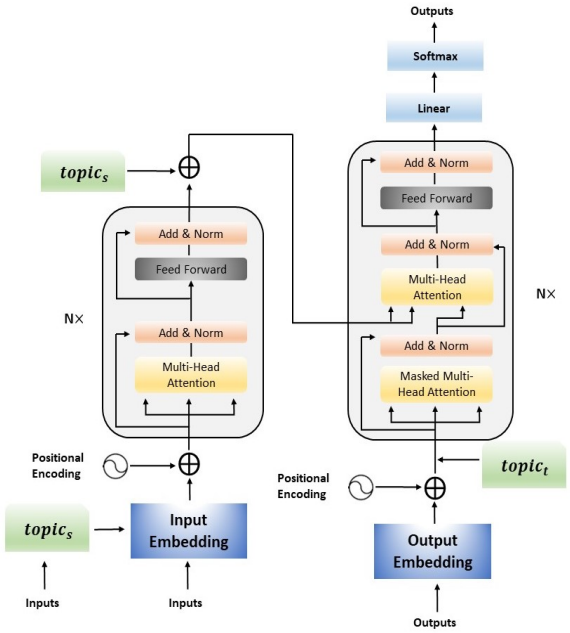
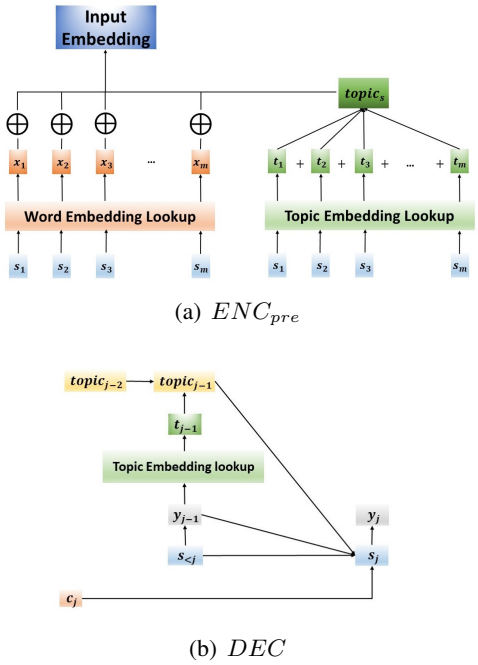


Figure 1: The illustration of the topic-enhanced NMT model. The $\oplus$ is a sum operation. The $topic_s$ is the encoder topic information obtained from the source words and the process of computation is illustrated in Figure 2(a). The $topic_t$ is the decoder topic information computed from the target words, as illustrated in Figure 2(b).



(a) ENC_{pre}

(b) DEC

Figure 2: (a) The $+$ represents adding up all the topic embedding of the source sentence. The $\oplus$ is to add the sentence topic information to the word embedding to generate the input embedding. (b) The $s_{<j}$ denotes the hidden state of decoder, y_{j-1} is the output token at the $j-1$ step, the t_{j-1} is the topic embedding for token y_{j-1}, and the c_j is a context vector.

adding the topic representation t_{j-1} to the previous topic embedding $topic_{j-2}$. By looking up the output word y_{j-1} in the target language topic embedding table, we get the topic representation t_{j-1}. Then the topic decoder embedding at time $j-1$ $topic_{j-1}$ is added to the previous output token y_{j-1} to participate in the decoding process. At time step j, the topic decoder is used to generate the target word y_j. Accordingly, the j-th hidden state of the topic decoder s_j is updated as:

$$s_j = f(y_{j-1}, s_{<j}, c_j, topic_{j-1}) \quad (4)$$

$$topic_{j-1} = topic_{j-2} + e(y_{j-1}) \quad (5)$$

MODEL	BLEU
Transformer (base) (Vaswani et al., 2017)	27.3
Transformer (big) (Vaswani et al., 2017)	28.4
Evolved Transformer (So et al., 2019)	28.4
DPE-NMT (Li et al., 2020)	27.61
Transformer base + PR (Xu et al., 2020)	28.67
Fairseq (baseline) (Ott et al., 2019)	27.44
BLT-NMT (Wei et al., 2019)	27.93
LTR-NMT (Chen et al., 2019)	28.18
Topic-enhanced NMT (ours)	**29.01**

Table 1: Evaluation of the WMT14 EN $\rightarrow$ DE translation using case-sensitive BLEU scores.

where c_j is the context vector obtained by attention mechanism, $e(\cdot)$ is the topic embedding table at the target side, $f(\cdot)$ is the non-linear calculation function of decoder. Consequently, the topic decoder can utilize the topic knowledge of previously generated target words with the topic information of the source sentence to increase the likelihood of selecting words from the same topic.

4 Experiments

Datasets Embedding topic model: We use the 20 Newsgroups corpus for training the English embedding table and use the WMT14 German monolingual dataset to train the German embedding table. To verify the effect of the proposed method on the English $\rightarrow$ French (EN $\rightarrow$ FR) task, we sample only 10 million (M) sentences representing less than one third of the training data randomly from WMT14 French monolingual dataset to train the French embedding table. The experiments are conducted on the standard WMT 14 English $\rightarrow$ German (EN $\rightarrow$ DE) and EN $\rightarrow$ FR training corpus as previous work (Wu et al., 2016). We evaluate the models on the newstest 2014, while the concatena-

MODEL	EN → DE				EN → FR
	newstest 2014	newstest 2016	newstest 2017	newstest 2019	newstest 2014
Fairseq	27.44	29.71	27.74	31.98	42.32
ENC_{pre}	28.72 (+1.28)	31.08 (+1.37)	28.64 (+0.90)	32.69 (+0.71)	42.57 (+0.25)
ENC_{post}	28.96 (+1.52)	30.92 (+1.21)	28.68 (+0.94)	33.18 (+1.20)	42.60 (+0.28)
DEC	28.59 (+1.15)	30.80 (+1.09)	28.41 (+0.67)	31.97 (-0.01)	42.80 (+0.48)
ENC_{pre} + DEC	28.75 (+1.31)	30.96 (+1.25)	28.49 (+0.75)	32.78 (+0.80)	42.94 (+0.62)
ENC_{pre} + ENC_{post}	**29.01 (+1.57)**	31.04 (+1.33)	28.65 (+0.91)	33.35 (+1.37)	42.89 (+0.57)
DEC + ENC_{post}	28.99 (+1.55)	30.94 (+1.23)	28.58 (+0.84)	32.91 (+0.93)	43.15 (+0.83)
ENC_{pre} + DEC + ENC_{post}	28.98 (+1.54)	31.05 (+1.34)	28.67 (+0.93)	33.28 (+1.30)	43.35 (+1.03)

Table 2: Case-sensitive BLEU (Papineni et al., 2002) scores evaluated on EN → DE translation task for topic NMT on newstest 2014, 2016, 2017, and 2019 and on EN → FR translation task on newstest 2014 with different settings. The numbers in parentheses represent the improvements of BLEU scores over the baseline BLEU score.

tion of newstest 2012 and newstest 2013 is used for the development set. The training corpus contains 4.5M sentence pairs for DE, and 35.7M sentence pairs for FR. We use the truecasing model (Lita et al., 2003) and Moses (Koehn et al., (2007) to tokenize all the data. Besides, we use both source and target vocabularies with 32K most frequent words for DE and 44K words for FR.

Training Details We preprocess the corpus for all experiments of ETM. We set the number of the topics to 50 and epoch number to 500, which are empirical values adopted from ETM. After preprocessing, we further remove one-word documents from the validation and test sets. For all NMT experiments, we train our models on one machine with 4 NVIDIA V100 GPUs and follow Vaswani et al. (2017) base model to set the hyper-parameters with model configurations. The number of parameters is 129M. We compare our topic model against the following models: **Fairseq** (base) is a sequence modeling toolkit (Ott et al., 2019). **BLT-NMT** is a topic enhanced model with incorporated bilingual topic knowledge into NMT (Wei et al., 2019). **LTR-NMT** is a topic-based NMT model using a CNN model (Chen et al., 2019).

Results The experimental results of various existing state-of-the-art (SOTA) models on the same dataset, including Base Transformer and Big Transformer (Vaswani et al., 2017), Evolved Transformer (So et al., 2019), Dynamic Programming Encoding NMT (Li et al., 2020), Phrase Representations Transformer (Xu et al., 2020), are quoted as a reference. For a fair comparison, we list the single best result reported in their papers.

The experimental results on EN → DE are depicted in Table 1. Compared to other NMT models, our baseline model based on the Transformer base architecture implemented in Fairseq achieves a BLEU score of 27.44, equivalent to the one for Vaswani et al. (2017). Our topic NMT model achieves 29.01 BLEU scores, significantly outperforming the baseline Fairseq by +1.57 BLEU points. Compared to BLT-NMT and LTR-NMT, our model is +1.08 and + 0.83 BLEU score higher.

To further investigate the effectiveness of our topic NMT model and study the main factor that influences the experiment results, we also compare different topic NMT on the newstest 2014, 2016, 2017, and 2019 dataset for EN → DE and the newstest 2014 for EN → FR. Ablation tests are performed to investigate the effects of three topic embedding options: ENC_{pre}, ENC_{post}, and DEC. The experimental results are shown in Table 2. It is noted that NMT with ENC_{pre}, ENC_{post}, and DEC achieve BLEU improvements of +1.28, +1.52 and +1.15, respectively over the baseline score in the newstest 2014 for EN → DE. The NMT models with four different combinations score +1.31, +1.57, +1.55, +1.54 BLEU points higher than that of the baseline in the newstest 2014. It can be observed that almost all experiments achieve higher BLEU scores over those of the baselines across different test sets. A consistent finding is confirmed in the EN → FR translation direction, indicating the effectiveness of the proposed method.

Examples of topic-enhanced NMT for EN → DE are shown in Table 3. For example, the base NMT model mistranslates "Systematic Theology" to "Systemtheorie" (systems theory in English), which is accurately translated to "Systematische Theologie" by the topic-enhanced NMT model.

5 Conclusion

In this paper, we propose heterogeneous ways of incorporating topic information as prior knowledge into the Transformer architecture to improve trans-

Example 1	
source	Since 2010, Johanna Rahner has occupied a Chair Systematic Theology in the Institute for Catholic Theology at the University of Kassel.
target	Seit 2010 hat Johanna Rahner einen Lehrstuhl für Systematische Theologie am Institut für Katholische Theologie der Universität Kassel inne.
base NMT	Johanna Rahner hat seit 2010 eine Lehrstuhl für Systemtheorie am Institut für katholische Theologie der Universität Kassel inne.
topic NMT	Seit 2010 hat Johanna Rahner einen Lehrstuhl für Systematische Theologie am Institut für Katholische Theologie der Universität Kassel inne.
Example 2	
source	An Obama voter's cry of despair.
target	Verzweiflungsschrei eines Obama Wählers.
base NMT	Obamas Ruf der Verzweiflung.
topic NMT	Der Schrei der Wähler eines Obama.
Example 3	
source	The previous silence was indeed a reaction to the events of previous days.
target	Das Schweigen zuvor war wohl eine Reaktion auf die Geschehnisse der vergangenen Tage.
base NMT	Das vorherige Schweigen war in der Tat eine Reaktion auf die Ereignisse der Vortage.
topic NMT	Das vorhergehende Schweigen war in der Tat eine Reaktion auf die Ereignisse der vergangenen Tage.

Table 3: Examples of improved translation quality when topic information is integrated to the NMT model as prior knowledge.

lation performance. The topic information can be incorporated as pre-encoder topic embedding, post-encoder topic embedding, and decoder topic embedding. Experimental results demonstrate that the proposed method can significantly improve translation quality by boosting the BLEU scores over the Transformer baselines on the English → German and English → French translation tasks.

Acknowledgements

We express our gratitude to colleagues from HUAWEI AI Application Research Center and Noah's Ark Lab for their continuous support. We also appreciate the anonymous reviewers for their constructive criticism.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Neural machine translation with sentence-level topic context. IEEE ACM Trans. Audio Speech Lang. Process., 27(12):1970–1984.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. CoRR, abs/1607.01628.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, pages 103–111. Association for Computational Linguistics.

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. Trans. Assoc. Comput. Linguistics, 8:439–453.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic topic adaptation for phrase-based MT. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden, pages 328–337. The Association for Computer Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, pages 1746–1751. ACL.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017, pages 372–378. INCOMA Ltd.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? A case study on context-aware neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 3512–3518. Association for Computational Linguistics.

Lucian Vlad Lita, Abraham Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. truecasing. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan, pages 152–159. ACL.

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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations, pages 48–53. 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, July 6-12, 2002, Philadelphia, PA, USA, pages 311–318. ACL.

David R. So, Quoc V. Le, and Chen Liang. 2019. The evolved transformer. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 5877–5886. PMLR.

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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Xiangpeng Wei, Yue Hu, Luxi Xing, Yipeng Wang, and Li Gao. 2019. Translating with bilingual topic knowledge for neural machine translation. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7257–7264. AAAI Press.

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

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A topic similarity model for hierarchical phrase-based translation. In The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers, pages 750–758. The Association for Computer Linguistics.

Deyi Xiong, Min Zhang, and Xing Wang. 2015. Topic-based coherence modeling for statistical machine translation. IEEE ACM Trans. Audio Speech Lang. Process., 23(3):483–493.

Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu, and Jingyi Zhang. 2020. Learning source phrase representations for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 386–396. Association for Computational Linguistics.

Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior knowledge integration for neural machine translation using posterior regularization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 1514–1523. Association for Computational Linguistics.

Jian Zhang, Liangyou Li, Andy Way, and Qun Liu. 2016. Topic-informed neural machine translation. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pages 1807–1817. ACL.