A Detection-based Attention Alignment Method for Document-level Neural Machine Translation

Kang Zhong[1] Wu Guo[1] Bin Gu[1] Weitai Zhang[2] Chao Lin[2]
[1]NERC-SLIP, University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Hefei, China
E-mail: Corresponding author: Wu Guo, guowu@ustc.edu.cn

Abstract

Previous works have shown that inter-sentential contextual information can lead to substantial improvements in document-level neural machine translation (DocNMT). Most existing DocNMT models focus on methods of introducing inter-sentential contextual information through attention mechanisms. Compared to intra-sentential attention, however, the long-range dependency in document-level attention calculation inevitably introduces meaningless contextual noise, resulting in significant performance deterioration. To address this problem, this paper proposes a detection-based attention alignment method, to help each translating word focus on relevant informative contextual words. We first introduce a context detector that automatically evaluates each source-side word's effect on the model's prediction. Based on the detection results, we align the original attention weights by integrating the cosine similarity between the aligned and original attention weights into the loss function, under a multi-task framework, which allows DocNMT to more effectively capture the document-level context. The results for three English-German (En-De) public translation datasets show that the proposed method can obtain consistent improvements over a strong G-Transformer baseline.

Index Terms: *document-level neural machine translation, contextual information, attention alignment*

1. INTRODUCTION

During the last decade, neural machine translation (NMT) has made remarkable progress to become a state-of-the-art method, especially for sentence-level translation [1, 2]. In document-level translation, it is widely accepted that the introduction of discourse dependencies between sentences can improve the coherence and quality of the translated text [3, 4]. Like those for sentence-level NMT, most existing document-level NMT (DocNMT) models integrate contextual information using an attention mechanism. A general method for capturing this information is to encode a limited number of previous or neighboring sentences [5–8]. Intuitively, encoding the whole input document [9–11] as a single unit should provide the best integration of context information. As Bao *et al.* [9] pointed out, however, it is difficult to train such a DocNMT model. As inputs are sparse for long sequences in DocNMT, the attention weights from target to context in the source-side are flat, with large entropy values. This indicates that long context sequences confound attention on meaningful portions of current translating words, and the model is difficult to select valuable inputs from context sequences to navigate the redundancy of information.

The simplest way to tackle this is to shorten the contextual sequence in attention calculation. Some works proposed selecting several sentences in a document as context, through integrating sparse attention [12] or using an extra selection module [13, 14]. These methods, however, treat all the words in a sentence equally, without discrimination, and select the same sentences for them. This inevitably creates meaningless contextual noise. To address this problem, Xu *et al.* suggest that each word should focus on its own informative words, spread throughout the document [10].

Attention mechanism plays an important role in NMT by dynamically selecting relevant inputs for different predictions. Based on the assumption that the translation is improved if attention values are more accurate, Li *et al.* and Lu *et al.* proposed the refinement of attention distribution in sentence-level NMT models, through word alignment [15] and attention calibration [16] respectively.

Inspired by the works of Xu *et al.* [10] and Lu *et al.* [16], this paper proposes a novel detection-based attention alignment method, to refine target-to-context attention for DocNMT. First, we introduce a context detector (CD), which is a lightweight network attached to the backbone, and only exists in the training procedure. The CD automatically de-

DOI reference number: 10.18293/SEKE2023-165

tects the significance of each source-side word for model prediction. Based on the detection results, we then align the original document-level attention weights, and use the cosine similarity as an auxiliary function in model training. In the inference procedure, the CD is dropped, and only the conventional backbone is used. We carry out experiments on three commonly used DocNMT datasets for English-German (En-De) translation, covering the domains of TED talks, News, and Europarl from small to large. The results show that our proposed method can further improve DocNMT performance over a strong G-Transformer [9] baseline.

Main contributions of this paper can be summarized as follows:

- Our method can automatically evaluate the significance of each contextual word for model prediction by introducing a lightweight perturbation noise detector, which can be jointly optimized and removed when inference.
- We propose a attention alignment loss to help making each translating word more focus on its own relevant essential contexts, which is experimentally proved to reduce the entropy of contextual attention weights and outperform strong baseline.
- Our code is publicly available[1].

2. BACKGROUND

2.1. Document-level Neural Machine Translation

Formally, $X = \{x_1, x_2, ..., x_S\}$ denotes the source document with S sentences, and $Y = \{y_1, y_2, ..., y_T\}$ denotes the target document with T sentences. S and T are usually assumed equal, because sentences can be merged with sentence alignment algorithms [17] to fix mismatches. Compared with sentence-level NMT, DocNMT not only gets the benefit of intra-sentential information in parallel pair $\{x_i, y_i\}$, but also takes advantages of the document's contextual information. The translation probability from X to Y can therefore be represented as:

$$P(Y|X) = \sum_{t=1}^{T} P(y_t|y_{<t}, X) \tag{1}$$

2.2. G-Transformer

The G-Transformer is inherited from the Transformer [1]. After analyzing the training failure of the Transformer's direct application into DocNMT, Bao *et al.* [9] attributed the problem to the huge complexity of target-to-source attention. They proposed the G-Transformer to reduce the hypothetical space of attention from target to source.

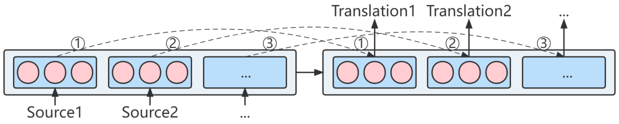


Figure 1. Overview of G-Transformer structure

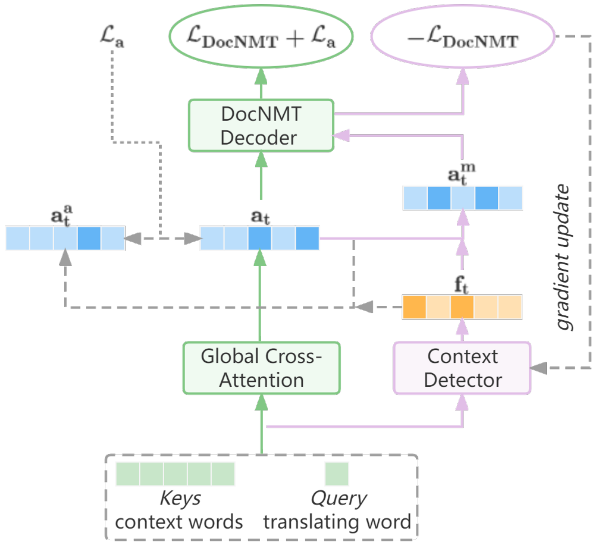


Figure 2. The proposed method's training framework. For simplicity, only the 6th layer is depicted. The proposed method is applied in the global cross-attention of 5th and 6th layers. The context detector is trained to find informative and detrimental contexts, and alignment loss $\mathcal{L}_a$ aims to make the attention weights focus more on informative contexts.

Figure 1 shows that on lower layers, the G-Transformer constrains both self-attention and target-to-source attention to concentrate on the current sentence being translated with a simple group tag method (*GroupAttn*).

$$\begin{aligned} &args = \{Q, K, V, tag\} \\ &\text{GroupAttn}(args) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + M(tag)\right)V \end{aligned} \tag{2}$$

where function $M(\cdot)$ works as an attention mask, excluding all tokens outside the sentence. This method allows local sentence representations to be obtained successfully.

On top layers, the model exploits standard multi-head attention as global attention (*GlobalAttn*), to capture contextual information of the document as a whole. GroupAttn and GlobalAttn are combined using a gate-sum module for integrating both intra-sentential and contextual information.

[1] Code for the proposed attention alignment method is open-sourced at https://github.com/KangZhong99/attention-alignment

3. METHODOLOGY

To tackle the inaccuracy and flatness of attention distribution in DocNMT, we propose a detection-based attention alignment method to enhance attention weights, focused on the informative context in model training. As shown in Figure 2, we first design a context detector (CD) to deteriorate performance by modifying the original attention weights generated by the backbone network. Specifically: we obtain learnable flags, and then apply them to the original attention weights. The CD parameters are updated based on the loss function that maximizes performance deterioration. This allows us to locate source-side words that are informative or meaningless for translation. Finally, we design an attention alignment strategy to calibrate the original attention weights, and integrate the cosine similarity between aligned and original attention weights into the loss function under the multi-task framework.

3.1. Context Detector

The CD is a lightweight network attached to the G-Transformer backbone. The basic assumption of CD is that a decrease in the attention weights of informative words and an increase in those of meaningless words reduces translation performance. In this paper, we make a tentative modification on the original attention weights using learnable flags, which results in a deterioration of performance. Specially, at the time step to produce t-th word in a sentence being translated, the learnable flag f_t is generated, based on the hidden states h_t^d from d-layer of DocNMT decoder ($d = 5\ or\ 6$ for G-Transformer):

$$f_t = tanh(W^f \cdot h_t^d + b^f) \tag{3}$$

where $tanh(\cdot)$ is the hyperbolic tangent function. W^f and b^f are trainable parameters vary among different attention layers and heads. The original attention weight a_t can then be modified to a_t^m:

$$a_t^m = a_t + f_t \odot \bar{a} \tag{4}$$

where $\odot$ denotes an element-wise multiplication, and $\bar{a}$ denotes a uniform distribution (an average vector of attention heads). Qualitatively, a positive flag results in the increase of the attention weight, while a negative flag results in a decrease. The modification operation aims to make smallest variations of attention weights that lead to most obvious performance deterioration. According to this purpose, the objective function of CD is designed as:

$$\mathcal{L}\left(\theta^f\right) = \underset{f_t \in [-1,1]}{\arg\min} \alpha \| \ f_t \ \|_2 - \mathcal{L}_{\text{DocNMT}}\left(a_t^m, \theta\right) \tag{5}$$

where $\theta^f = \left\{W^f, b^f\right\}$ is the parameters of CD and θ is the parameters of the original G-Transformer backbone. $\mathcal{L}_{DocNMT}\left(a_t^m, \theta\right)$ is the corss entropy loss of the DocNMT model based on the modified attention weight a_t^m. The hyper-parameter α serves as a regular term, encouraging most of the flags to be turned off, and alleviating the vanishing gradient problem of the hyperbolic tangent function. According to the second term, in order to deteriorate model prediction, the CD decreases the attention weights of informative words and increases those of meaningless words by setting corresponding flags to negative or positive, respectively.

3.2. Attention Alignment in Backbone Training

The CD can evaluate the importance of the source-side inputs for each output word. As analyzed above, a more negative (or positive) flag means the more beneficial (or harmful) the impact of corresponding contextual word. Here, we use the flag matrix f_t as supervised information to align the original attention weight a_t to *aligned attention weight* a_t^a.

$$a_t^a = a_t \odot e^{-f_t} \tag{6}$$

It is clear that the values of a_t^a increase in informative contextual words and decrease in meaningless ones. This is consistent with our expectations for accurate attention. a_t^a differs from a_t^m in Eq.4: the former has a positive effect on the final translation, while the latter has a negative effect.

We then use cosine similarity between a_t^a and a_t to construct an auxiliary alignment loss function:

$$\mathcal{L}_a\left(\theta\right) = -\mathcal{CS}\left(a_t^a, a_t\right) \tag{7}$$

Considering the attention mechanism is not well-trained at the early stage, and the final performance evaluation is contributed by the main training objective $\mathcal{L}_{DocNMT}\left(a_t, \theta\right)$, the DocNMT model is optimized by:

$$\begin{aligned} &\mathcal{L}\left(\theta\right) = \mathcal{L}_{DocNMT}\left(a_t, \theta\right) + \beta\left(s\right) * \mathcal{L}_a\left(\theta\right) \\ &\beta\left(s\right) = e^{-s/\tau} \end{aligned} \tag{8}$$

where s donates the training step, and τ is the decay time constant. A lager τ forces the model optimization to be affected by $\mathcal{L}_a$ for a longer time.

3.3. Training and Inference Issues

The proposed CD and backbone DocNMT model can be trained jointly. In a mini-batch, we first train the CD to find informative and meaningless contexts with the loss function in Eq.5, and then optimize the global attention mechanisms of backbone DocNMT following the objective function in Eq.8.

Model	TED		News		Europarl		Avg.	
	s-BLUE	d-BLEU	s-BLEU	d-BLEU	s-BlEU	d-BLEU	s-BlEU	d-BLEU
Transformer [1]	23.10	-	22.40	-	29.40	-	24.97	-
HAN [5]	24.58	-	25.03	-	28.60	-	26.07	-
SAN [12]	24.42	-	24.84	-	29.75	-	26.34	-
Flat-Transformer [8]	24.87	-	23.55	-	30.09	-	26.17	-
Context-Adaptive [13]	**26.77**	-	25.81	-	31.13	-	27.90	-
Our Sentence Baseline	24.79	-	25.28	-	31.33	-	27.13	-
G-Transformer [9]	25.10 (+0.31)	27.17	25.58 (+0.30)	27.11	32.34 (+1.01)	34.08	27.67 (+0.54)	29.45
Attention Calibration [16]	24.97 (+0.18)	27.12	25.20 (-0.08)	26.89	32.41 (+1.08)	34.12	27.53 (+0.40)	29.38
Proposed Method	25.47 (+0.68)	**27.65**	**26.09** (+0.81)	**27.66**	**32.67** (+1.34)	**34.35**	**28.08** (+0.95)	**29.89**

Table 1. Experiment results on three EN-DE datasets, where s-BLEU represents BLEU score for sentences, and d-BLEU [18] is the score for documents. Improvements over the sentence-level baseline are reported as “()”.

In practice, the parameter increments of CD are negligible (2M per layer and 4M in total), and can be removed in the inference procedure.

4. EXPERIMENTS

4.1. Datasets and Settings

We conduct experiments on widely used document-level parallel benchmark datasets, including three domains for English-German (En-De) translation: *TED* [19], *News* [20] and *Europarl* [21]. We split the documents into instances with up to 512 tokens. Moses [22] is used for data processing, and BPE [23] is used with vacab-size of 30K merges. Detailed statics for these datasets are shown in Table 2.

Data	**# of Docs**	**# of Sents/Doc**	**# of sents/Doc in our Inst**
TED	1.7K/92/22	123/98/105	18.3/18.5/18.3
News	6K/80/154	40/25/20	12.8/12.6/11.3
Europarl	118K/239/359	14/15/14	10.3/10.4/10.3

Table 2. Dataset statistics for Train/Valid/Test.

The DocNMT model is trained in two stages: first, a vanilla Transformer [1] base model is trained for sentence-level translation, then the DocNMT model is finetuned based on sentence baseline with document-level data. We use the same model configuration as Bao *et al.* [9] and train all models on 2 GeForce RTX 3090 GPUs. To finetune on the sentence baseline, we employ Adam ($\beta_1 = 0.9, \beta_2 = 0.998$) for parameters optimization with warmup steps of 2,000 for TED and 4,000 for News and Europarl. The dropout rate is 0.3 except for News (0.4). The models are trained with the batch size of 32K tokens for all datasets. The batch size of 32K tokens is achieved by using the batch size of 4096 tokens and updating the model for every 8 batches. Following Bao *et al.* [9], we apply word dropout [24, 25] to the inputs with p = 0.1. We set hyper-parameters as $\alpha = 1.5$ and $\tau = 10^4$ in Eq.5 and Eq.8 respectively.

We mainly build four systems for comparison: the Sentence Baseline, the G-Transformer [9], Attention Calibration [16] and the Proposed Method.

4.2. Results and Analysis

The results of our experiments are listed in Table 1, along with those of other DocNMT models. Finetuned on a strong sentence-level baseline, the performance of G-Transformer improves on three datasets by 0.54 s-BLEU points on average. Applying the proposed method to G-Transformer to refine contextual attention calculation further improves the gain from 0.54 to 0.95 s-BLEU points, and achieves the best results among all those listed on the two lager datasets, News and Europarl.

The hyper-parameter α in Eq.5 decides the values of generated flags, and influences the concentration degree of global attention by Eq.8. The entropy values of global attention and s-BLEU on News, with respect to different α values, are depicted in Figure 3. The s-BLEU score presents a negative association with contextual attention entropy, and reaches its climax when α is 1.5. This verifies our hypothesis that translating the current word only replies on very few contextual words and that sharpening the focus of document-level attention is beneficial to model prediction.

Figure 4 (a) depicts the entropy curves, with respect to different training epochs on vanilla G-Transformer and the proposed method. Both curves show a convergence trend in global attention. The curve of the former retains stable during training, however, which means the model captures a significant amount of meaningless contextual information. As for the proposed method, the curve falls sharply for about 20 epochs, and then retains at a relatively low level. Figure 4 (b) plots the attention distribution of final converged models. In our method, the model is more confident about the choice of informative contextual words, and focuses on very few of them compared to the vanilla G-Transformer.

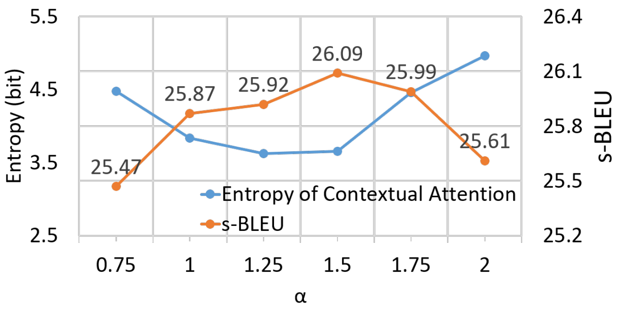
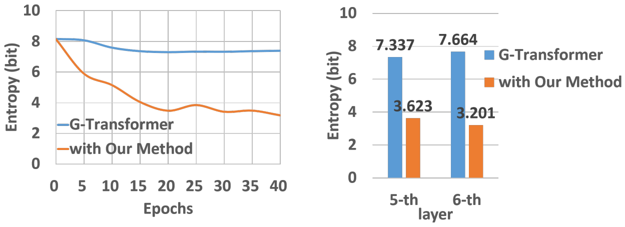


Figure 3. Entropy of global cross-attention on valid set and s-BLEU on test set, with a different hyper-parameter α on News.



(a) Development (b) Distribution

Figure 4. Comparison of global cross-attention entropy with G-Transformer on News.

To futher investigate whether our model is able to improve the translation of discourse phenomena, we conducted linguistic evaluation on deixis and ellipsis using a English-Russian (En-Ru) linguistic contrastive test set [26]. To make a fair comparison, we follow [26] to use 6M sentence-level instances to train sentence-level baseline and 1.5 M document-level instances to train our models (G-Transformer and the proposed method). Results are reported in Table 3. The proposed method achieves improvements on deixis and ellipsis(VP) compared with G-Transformer. This indicates that our method can make better use of context to deal with discourse phenomena.

The upper box of Table 4 shows an example where our method correctly translates the pronoun "it" (highlighted in bold). To analyze the effect of our attention alignment approach, we print two particular attention heads, one from the G-Transformer and one from our method. The source-side surrounding contexts with assigned attention weights at the time step after the prediction "teilt" are shown in the lower box of Table 4. Both heads focus more on the contextual reference "love", which is related to the query "it". In contrast, our method reduces the disturbance of irrelevant context words, like "Remi", and "differences".

Model	deixis	ellipsis(infl.)	ellipsis(VP)
sent[†]	50.0	53.0	28.4
concat[†]	83.5	47.5	76.2
CADec[†]	81.6	58.1	76.2
G-Transformer	89.7	84.7	82.2
proposed method	**90.3**	84.9	**83.5**

Table 3. Linguistic evaluation results on the contrastive test set (accuracy). [†] indicates that the results are borrowed from the original paper [26].

Src: He shares **it** regardless.
Ref: Er teilt **sie** unbekümmert.
G-Transformer: Er teilt **es** unabhängig davon.
Proposed Method: Er teilt **sie** unabhängig davon.

Attention to reference word *love*
G-Transformer (Head 3):
... Remi knows what love is . He shares it regardless .
He doesn't care about religious differences , and gets ...
Proposed Method (Head 3):
... Remi knows what love is . He shares it regardless .
He doesn't care about religious differences , and gets ...

Table 4. Example of pronoun disambiguation. In German, if the reference of "it" is feminine (e.g., love), "sie" is used. Otherwise, it is "er" (masculine) or "es" (neutral). The intensity of color represents the attention given to a specific word.

5. CONCLUSION

In this paper, we propose a detection-based attention alignment method to force the DocNMT model to increase the focus of each word on relevant key contextual words. We design a lightweight context detector to evaluate the importance of each contextual word, which proves simple yet effective. Then, we introduce a strategy that uses this supervised information to guide attention alignment. As demonstrated, the proposed method observably reduces the entropy values of document-level attention, and consistently improves on the strong G-Transformer baseline. More encouragingly, our work provides valuable reference on self-supervised learning for improved or more-focused attention in other long-input generation frameworks.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving hu-

man parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

[3] Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3983–3989, 2021.

[4] Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, 2022.

[5] Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 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, 2018.

[6] Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, 2019.

[7] Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 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 (EMNLP-IJCNLP)*, pages 1527–1537, 2019.

[8] Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3505–3511, 2020.

[9] Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, 2021.

[10] Mingzhou Xu, Liangyou Li, Derek F Wong, Qun Liu, and Lidia S Chao. Document graph for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8435–8448, 2021.

[11] Marcin Junczys-Dowmunt. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, 2019.

[12] Sameen Maruf, André FT Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL-HLT*, pages 3092–3102, 2019.

[13] Linlin Zhang, Zhirui Zhang, Boxing Chen, Weihua Luo, and Luo Si. Context-adaptive document-level neural machine translation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6232–6236. IEEE, 2022.

[14] Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, 2020.

[15] Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, 2019.

[16] Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1288–1298, 2021.

[17] Rico Sennrich and Martin Volk. Iterative, mt-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic conference of computational linguistics (NODALIDA 2011)*, pages 175–182, 2011.

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[19] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 28–30 2012. European Association for Machine Translation.

[20] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.

[21] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005.

[22] Philipp Koehn et al. 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, 2007.

[23] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[24] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, August 2016. Association for Computational Linguistics.

[26] E Voita, R Sennrich, and I Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. The Association for Computational Linguistics, 2019.