applied sciences

MDPI

Article

Grammatically Derived Factual Relation Augmented Neural Machine Translation

Fuxue Li [1,2,*], Jingbo Zhu [1], Hong Yan [2,*] and Zhen Zhang [2]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China; zhujingbo@mail.neu.edu.cn
[2] College of Electrical Engineering, Yingkou Institute of Technology, Yingkou 115014, China; zhangzhen@yku.edu.cn
* Correspondence: lifuxue119@163.com (F.L.); yanhong@yku.edu.cn (H.Y.); Tel.: +86-0417-3588555 (F.L. & H.Y.)

Featured Application: This paper introduces factual relation information into Transformer-based neural machine translation to improve translation quality.

Abstract: Transformer-based neural machine translation (NMT) has achieved state-of-the-art performance in the NMT paradigm. This method assumes that the model can automatically learn linguistic knowledge (e.g., grammar and syntax) from the parallel corpus via an attention network. However, the attention network cannot capture the deep internal structure of a sentence. Therefore, it is natural to introduce some prior knowledge to guide the model. In this paper, factual relation information is introduced into NMT as prior knowledge, and a novel approach named Factual Relation Augmented (FRA) is proposed to guide the decoder in Transformer-based NMT. In the encoding procedure, a factual relation mask matrix is constructed to generate the factual relation representation for the source sentence, while in the decoding procedure an effective method is proposed to incorporate the factual relation representation and the original representation of the source sentence into the decoder. Positive results obtained in several different translation tasks indicate the effectiveness of the proposed approach.

Keywords: factual relation; augmented; neural machine translation; prior knowledge

Citation: Li, F.; Zhu, J.; Yan, H.; Zhang, Z. Grammatically Derived Factual Relation Augmented Neural Machine Translation. *Appl. Sci.* **2022**, *12*, 6518. https://doi.org/10.3390/app12136518

Academic Editor: Valentino Santucci

Received: 30 May 2022
Accepted: 24 June 2022
Published: 27 June 2022

1. Introduction

In recent years, neural network-based models have consistently delivered better-quality translations than those generated by phrase-based systems. Transformer-based [1] neural machine translation has achieved state-of-the-art performance in neural machine translation, and it outperforms recurrent neural network (RNN)-based models [2–4]. However, recent work [5–7] has shown that the Transformer may not learn the linguistic information to the greatest extent possible due to the characteristics of the model, especially in low-resource scenarios. On the other hand, prior knowledge has been proved to be an effective way to improve the quality of statistical machine translation. Therefore, this issue has become a hot spot in the field of NMT research, with existing linguistic knowledge being utilized to alleviate the inherent difficulties faced by NMT.

In addition, Sennrich [8] has proved that linguistic information is beneficial to neural translation models. Many researchers have focused on incorporating prior knowledge into neural machine translation to improve translation quality, such as character or word structure [9–11], phrase structure [12,13], syntactic structure [14–16], and so on. However, previous methods based on linearized or convolutional neural networks often face difficulties in choosing suitable sequences to balance training efficiency and sufficient syntactic information [17–19]. Moreover, most syntax-aware NMT models are limited to a qualitative use of syntactic information, while using syntax distance [20] to quantify syntactic relationships coarsely will bring quantization noise into NMT. Different from previous work,

this paper proposes a Factual Relation Augmented method which avoids the noise caused by the quantization of syntactic information to improve Transformer-based NMT. Factual relations in a sentence depict the relationship of several different entities, which can be regarded as the core meaning of the sentence.

In fact, factual relation information is the target output of the information extraction [21] model, which can derive entity, relationship, event, and other factual information from a given sentence, and the result can be depicted by multiple relational tuples. In other words, the factual information in relation tuples describes the semantic relationship between two or more entities in a sentence. The factual relation tuples are given for an example sentence in Figure 1.

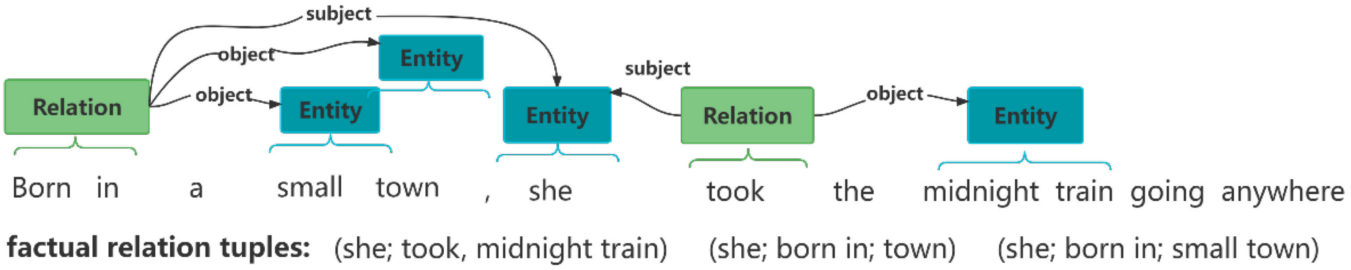


Figure 1. Factual relation tuples extracted from a sentence.

From the perspective of the NMT model, the generation of the current target word only focuses on target words that have been generated and all source words. Source words make the same contribution to the model regardless of their importance. In other words, the model lacks a mechanism to guarantee that it can pay more attention to essential words in the source sentence. Chen [22] proposed an approach which makes the model pay more attention to content words than functional words. Inspired by this, we improve the NMT model by paying more attention to the factual relation information in a source sentence which captures the core words more than the content words. The method is simple yet effective, improving translation quality with few additional parameters and less computational overhead. The main contributions of this paper are as follows:

1. To our best knowledge, it is the first attempt at using factual relation information to improve Transformer-based neural machine translation.
2. This paper introduces an effective method (Factual Relation Augmented, abbreviated as FRA) for the NMT model which utilizes factual relation information in source sentences to improve the translation quality of the NMT model.
3. This FAR method can improve the translation quality of the Transformer-based model, especially for complex sentences (e.g., complex clause sentences).

The main steps of FRA are as follows: First, the factual relation tuples are extracted by Stanford CoreNLP [23] from the source sentence in a parallel corpus. Secondly, a factual relation mask matrix is constructed by factual relation tuples. Next, the factual relation representation is generated in the encoding procedure. Finally, when predicting the translation, factual relation-decoder attention is introduced to guide the decoder.

This paper is structured as follows: Section 2 introduces Transformer-based neural machine translation. Section 3 describes the Factual Relation Augmented approach (FRA). Section 4 gives an overview of the experiments and results, and Section 5 introduces related work. Section 6 concludes and offers prospects for future work.

2. Transformer-Based Neural Machine Translation

In Transformer-based NMT [1], the encoder is composed of a stack of N identical layers, and each layer is composed of a multi-headed self-attention network (ATT) and a fully connected feed-forward network (FFN). A residual connection [24] is applied between the sub-layers, and layer normalization (LayNorm) [25] is performed. Formally, the i-th identical layer of this stack is as follows:

$$\widetilde{H}^i = LayNorm\left(ATT^{i-1}\left(Q^{i-1},\ K^{i-1},\ V^{i-1}\right) + H^{i-1}\right) \tag{1}$$

$$C^i = LayNorm\left(FFN^i\left(\widetilde{H}^i\right) + H^i\right) \tag{2}$$

where Q^{i-1}, K^{i-1} and V^{i-1} represent the query, key and value vectors transformed from the $(i-1)$-th layer. Similar to the encoder, the decoder shares a similar architecture, having an additional encoder–decoder attention block sandwiched between the self-attention and FFN blocks.

$$\widetilde{S}_t^i = LayNorm\left(ATT_d^i\left(Q_t^{i-1}, K_t^{i-1},\ V_t^{i-1}\right) + S_t^{i-1}\right) \tag{3}$$

$$C_t^i = LayNorm\left(ATT_c^i\left(\widetilde{S}_t^i, K_e^L,\ KV_e^L\right) + \widetilde{S}_t^i\right) \tag{4}$$

$$S_t^i = LayNorm\left(FFN_d^i\left(C_t^i\right) + C_t^i\right) \tag{5}$$

where Q_t^{i-1}, K_t^{i-1} and V_t^{i-1} are transformed from the $(i-1)$-th layer S^{i-1} into time-step t. K^L and V^L are transformed from the L-th layer of the encoder. The top layer of the decoder S_t^i is used to predict the next target word:

$$P(y_i|y_{<i}) \propto \exp\left(W_o \tan h\left(W_w S_t^i\right)\right) \tag{6}$$

where W_o and W_w are weight matrices which can be learned during the training procedure.

3. Factual Relation Augmented Approach

In the open information extraction task, factual relation tuples are extracted from structured and unstructured data. In this paper, this information is extracted from source sentences in the parallel corpus by Stanford CoreNLP. In order to design a neural machine translation model that is efficient in training and exploits factual relation tuples while producing high-quality translations, we based our model on the Transformer architecture [1].

The architecture of the Factual Relation Augmented method is shown in Figure 2. First of all, a factual relation mask matrix is conducted by the factual relation tuples extracted by Stanford CoreNLP. Next, the encoder utilizes the factual relation mask matrix to generate a factual relation representation which can be seen as an enhanced representation of the original representation in the encoding procedure. Finally, both the factual relation representation and the original representation of the source sentence are fused into the integrated layer, and four methods are proposed for the integrated layer.

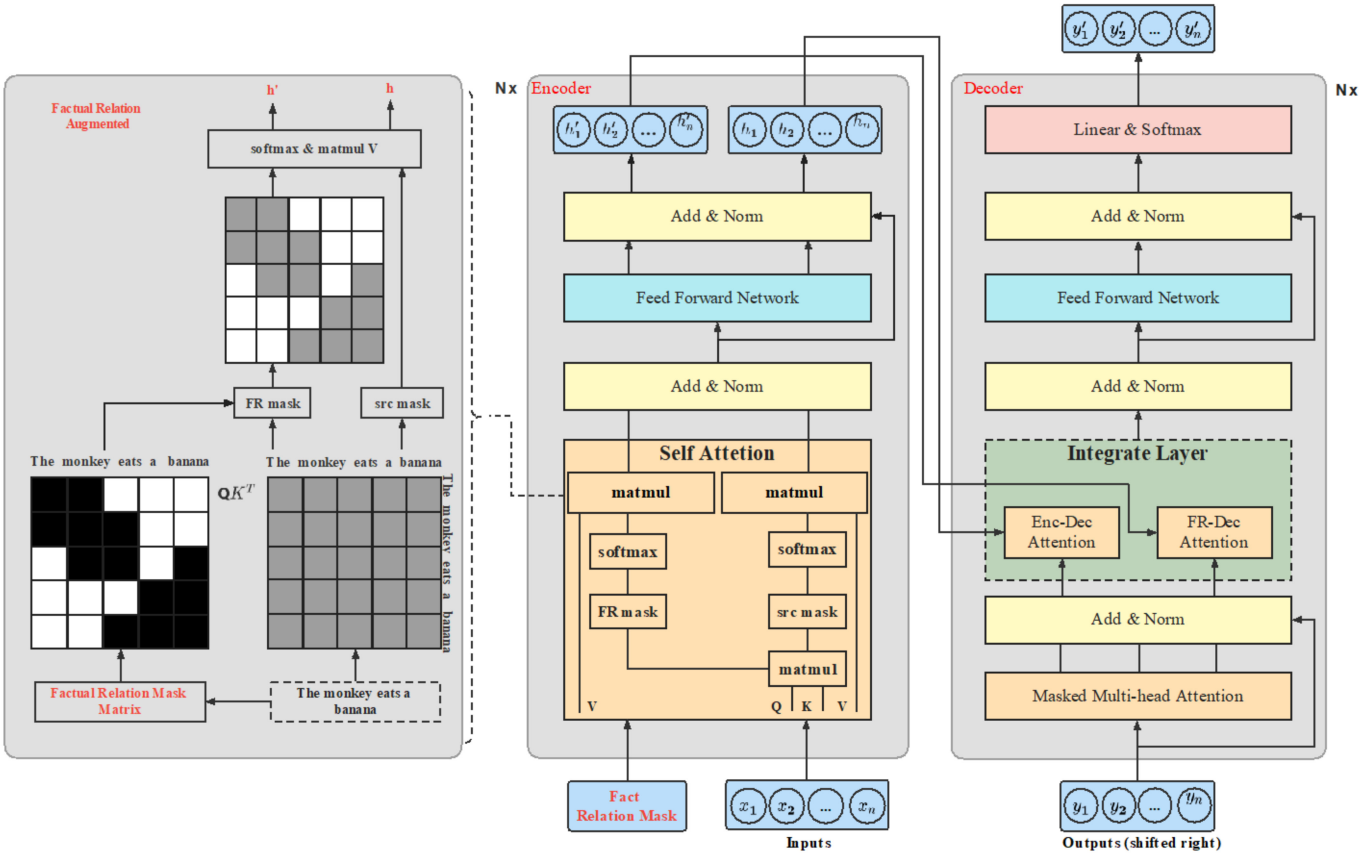


Figure 2. The architecture of the Factual Relation Augmented Transformer.

3.1. Generating the Factual Relation Mask Matrix (FRMM)

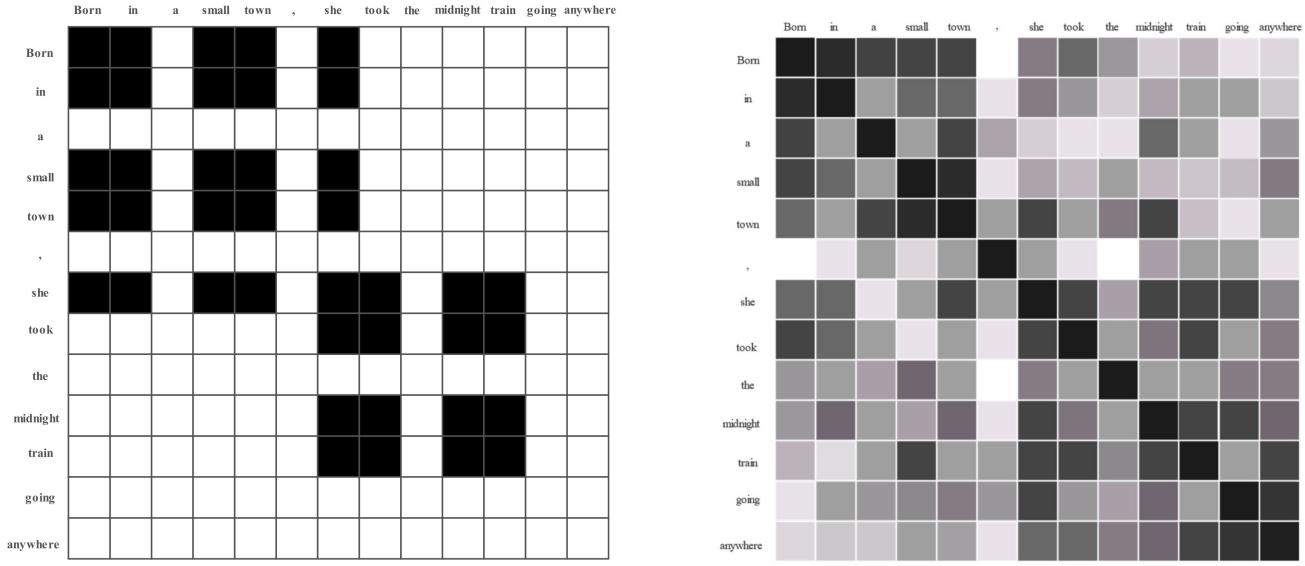
When the source-side factual relation tuples are generated by Stanford CoreNLP, the challenge is how to incorporate the factual relation tuples into the Transformer-based NMT. Inspired by masked multi-headed attention, this paper introduces a factual relation perception module in the encoder. Specifically, a factual relation mask matrix (FRMM) is constructed by the factual relation tuples extracted by Stanford CoreNLP, and the generation steps of the FRMM are as follows:

Step 1: Obtain all the factual relation tuples set from the source sentence which contains words;

Step 2: Construct a matrix M^s, which has l rows and l columns;

Step 3: For each word in a factual relation tuple, retain the relevant part (value is set as one) and discard the irrelevant part (value is set as zero) in M^s.

A factual relation mask matrix derived from the example in Figure 1 is shown in Figure 3a. Different from the attention matrix in Figure 3b, which is derived by the self-attention mechanism in the Transformer, the factual relation mask only focuses on the words which exist in the factual relation tuples, while the attention matrix pays attention to all words in this sentence. In the attention matrix, the darker the color in the matrix, the higher the correlation is, and the correlation is represented by a value which ranges from 0 to 1.



(**a**) (**b**)

Figure 3. An example of factual relation mask matrix (**a**) in left and attention matrix (**b**) in right.

3.2. Factual Relation Augmented Encoder

After the FRMM is ready, it works on the attention weight matrix M^A, which has l rows and l columns, and then a new factual relation attention matrix M' is generated:

$$M' = M^s \odot M^A \tag{7}$$

where $\odot$ denotes a point-wise multiplication operation. M' can be seen as an attention matrix which incorporates factual relation information from the source sentence into the encoder for the Transformer. The formula is as follows:

$$Attention(Q,\ K,\ V) = softmax\left(\frac{FR(QK^T)}{\sqrt{d}}\right)V \tag{8}$$

where FR denotes the factual relation mask operation. The output of the encoder consists of two parts, the representation of the source sentence $h = h_1, h_2 \ldots h_n$ and the representation of the factual relation of the source sentence $h' = h'_1, h'_2 \ldots h'_n$. In practice, some details should be noticed. First of all, multiple factual relation tuples may be extracted from one sentence. Naturally, an entity may exist in different factual relation tuples; for instance, in the example in Figure 1, the entity "she" exists multiple times in three factual relation tuples. However, the proposed method only considers whether the word appears when generating the factual relation mask matrix and does not consider the number of times it appears. Second, for the calculation of the factual relation representation embedding h', no additional parameters are introduced during the training procedure for the Transformer. All parameters in the encoder are shared with the original encoder in the Transformer.

3.3. Factual Relation Augmented Decoder

The factual relation decoder has a similar structure to the decoder in the Transformer, except for the introduction of an integrated layer between the masked multi-headed attention mechanism and the feed-forward network layers. The integrated layer consists of two parts: FR-Dec attention and Enc–Dec attention, and the structure of the integrated layer is shown in Figure 2. The factual relation-decoding interaction attention is introduced to guide the decoder. In other words, the decoder utilizes the representation of factual relation augmented in the source sentence to optimize the generation of the target translation dur-

ing the decoding procedure. Give the source sentence and the target translation sentence $t = t_1, t_2 \ldots t_n$, the factual relation augmented decoding procedure can be summarized as follows:

$$S = Attention_{self}\left(E_t W^Q,\ E_t W^K, E_t W^V\right) \tag{9}$$

$$\hat{S} = Attention_{ed}(h, h, S) \tag{10}$$

$$\widetilde{S} = Attention_{fr}(h', h', S) \tag{11}$$

where E_t denotes the word embedding of the target input, $Attention_{self}$ denotes self-attention, $Attention_{ed}$ denotes encoding–decoding attention and $Attention_{fr}$ denotes factual relation-decoding attention. The challenge we encounter is how to incorporate $\widetilde{S}$ and $\hat{S}$ in the integrated layer. To be specific, four interpolation approaches are proposed to address the issue.

Linear Interpolation. $\widetilde{S}$ can be seen as an enhancement representation derived from prior knowledge for $\hat{S}$. In this manner, the interpolation can be calculated as follows:

$$H = \hat{S} + \lambda * \widetilde{S} \tag{12}$$

where λ is a hyper-parameter whose value is [0, 1]. Note that no additional parameters are introduced during the training procedure in this method.

Gate Learning. Compared with linear interpolation, we propose a general method which can adjust the hyper-parameter dynamically during the training procedure rather than in empirical settings.

$$g = sigmoid\left(W_{\hat{S}}\left(\hat{S}; \widetilde{S}\right) + b_{\hat{S}}\right) \tag{13}$$

$$H = g * \hat{S} + (1 - g) * \widetilde{S} \tag{14}$$

where $\left(\hat{S}; \widetilde{S}\right)$ denotes the operation of concatenation, and $W_{\hat{S}}$ and $b_{\hat{S}}$ are the introduced parameters which can be learned during the training procedure.

Concat Gate Learning. Inspired by LSTMM [26], we propose an output gate and a forget gate to learn the final representation of the source context. It can be summarized as:

$$f = sigmoid\left(W_f\left(\hat{S}; \widetilde{S}\right) + b_f\right) \tag{15}$$

$$o = W_o\left(\hat{S}; \widetilde{S}\right) + b_o \tag{16}$$

$$H = o * f + W_f\left(\hat{S}; \widetilde{S}\right) * (1 - f) \tag{17}$$

where W_f and W_o are weight matrixes, b_f and b_o are biases, and all parameters can be learned during the training procedure.

Linear Transformation Linear Transformation can be described as follows:

$$H = W_{\hat{S}}\left(\hat{S}; \widetilde{S}\right) + b_{\hat{S}} \tag{18}$$

where $W_{\hat{S}}$ and $b_{\hat{S}}$ can be learned during the training procedure. Note that H, $\hat{S}$ and $\widetilde{S}$ have the same dimensions.

4. Experiment

In order to demonstrate the effectiveness of the approach proposed in this paper, several experiments were carried out for different translation tasks. The proposed models were implemented using the fairseq toolkit [27]. All models were run on a machine with GeForce GTX 3080.

4.1. Datasets and Settings

Datasets The proposed methods were evaluated on IWSLT15 English–Vietnamese (En–Vi), WMT18 English–Turkish (En–Tr) and IWSLT14 German–English (De–En). For IWSLT15 En–Vi, we used the pre-processed datasets provided by Luong [28] and used tst2012 as the dev set and test on tst2013. For NC11 (News Commentary v11) En–De, De–En and WMT18 En–Tr, the settings of the experiment were consistent with [29]. For IWSLT14 De–En, we followed the pre-processing steps described in [30]. The number of sentences in each dataset is shown in Table 1.

Table 1. Number of sentences in each dataset.

Corpus	IWSLT14	IWSLT15	WMT18
Training Set	160,239	133,158	207,678
Valid Set	7283	1553	3000
Test Set	6750	1268	3007

Settings The byte pair encoding algorithm [31] was applied to encode all sentences to limit the size of the vocabulary to 40 K. In order to alleviate the problem of large vocabulary and sparse vocabulary, this paper adopts the factual relation to the sub-words, and it defines the relationship between sub-words as related when the corresponding original word is related. In terms of the parameters for the Transformer, we choose the Adam optimizer, with $\beta 1 = 0.9$, $\beta 1 = 0.98$, $\varepsilon = 10^{-9}$ and the learning rates setting strategy, which are all the same as [1]; the configurations were identical to those in [32]. We used a beam search decoder for all translation tasks with a beam width of 5. For all translation tasks, we used the checkpoint that had the best valid performance on the valid set and the case-insensitive 4-g BLEU (Bilingual Evaluation Understudy) score as the primary evaluation metric, adopting the script multi-bleu.perl in the Mose toolkit.

This paper compares the proposed approach with some related work, e.g., the Mixed Enc and Multi-Task systems proposed by Currey [33], PASCAL [29] (Parent-Scaled Self-Attention) proposed by Bugliarello and Okazaki, parameter optimization for Multi-Task, incorporating the syntactic information as label-dependent into the Transformer encoder word-embedding matrix proposed by Sennrich and Haddow [8], and LISA (combining self-attention with syntactic parsing), proposed by Strubell [34].

In addition, this paper compares some other machine translation methods on the IWSLT task, e.g., ELMo (Embeddings from Language Models), CVT (Cross-View Training) [32], SAWR (Syntax-Aware Word Representations) [35] and Dynamic Conv [30]; Tied-Transform [36] and Macaron [37]; C-MLM (Conditional-Masked Language Modeling) [38] and the BERT (Bidirectional Encoder Representations from Transformers)-fused model [39].

4.2. Main Results and Analysis

4.2.1. Performance on Different Datasets

As shown in Table 2, incorporating the dependent labels (+S&H) and the multi-task approach (+Multi-Task) in the word-embedding representation did not issue in significant improvements to the baseline model. However, the methods (+LISA and +PASCAL) utilized the syntactic information to improve the attention and showed better results. This paper utilizes the factual relation tuples that can be seen as prior knowledge to improve the NMT model. Compared with the strong baseline, the experiments showed surprising results and achieved BLEU improvements of +1.6, +0.91, and +1.5 on WMT18 (En–Tr) and NC11 (En–De and De–En) translation tasks, respectively. The reason for this was that the factual relation information includes the core meaning of the sentence and it is integrated into the decoder during the decoding procedure directly. With the help of this valuable prior knowledge, the decoder generates better translation results.

Table 2. Performance of different machine translation approaches on various datasets.

Model	WMT18	NC11		Model	IWSLT14 (De–En)		IWSLT15 (En–Vi)	
	En–Tr (Test Set)	En–De (Test Set)	De–En (Test Set)		Valid Set	Test Set	Valid (Test 2012)	Test (Test 2013)
/	/	/	/	ELMo [32]	/	/	/	29.30
/	/	/	/	SWAR [35]	/	/	/	29.09
Mixed Enc [33]	9.60	/	/	CVT [32]	/	/	/	29.60
Multi-Task [33]	10.60	/	/	C-MLM [38]	36.93	35.63	27.85	31.51
Transformer [1]	13.13	25.00	26.60	Transformer [1]	35.27	34.41	27.45	30.76
+Multi-Task [8]	14.00	24.80	26.70	Tied-Transform [36]	/	35.53	/	/
+S&H [8]	13.00	25.50	26.80	Dynamic Conv [30]	/	35.20	/	/
+LISA [24]	13.60	25.30	27.10	Macaron [37]	/	35.40	/	/
+PASCAL [29]	14.00	25.90	27.40	BERT-fused [39]	/	36.11	/	/
Our Approach	14. 73$^{\uparrow}$	25.91$^{\uparrow}$	28.10$^{\uparrow}$	Our Approach	36.89$^{\uparrow}$	36.10	27.90$^{\uparrow}$	31.53$^{\uparrow}$

In terms of the IWSLT translation task, the approach proposed in this paper also performed well, achieving competitive results compared to other machine translation models that are well designed. Tied-Transformer conducts a slight model by sharing the encoder, while it takes a long time for the model to converge. In contrast, the FRA approach proposed in this paper achieved comparable performance without additional training time. For Macaron, adding a feed-forward network before the attention of each layer resulted in a large number of parameters being introduced for this model, increasing the cost of calculation. In contrast, FRA showed better performance while only a few parameters were introduced, but it achieved the best BLEU score of 31.53 on the IWSLT15 En–Vi translation task. Compared with FRA proposed in this paper, C-MLM and BERT-fused incorporated the pre-training language model (BERT) into the Transformer, resulting in a long period of training being required for the model.

4.2.2. Evaluating Hyper-Parameters for Linear Interpolation

For the Linear Interpolation method, we studied the effect on WMT18 (En–Tr) and NC11 (En–De and De–En) by varying the value of λ. The results of three test sets with different hyper-parameter values for λ are shown in Figure 4. When λ increased from 0 to 0.4, the BLEU scores improved by +1.23, +0.57 and 1.11 over the baseline model, respectively. This means that the proposed method is effective for the NMT model. Subsequently, larger values of λ reduced the BLEU scores, which indicated that excessive biased factual relation information may be weak when translating the target sentence.

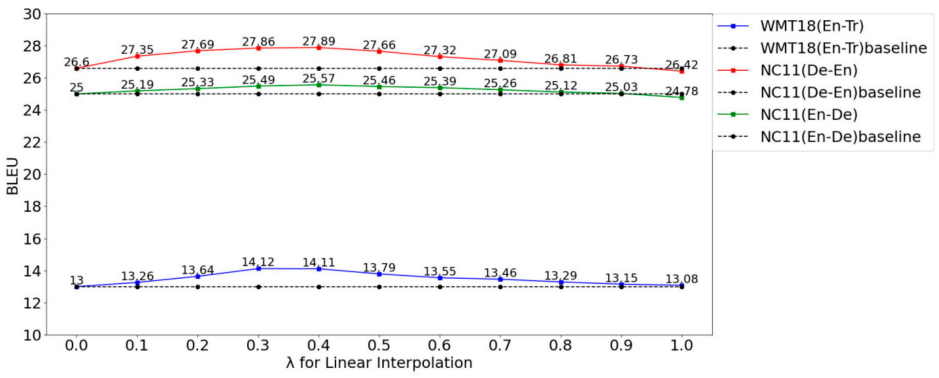


Figure 4. BLEU scores of the Linear Interpolation method on the WMT18 (EN-Tr), NC11 (En–De) and NC11 (De–En) test sets with different values of λ. The dashed line denotes the result of the baseline.

4.2.3. Comparison of Different Interpolation Methods

The linear interpolation method is effective, but whether this method is the best cannot be guaranteed. Therefore, four methods were proposed to incorporate factual relation-decoding interaction attention and encoder–decoder attention into the integrated layer. As shown in Table 3, CGL damaged the model's performance both in the valid set and test set, but LI, GL and LT improved translation quality. LT showed outstanding performance compared to the other methods. A possible reason for this is as follows: the LI method improved the blue score slightly. The simple interpolation method may have resulted in losing the diversity of the source sentence representation and did not improve the model obviously. CGL may not distinguish the representation of concatenation between two representations, and it does not learn which parts of the representation need to be activated and which parts need to remain unchanged. In contrast, GL and LT succeeded in incorporating factual relation information into the integrated layer and achieved the goal of guiding the decoder by means of factual relation. In short, these two methods optimized the representation of the model and improved translation quality. Therefore, LT was selected to perform other experiments in this paper.

Table 3. BLEU values for different interpolation methods on the IWSLT15 (En–Vi) task.

	Method	Valid Set (tst2012)	Test Set (tst2013)
1	Transformer (baseline)	27.45	30.76
2	Linear Interpolation (LI)	27.49	30.79
3	Gate Learning (GL)	27.13$^{\uparrow}$	31.34$^{\uparrow}$
4	Concat Gate Learning (CGL)	27.19	30.43
5	Linear Transformation (LT)	27.91$^{\uparrow}$	31.53$^{\uparrow}$
6	Compare Fusion	27.57	31.12

4.2.4. Performance by Sentence Length

As shown in Figure 5, the FRA proposed in this paper is more useful when translating long sentences; it obtained more than +1.5 BLEU points when translating long sentences in all the experiments and achieved +2.51 BLEU points on the En–Tr pair, although, admittedly, only a few sentences (1.9% 3.2%) in the evaluation datasets were long. This phenomenon is consistent with our expectations. Generally speaking, the structure of long sentences is more complicated, and factual relation information can guide the model so that it pays more attention to the important clues about sentence relations between several entities, thereby improving performance.

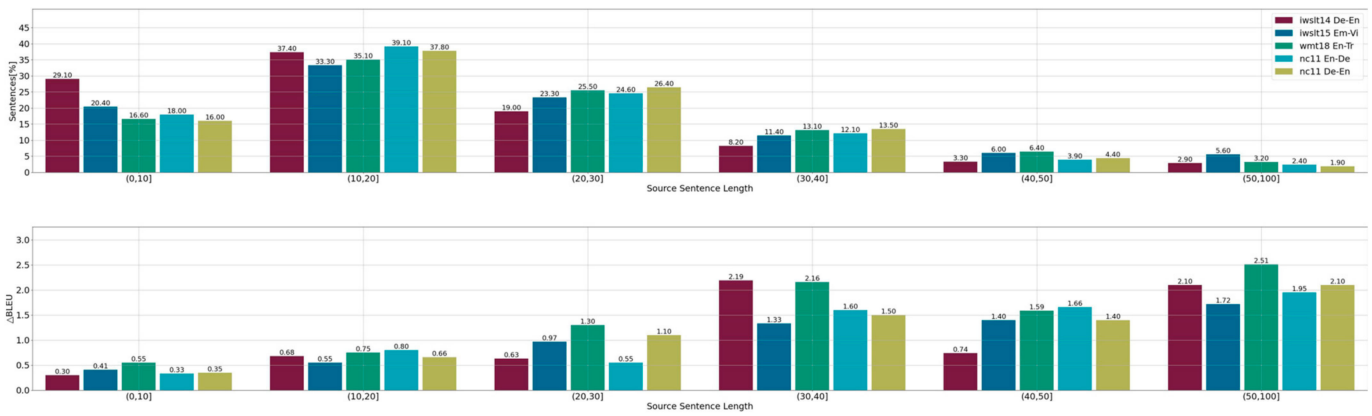


Figure 5. Percentage data (**above**) and ΔBLEU scores (**below**) for the Transformer.

4.2.5. Effect of Factual Relation Augmented

Although the proposed method shows good performance, it cannot be proved that the improvement is contributed by factual relation information. A possible candidate is the Linear Transformation operation in the decoder. To address this issue, we replaced the factual relation representation with the original representation of the source sentence (let $h' = h$ in Figure 2) and tested the performance of the proposed method. The experimental result is shown in Table 3 (line 6: Compare Fusion). Linear Transformation shows better performance than baseline and poorer performance than the Factual Relation Augmented method. There is a large gap between the two methods. Consequently, it has been demonstrated that the Factual Relation Augmented method is effective for the NMT model.

4.2.6. Performance on Different Layers

Previous studies have pointed out that different features can be captured by different layers [40]. As a result, some experiments were carried out on the IWSLT15 (En–Vi) translation task to test the performance on different layers, and the results are shown in Table 4. Note that 1~6 denotes incorporating factual relation information into all layers in the decoder.

Some details should be noticed. Compared with the baseline, all methods showed improved performance to different degrees. The best performance for all methods was achieved by incorporating factual relation knowledge into the sixth layer (top layer). The improvement in BLEU score achieved was +0.77, which demonstrates the effectiveness of the proposed method. In addition, the integration of the top layer is of greater benefit more than that of the bottom layer. This is mainly because knowledge of factual relations can provide more context, which can be regarded as an effective enhancement to the representation of the source language. This is consistent with the discovery that the lower layer is biased toward paying attention to semantics, while the higher layer is biased toward paying attention to contextual information [40].

Table 4. Performance of different integration layers on the IWSLT15 En–Vi task.

Layer	Tst2013	Layer	Tst2013
Baseline	30.76	1~6	31.02
1 (bottom)	31.02	1~2	31.13
2	31.22	1~3	31.00
3	31.17	1~4	31.07
4	31.29	1~5	31.08
5	31.51	4~6	31.40
6 (upper)	31.53$^{\uparrow}$	5~6	31.47

The effect of single-layer interpolation is more obvious than that of multi-layer interpolation. With the increase in the layers from bottom to top, the performance of the model does not change obviously, but there is a decreasing contrast. The phenomenon shows that the integration of multi-layers cannot improve the performance of the model. However, this introduces more parameters, which leads to redundancy in the model, and this is not conducive to learning the information contained in training data. Therefore, in the other experiments in this paper, the top layer (the sixth) was selected to incorporate the factual relation knowledge into the integrated layer. In contrast, the other layers in the decoder were consistent with the original decoder in the Transformer.

4.2.7. Analysis and Thinking

The Transformer is based on a standard end-to-end structure and only relies on the parallel corpus. It assumes that all the information can be learned by the self-attention mechanism automatically. From the perspective of self-attention, we hold that self-attention can be seen as wide and soft attention, which ensures the generalization ability of the model. On this bias, the explicit addition of factual relation information can be regarded as

restricting and hard attention to focus on the sentence itself. The experiments show that integrating factual relations and the original soft attention optimizes the representation of sentences without compromising the model's generalization.

5. Related Work

The incorporation of prior knowledge into the NMT model is a hot topic in the field of neural machine translation research. Chatterjee [41] introduced the possibility to guide the translation procedure, with constraints provided as XML annotations of the source words with the corresponding translations. For RNN-based models, Sennrich and Haddow [8] utilized linguistic information (lemmas, morphological features, POS tags, syntactic dependency labels) to enrich the embedding layer of the encoder in the attentional encoder–decoder architecture. Cohn [42] incorporated structural alignment bias information (absolute positional bias, fertility, relative positions bias, alignment consistency) into the attention model. Eriguchi [14] proposed an end-to-end syntactic NMT model to explicitly take the syntactic structure into consideration in a tree-based encoder. Chen [20] created a dependency unit for each source word to capture long-distance dependency constraints and then designed an encoder with a convolutional architecture to jointly learn SDRs and source dependency annotations. Different from the studies mentioned above, in our system, prior knowledge was introduced as a feature into the translation model, while factual relation information was vectorized and integrated into the Transformer, and the performance of the proposed approach outperformed that described in [8] (+S&H and +Multi-Task).

Some methods have introduced syntactic information concisely without changing the structures of models. Saunders [43] interleaved words with syntax representation, resulting in longer sequences. Currey and Heafield [33] introduced constituency parsing information into the Transformer NMT model, then proposed a multi-task model for low-resource data and a mixed encoder model for rich-resource data. Zhang [35] vectorized the source-side syntax to embedding and concatenated it with the intermediate representations. Bugliarello and Okazaki [29] optimized the attention weight when encoding the source sentences according to the syntactic distance between words (+PASCAL) and achieved state-of-the-art results. Wu [19] proposed an approach to incorporate syntax into NMT with a Transformer model which utilized source-side and target-side dependency relations to improve NMT. Peng [44] introduced external syntax knowledge to guide the learning of the attention network. Experimental comparisons with the approach proposed here are shown in Table 2. Our approach showed better performance than Mixed Enc and Multi-Task, especially in the translation of longer sentences.

Chen [22] proposed a content word-aware model which utilizes word frequency information to distinguish between content and function words. Similar to this work, both methods pay attention to essential words in the encoding procedure. The content word-aware model estimates the importance of words by the TF-IDF algorithm, while our approach pays more attention to the words that exist in factual relation information in sentences, which is simple and effective.

6. Conclusions

This paper has proposed a Factual Relation Augmented (FRA) approach to improve Transformer-based machine translation. Furthermore, we have proposed a method to integrate factual relation information (prior knowledge) in the Transformer and utilize factual relations to guide the decoder to generate target translations during the decoding procedure. Experiments were carried out to prove the usefulness of the Factual Relation Augmented method in several translation tasks, and the results showed good performance in several datasets, especially for longer sentences. Although the proposed approach is simple, it is efficient.

In future work, much valuable research could be undertaken, e.g., comparisons could be made with other models and approaches, the training time and efficiency of translation

using the FRA approach could be further assessed, and questions of how factual relation information affects the translation procedure and how factual relation information in target sentences can be utilized to improve translation quality with the NMT model can be explored.

Author Contributions: Conceptualization, F.L. and Z.Z.; methodology, F.L.; validation, H.Y. and Z.Z.; formal analysis, F.L. and J.Z.; writing—original draft preparation, F.L. and H.Y.; writing—review and editing, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Liaoning Province of China, grant number: 2021-YKLH-12; the Scientific Research Foundation of Liaoning Province, grant number: L2019001; the High-level talents research project of Yingkou Institute of Technology, grant number: YJRC202026; and the Doctor Startup Foundation of Liaoning Province, grant number: 2020-BS-288.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: The study did not involve humans.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here (https://www.statmt.org/wmt14/index.html; https://www.statmt.org/wmt15/index.html; https://www.statmt.org/wmt18/index.html; https://wit3.fbk.eu/).

Acknowledgments: We would like to thank the anonymous reviewers for their insightful comments, as well as Chuncheng Chi, Zhongchao Zhao and Beibei Liu for their helpful advice on how to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017; pp. 6000–6010.
2. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7 May 2015.
3. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, Montreal, QC, Canada, 8 December 2014; pp. 3104–3112.
4. Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014; pp. 103–111.
5. Raganato, A.; Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 287–297.
6. Tang, G.; Müller, M.; Gonzales, A.R.; Sennrich, R. Why self-attention? A targeted evaluation of neural machine translation architectures. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 4263–4272.
7. Tran, K.M.; Bisazza, A.; Monz, C. The importance of being recurrent for modeling hierarchical structure. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 4731–4736.
8. Sennrich, R.; Haddow, B. Linguistic input features improve neural machine translation. In Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, Berlin, Germany, 7 August 2016; pp. 83–91.
9. Chen, H.; Huang, S.; Chiang, D.; Dai, X.; Chen, J. Combining character and word information in neural machine translation using a multi-level attention. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1 June 2018; pp. 1284–1293.
10. Wang, F.; Chen, W.; Yang, Z.; Xu, S.; Xu, B. Hybrid attention for Chinese character-level neural machine translation. *Neurocomputing* **2019**, *358*, 44–52. [CrossRef]
11. Zhang, L.; Komachi, M. Neural machine translation of logographic language using sub-character level information. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, 31 October 2018; pp. 17–25.
12. Wang, X.; Tu, Z.; Xiong, D.; Zhang, M. Translating phrases in neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7 September 2017; pp. 1421–1431.
13. Dahlmann, L.; Matusov, E.; Petrushkov, P.; Khadivi, S. Neural machine translation leveraging phrase-based models in a hybrid search. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7 September 2017; pp. 1411–1420.

14. Eriguchi, A.; Hashimoto, K.; Tsuruoka, Y. Tree-to-sequence attentional neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7 August 2016; pp. 823–833.
15. Chen, H.; Huang, S.; Chiang, D.; Chen, J. Improved neural machine translation with a syntax-aware encoder and decoder. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July 2017; pp. 1936–1945.
16. Aharoni, R.; Goldberg, Y. Towards string-to-tree neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July 2017; Short Papers. Volume 2, pp. 132–140.
17. Chen, K.; Wang, R.; Utiyama, M.; Liu, L.; Tamura, A.; Sumita, E.; Zhao, T. Neural machine translation with source dependency representation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7 September 2017; pp. 2846–2852.
18. Hashimoto, K.; Tsuruoka, Y. Neural machine translation with source-side latent graph parsing. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7 September 2017; pp. 125–135.
19. Wu, S.; Zhang, D.; Zhang, Z.; Yang, N.; Li, M.; Zhou, M. Dependency-to-dependency neural machine translation. *IEEE/ACM Trans. Audio Speech Lang. Processing* **2018**, *26*, 2132–2141. [CrossRef]
20. Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Zhao, T. Syntax-directed attention for neural machine translation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2 February 2018; pp. 4792–4799.
21. Angeli, G.; Premkumar, M.J.J.; Manning, C.D. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26 July 2015; pp. 344–354.
22. Chen, K.; Wang, R.; Utiyama, M.; Sumita, E. Content word aware neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5 July 2020; pp. 358–364.
23. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The stanford corenlp natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22 June 2014; pp. 55–60.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June 2016; pp. 770–778.
25. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *Stat* **2016**, *1050*, 21.
26. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
27. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. Fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, 2 June 2019; pp. 48–53.
28. Luong, M.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17 September 2015; pp. 1412–1421.
29. Bugliarello, E.; Okazaki, N. Enhancing machine translation with dependency-aware self-attention. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5 July 2020; pp. 1618–1627.
30. Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y.; Auli, M. Pay less attention with lightweight and dynamic convolutions. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 2018.
31. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7 August 2016; pp. 1715–1725.
32. Clark, K.; Luong, M.; Manning, C.D.; Le, Q. Semi-supervised sequence modeling with cross-view training. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 1914–1925.
33. Currey, A.; Heafield, K. Incorporating source syntax into transformer-based neural machine translation. In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1 August 2019; Research Papers. Volume 1, pp. 24–33.
34. Strubell, E.; Verga, P.; Andor, D.; Weiss, D.; McCallum, A. Linguistically-informed self-attention for semantic role labeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 5027–5038.
35. Zhang, M.; Li, Z.; Fu, G.; Zhang, M. Syntax-enhanced neural machine translation with syntax-aware word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2 June 2019; pp. 1151–1161.
36. Xia, Y.; He, T.; Tan, X.; Tian, F.; He, D.; Qin, T. Tied transformers: Neural machine translation with shared encoder and decoder. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 2 February 2019; Volume 33, pp. 5466–5473.
37. Lu, Y.; Li, Z.; He, D.; Sun, Z.; Dong, B.; Qin, T.; Wang, L.; Liu, T. Understanding and improving transformer from a multi-particle dynamic system point of view. In Proceedings of the ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations, Addis Ababa, Ethiopia, 26 April 2020.
38. Chen, Y.; Gan, Z.; Cheng, Y.; Liu, J.; Liu, J. Distilling knowledge learned in bert for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5 July 2020; pp. 7893–7905.

39. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T. Incorporating bert into neural machine translation. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6 May 2019.
40. Anastasopoulos, A.; Chiang, D. Tied multitask learning for neural speech translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1 June 2018; pp. 82–91.
41. Chatterjee, R.; Negri, M.; Turchi, M.; Federico, M.; Specia, L.; Blain, F. Guiding neural machine translation decoding with external knowledge. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7 September 2017; pp. 157–168.
42. Cohn, T.; Hoang, C.D.V.; Vymolova, E.; Yao, K.; Dyer, C.; Haffari, G. Incorporating structural alignment biases into an attentional neural translation model. In Proceedings of the NAACL-HLT, San Diego, CA, USA, 12 June 2016; pp. 876–885.
43. Saunders, D.; Stahlberg, F.; de Gispert, A.; Byrne, B. Multi-representation ensembles and delayed sgd updates improve syntax-based nmt. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15 August 2018; Short Papers. Volume 2, pp. 319–325.
44. Peng, R.; Lin, N.; Fang, Y.; Jiang, S.; Zhao, J. Boosting neural machine translation with dependency-scaled self-attention network. *arXiv* **2021**, arXiv:2111.