Self-supervised Bilingual Syntactic Alignment for Neural Machine Translation

Tianfu Zhang,[1] **Heyan Huang,** [1] **Chong Feng,** [1*] **Longbing Cao,** [2]
[1] Beijing Institute of Technology
[2] University of Technology Sydney
{tianfuzhang,hhy63,fengchong}@bit.edu.cn, LongBing.Cao@uts.edu.au

Abstract

While various neural machine translation (NMT) methods have integrated mono-lingual syntax knowledge into the linguistic representation of sequence-to-sequence, no research is available on aligning the syntactic structures of target language with the corresponding source language syntactic structures. This work shows the first attempt of a source-target bilingual syntactic alignment approach SyntAligner by mutual information maximization-based self-supervised neural deep modeling. Building on the word alignment for NMT, our SyntAligner firstly aligns the syntactic structures of source and target sentences and then maximizes their mutual dependency by introducing a lower bound on their mutual information. In SyntAligner, the syntactic structure of span granularity is represented by transforming source or target word hidden state into a source or target syntactic span vector. A border-sensitive span attention mechanism then captures the correlation between the source and target syntactic span vectors, which also captures the self-attention between span border-words as alignment bias. Lastly, a self-supervised bilingual syntactic mutual information maximization-based learning objective dynamically samples the aligned syntactic spans to maximize their mutual dependency. Experiment results on three typical NMT tasks: WMT'14 English→German, IWSLT'14 German→English, and NC'11 English→French show the SyntAligner effectiveness and universality of syntactic alignment.

Introduction

Neural Machine Translation (NMT) has made significant progress by developing MT-oriented deep neural translators, including recurrent neural network (RNN) (Hochreiter and Schmidhuber 1997), convolutional neural network (CNN) (Kim 2014), Transformer (Vaswani et al. 2017), embedding (Mukherjee et al. 2020; Piazza 2020) and their variants. In MT, syntactic knowledge has shown essential for extracting and learning the effective linguistic representations from both source-target sequences as in both statistical machine translation (SMT) (Koehn et al. 2003) and NMT (Bugliarello et al. 2020; Eriguchi et al. 2016; Hao et al. 2019). For example, in (Eriguchi et al. 2016), a tree-to-sequence model with the source phrase structure explicitly exploits a constituency tree to guide the decoder to generate a translated word by weakly aligning it with phrases and words of source sentences. The graph-convolutional networks (GCN) (Bastings et al. 2017) adds layers to the standard encoder of NMT models to explicitly model source dependency-based word representation. A multi-granularity self-attention mechanism (Hao et al. 2019) randomly modifies several attention heads in Transformer to attend to phrase modeling in either n-gram or syntactic formalization. The above work shows improved translation by leveraging the source syntax information.

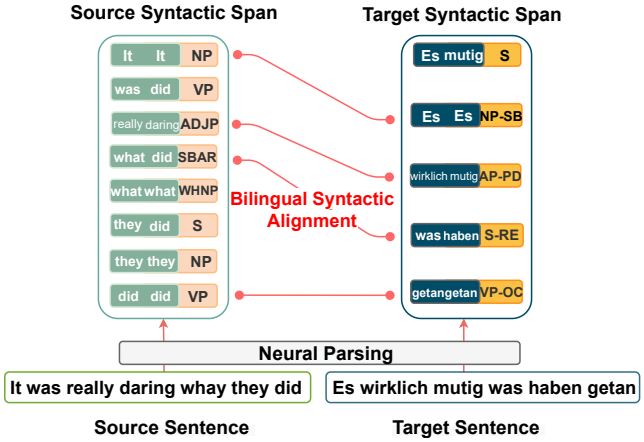

Figure 1: An example shows the deep structural mapping between source-target syntactic spans-based embeddings in their latent spaces.

However, the existing syntax-aware NMT models only integrate monolingual syntax, omitting the correspondence and divergence between source and target's syntactic structures. Inspired by linguistics, there may be certain bilingual syntactic consistency and alignment between source and target languages, which could be used as a unique linguistic characteristic of sequence-to-sequence NMT. Fig. 1 illustrates the bilingual syntactic alignment between the source (English) and target (German) sentences in the granularity of their syntactic spans. A span is a triple-set syntactic unit

*Corresponding author

with border-word positions and a constituent label to represent a constituency tree. These syntactic spans are generated by a neural constituency parsing model (Kitaev et al. 2018). Each source or target syntactic span vector is concatenated by border-word hidden states and the constituent label embedding. While each language has its own syntactic span collocation and constituency labels, there exists aligned (e.g, between (*really*, *daring*, 'ADJP') in English and (*wirklich*, *mutig*, 'AP-PD') in German) and unaligned (e.g, (*was*, *did*, 'VP') in English and (*Es*, *mutig*, 'S') in German) between the source-target linguistic syntactic structures.

In fact, to date, our understanding and application of the source-target bilingual syntactic relationship in NMT are very limited, and it is challenging to capture the bilingual syntactic relations. There usually exists complex and flexible syntax divergences across different languages, such as Subject-verb-object orders (German V.S English), Special Interrogative Word orders (Chinese V.S English), and Pronoun Shedding (Japanese). Hence, there is no universal linguistic rule or golden-data to characterize the bilingual syntactic relationship. The existing monolingual syntactic NMT models cannot handle these challenges. Fortunately, the sequence-to-sequence NMT architecture is naturally suitable to be a platform to align the bilingual syntactic structure in unsupervised way. Inspired by Fig. 1, it is potential to identify and align bilingual syntactic spans and maximize their mutual dependency for syntactic alignment. Then, the bilingual syntactic relationships could be modeled by means such as between Transformer encoder and decoder word hidden states, and an abstract linguistic alignment could be made between source and target sentences on top of the traditional word alignment for NMT.

Motivated by the above analysis, this paper proposes a self-supervised bilingual syntactic alignment approach SyntAligner to precisely align the source-target bilingual syntactic structures in the high-dimensional deep space, and then maximizes the mutual information between the aligned bilingual syntactic structure samples for translation. First, SyntAligner represents the respective syntactic span vectors of source and target sentences by combining sequential representation and syntactic representation. Second, SyntAligner introduces a Border-Sensitive Span Attention (BS-SA) mechanism to characterize the alignment relationship between source and target syntactic spans. Considering the potential noise incorporated by the syntax parser, we evaluate the confidence of syntactic span by leveraging the self-attention between the syntactic span start and end border words from the Transformer encoder (and decoder). These attention weights are then to bias the alignment attention distribution to obtain the final alignment attention matrix. Subsequently, SyntAligner samples the aligned syntactic span pairs with high attention scores by a curriculum learning strategy to sample the aligned syntactic span pairs with a gradually increasing scale for better training the span alignment. Lastly, a self-supervised objective function Bilingual Syntactic Mutual Information Maximization (MIM) optimizes the NMT model for maximizing the bilingual syntactic mutual dependency.

We test SyntAligner on three widely-used translation tasks WMT'14 English→German, IWSLT'14 German→English, and NC'11 English→French. Extensive analyses reveal that the NMT with bilingual syntactic alignment effectively improves the translation performance. We also visualize the bilingual syntactic span alignment and the mutual information variation process as to why the SyntAligner insight.

Background

Transformer

Transformer is a primary sequence-to-sequence model for NMT. It consists of a stack of layers both in the encoder and decoder. Each layer first learns the scale-dot product self-attention to extract information from the whole sentence, and then forms a point-wise feed-forward network to provide nonlinearity. The self-attention is formulated as:

$$Attention(Q, K, V) = softmax(\frac{QK^{\intercal}}{\sqrt{d_e}})V, \quad (1)$$

where d_e is the dimension of the hidden representation and is set as the embedding size. For the self-attention inside the encoder, $Q, K, V \in \mathbb{R}^{M \times d_e}$, while for the self-attention inside the decoder, $Q, K, V \in \mathbb{R}^{N \times d_e}$. For the attention that bridges the encoder and decoder, $Q \in \mathbb{R}^{N \times d_e}$ and $K, V \in \mathbb{R}^{M \times d_e}$. The feed-forward network consists of two linear projections with a ReLU activation in between:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2, \quad (2)$$

where x is the input representation. Both the self-attention and feed-forward networks are wrapped by the residual connection (He et al. 2016) to form a sublayer:

$$Sublayer(x) = Block(LayerNorm(x)) + x, \quad (3)$$

$LayerNorm()$ is the layer normalization (Ba et al. 2016) for self-attention.

Our SyntAligner substantially extends Transformer to respectively leverage the word hidden states from encoder and decoder to construct the bilingual syntactic span sequences, as shown in Fig. 2. It consists of novel modules to incorporate syntax of bilingual spans and to maximize their mutual information to capture bilingual syntactic dependency.

Constituency Syntactic Span

To align the bilingual syntactic trees and measure their mutual dependency on a sufficient sample scale, we aim to find an appropriate syntactic granularity to represent the syntactic structure of a sentence. A constituency tree (Kasami 1965) $Tree$ is a collection of labeled spans over a sentence,

$$Tree := \{(l_i, (s_i, e_i))\}, \quad (4)$$

where $i = 1, ..., I$, s_i and e_i represent the start and end border words of the syntactic span $Tree_i$, and l_i is the corresponding constituent label. Recently, many new methods improve the span-based neural constituency parsing (Cross et al. 2016; Kitaev et al. 2018; Stern et al. 2017) by following an encoder-decoder architecture. In this paper, we adopt the syntactic span sequences generated by the well-performed Berkerly Neural Parser[1] as our constituency parsing tree.

[1]https://github.com/nikitakit/self-attentive-parser

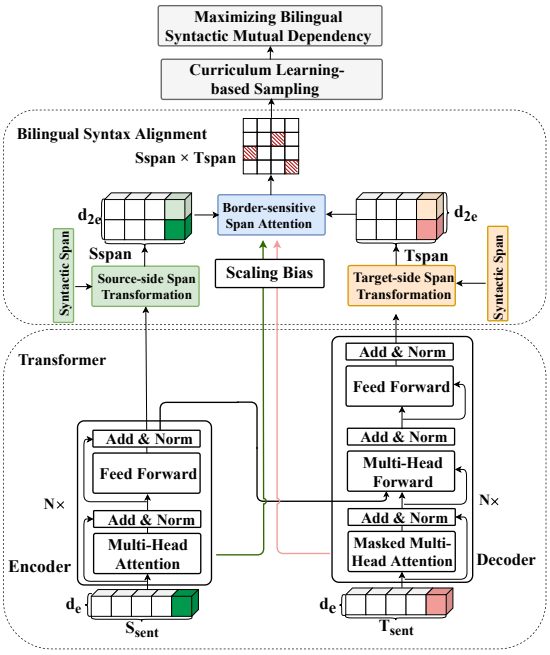

Figure 2: The architecture of SyntAligner for mutual information maximization-based bilingual syntactic alignment. The red blocks in the BS-SA matrix are the aligned syntactic span pairs.

The SyntAligner

Fig. 2 shows the architecture of our proposed SyntAligner for bilingual syntactic alignment. SyntAligner firstly represents bilingual syntactic spans and then aligns them by a border-sensitive span attention (BS-SA) mechanism with the border-word self-attention weights from Transformer's encoder and decoder as the scaling biases. SyntAligner further includes a self-supervised objective function for bilingual syntactic mutual information maximization between the transformed spans, which maximizes the lower bound of mutual information for sampling aligned bilingual syntactic span pairs, then jointly train SyntAligner with a cross-entropy training objective for NMT.

Aligning Bilingual Syntaxes

Syntactic Span Representation Given a source input word-based sequence $S_{sent} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, their hidden state-based sequence $H_{sent} = \{h_1, h_2, ..., h_M\}$ can be obtained from the top layer of Transformer encoder. Then, with the span sequence of source language by Eq. 4, we obtain the source syntactic span sequence $S_{span} = \{s_1, s_2, .., s_I\}$, where each source syntactic span vector $s_i^x = [(\mathbf{x}_{e_i} - \mathbf{x}_{s_i}) : l_i^x]$. We represent the syntactic span vector by combining the span border word hidden state $\mathbf{x}_{e_i} - \mathbf{x}_{s_i}$ with its constituency label embedding l_i^x of the source language, where $\mathbf{x}_{s_i}, \mathbf{x}_{e_i}, l_i^x \in \mathbb{R}^{d_e}$ and $s_i^x \in \mathbb{R}^{d_{2e}}$. Similarly, we can generate the target syntactic span sequence $T_{span} = \{s_1^y, s_2^y, .., s_J^y\}$ by using the output from the top-layer of decoder, where each target syntactic span vector $s_j^y = [(\mathbf{y}_{e_j} - \mathbf{y}_{s_j}) : l_j^y]$.

BS-SA: Border-sensitive Span Attention To align the source and target syntactic structures in the granularity of span, we propose the Border-sensitive Span Attention (BS-SA) mechanism, which reflects the alignment relationship between each source and target span pair's vectors s_i^x and s_j^y by using the scale-dot product attention method:

$$Span_{Attn}(S_{span}, T_{span}) = softmax(\frac{S_{span} \cdot {T_{span}}^{\intercal}}{\sqrt{d_{2e}}}), \tag{5}$$

where the attention weight matrix $Span_{Attn}(S_{span}, T_{span}) \in \mathbb{R}^{I \times J}$. Eq. 5 enforces to align bilingual syntactic spans, and we choose the aligned syntactic span pairs with higher attention scores than the alignment gate η. We will introduce the details later.

Furthermore, as noise is likely introduced by the syntax parser, BS-SA evaluates the confidence of each syntactic span S_{span} (or T_{span}) according to the correlation between its border words $\mathbf{x}_{e_i}$ and $\mathbf{x}_{s_i}$ (or $\mathbf{y}_{e_j}$ and $\mathbf{y}_{s_j}$). In Transformer, the self-attention mechanism can capture the linguistic relationship between words, especially in the top layers. A higher attention score represents a stronger relationship between words. We thus refer to the word self-attention scores to reflecting the confidence of relationship between syntactic span border words. Specifically, for the source word self-attention $S_{Attn} \in \mathbb{R}^{M \times M}$ (from Eq. 1), we firstly find the corresponding border- word attention scores for each source syntactic span and generate a confidence matrix $S_{Conf} \in \mathbb{R}^{I \times J}$, where each row i has the same J confidence biases of source syntactic span s_i. Next, we set a scaling weight W_s multiplying to S_{Conf} to widen the confidence gap between source syntactic spans. Then, we element-wisely multiply the S_{Conf} with $Span_{Attn}$ to obtain a BS-SA-Src attention matrix as follows:

$$\begin{aligned} BS_{Src}(S_{span}, T_{span}) = & softmax(W_s \times S_{Conf} \\ & \odot softmax(\frac{S_{span} \cdot {T_{span}}^{\intercal}}{\sqrt{d_{2e}}})). \end{aligned} \tag{6}$$

A higher confidence score will enlarge the difference in the distribution of each row of $Span_{Attn}$ after the outermost $softmax$ normalization. We further leverage the target syntactic span confidence vector $T_{Conf} \in \mathbb{R}^{I \times J}$ obtained from the target word-masked self-attention (the bottom sub-layer in decoder) $T_{Attn} \in \mathbb{R}^{N \times N}$, where each column j has the same I confidence biases of target syntactic span $\mathbf{y}_j$. In practice, for each row n in T_{Attn}, the attention of future words $[T_{attn_{n+1}}, .., T_{attn_N}]$ will be masked to avoid accessing the future information. To obtain the attention scores between any words, we use the unmasked version T'_{Attn} for BS-SA. Meanwhile, it also has a scaling weight W_t. We can bias each column of $Span_{Attn}$ to produce a BS-SA-Tgt attention matrix as follows:

$$\begin{aligned} BS_{Tgt}(S_{span}, T_{span}) = & softmax(W_t \times T_{Conf} \\ & \odot softmax(\frac{S_{span} \cdot {T_{span}}^{\intercal}}{\sqrt{d_{2e}}})). \end{aligned} \tag{7}$$

Compared to the source confidence bias, the target confidence bias changes the final span attention matrix more directly, because the outermost $softmax$ function normalizes the span attention matrix in the row dimension, and the target confidence bias adjusts the attention score for each column element. We deeply analyze the effect of scaling weight on BS-SA in the Experiment section.

Curriculum Learning-based Sampling To obtain more accurate syntactic span pairs as positive samples for better bilingual mutual dependency, we use the BS-SA mechanism to align the bilingual syntactic spans after the NMT pretraining. Further, we adopt the curriculum learning idea to sample syntactic pairs from easy to difficult and from less to more. Accordingly, we set a time-based exponential decay alignment gate η as follows:

$$\eta = \eta_0 \times \eta_d^{\left\lfloor \frac{max(step_n - step_s + step_d, 0)}{step_d} \right\rfloor}, \quad (8)$$

where $\eta_0 \in (0, 1]$ represents the initial gate, $\eta_d \in (0, 1]$ represents the decay gate, $step_n, step_s, step_d$ represent the current training steps, starting decay steps, and decay interval steps, respectively. Based on Eq. 8, at the initial training steps, we only sample the aligned syntactic pairs with the highest span attention (e.g., $\eta = 1$) to guarantee the quality of alignment at the cost of sampling scale. As the BS-SA network captures more accurate syntactic alignment, we lower the threshold of sampling to get more positive samples for mutual information maximization.

Finally, for the bilingual syntactic span sequences S_{span}, T_{span}, we can sample the aligned bilingual syntactic span pairs $B_{span} = \{(s_i^x, s_j^y)\}$.

Maximizing Bilingual Syntactic Mutual Dependency

To generate more syntax-consistent translations, it is necessary to strengthen the syntactic correspondence between the source and target syntactic span vectors. We leverage the concept of mutual information to measure the dependency between random variables and maximize their mutual dependency. Formally, the mutual information between an aligned syntactic pair (s_i^x, s_j^y) is:

$$I(s_i^x, s_j^y) = H(s_i^x) - H(s_i^x|s_j^y) = H(s_j^y) - H(s_j^y|s_i^x) \quad (9)$$

where $H(\cdot)$ denotes the entropy. We choose a particular lower bound InfoNCE, which is based on Noise Contrastive Estimation (NCE; (Gutmann et al. 2012)) . The InfoNCE bound of an aligned syntactic pair (s_i^x, s_j^y) is defined as:

$$\begin{aligned} &I(s_i^x, s_j^y) \geq \\ &\mathbb{E}_{p(s_i^x, s_j^y)}[f_\theta(s_i^x, s_j^y) \\ &- \mathbb{E}_{q(\widetilde{s}_j^y)} \log \sum_{\widetilde{s}_j^y \in \widetilde{\mathcal{S}}} exp f_\theta(s_i^x, \widetilde{s}_j^y)] + \log \left|\widetilde{S}\right|, \end{aligned} \quad (10)$$

where $f_\theta \in \mathbb{R}$ is a function parameterized by θ (e.g., a dot product between bilingual syntactic spans), and $\widetilde{s}_j^y$ is a target syntactic span from a sample set $\widetilde{\mathcal{S}}$ drawn from a proposal distribution $q(\widetilde{S})$. $\widetilde{\mathcal{S}}$ consists of the positive sample s_j^y (the aligned target syntactic spans) from the current target sentence and $|\widetilde{S}| - 1$ negative samples (unaligned target syntactic spans) from all of the sentences. In practice, we constrain the negative samples into the current training batch, hence the size of the negative set is still manageable.

We use the contrastive learning framework to design a task that maximizes the mutual information between the aligned source and target syntactic spans with unaligned target syntactic spans as negative samples. Accordingly, the self-supervised objective function $\mathcal{J}_{MIM}$ is:

$$\mathcal{J}_{MIM} = -\mathbb{E}_{p(s_i^x, s_j^y)} \left[s_i^{x\intercal} s_j^y - \log \sum_{\widetilde{s}_j^y \in \widetilde{\mathcal{S}}} exp(s_i^{x\intercal} \widetilde{s}_j^y) \right], \quad (11)$$

In comparison with the traditional NMT supervised cross-entropy training objective:

$$\mathcal{J}_{CE}(\theta) = \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \theta), \quad (12)$$

our bilingual syntactic alignment-oriented objective function $\mathcal{J}_{BLS-MIM}$ below combines the self-supervised objective function $\mathcal{J}_{MIM}$ with the above-supervised objective function $\mathcal{J}_{CE}$ to train the NMT model. It is a weighted combination of the two terms above:

$$\mathcal{J}_{BLS-MIM} = \lambda_{MIM}\mathcal{J}_{MIM} + \lambda_{CE}\mathcal{J}_{CE}(\theta) \quad (13)$$

where λ_{MIM} and λ_{CE} are hyperparameters that balance the contribution of each term.

The general mutual information calculation will be provided with stable positive samples and negative samples, then it maximizes the mutual dependency between positive samples and minimizes the relationship between negative samples. However, a big challenge is that we do not know which syntactic source-target pair is correctly aligned in advance. Hence, the confidence of our positive sample (s_i^x, s_j^y) is dependent on the BS-SA accuracy. To improve the accuracy of BS-SA, we firstly pretrain the NMT model by using E.q. 12 to convergence to obtain more well-trained word hidden states from the encoder and decoder to construct bilingual syntactic span vectors. Then we fine-tune the NMT model using the joint training objective as in E.q. 13.

Experiments

Settings

We test SyntAligner on three language translation tasks: WMT14 English→German (En→De), IWSLT14 German→English (De→En), and WMT14 News Commentary version 11 (NC11) English→French (En→Fr). For the En→De translation, the training data consists of 4.5M sentence pairs (newstest2013 and newstest2014 as the validation set and sets). For the De→En translation, the training set consists of 160K sentence pairs and we randomly draw 7K samples from the training set as the validation set. We concatenate dev2010, dev2012, tst2010, tst2011 and tst2012 as the test set. For the En→Fr translation, the training

Architecture	En→De		De→En		En→Fr	
	#Para	BLEU	#Para	BLEU	#Para	BLEU
Transformer (Vaswani et al. 2017)	88.0M	27.31	52.8M	33.63	97.6M	28.35
LightConv (Wu et al. 2019)	n/a	n/a	n/a	34.80	n/a	n/a
MG-SA (Hao et al. 2019)	89.9M	28.28	n/a	n/a	n/a	n/a
Transformer-Base	88.0M	27.61	52.8M	34.43	97.6M	28.54
+ SyntAligner-Base	90.2M	28.29	52.8M	$35.05^{\uparrow}$	97.6M	$28.85^{\uparrow}$
+ SyntAligner-Src	90.2M	28.22	52.9M	$35.01^{\uparrow}$	97.6M	$29.12^{\Uparrow}$
+ SyntAligner-Tgt	90.2M	**28.56**$^{\Uparrow}$	52.9M	**35.13**$^{\Uparrow}$	97.6M	$28.96^{\uparrow}$
+ SyntAligner-Bi	90.2M	$28.42^{\Uparrow}$	52.9M	$35.11^{\uparrow}$	97.6M	**29.28**$^{\Uparrow}$

Table 1: Testing Results of SyntAligner against Syntax-enhanced baselines Including Transformer for NMT on WMT14 En→De, IWSLT14 De→En, and NC11 En→Fr. "# Para" denotes the trainable parameter size of each model (M = million). Symbols "↑ /⇑" refer to the improvement significance level over the self-attention baseline ($p < 0.05/0.01$).

data consists of 180K sentence pairs (newstest2013 and newstest2014 as validation and test sets). We evaluate our approach in terms of different languages and data sizes.

The baselines include: **Transformer** as a strong baseline with the state-of-the-art performance; **LightConv** as a simpler but effective baseline; and **MG-SA** as a latest source syntax-integrated method which modifies the partial heads of the self-attention networks of Transformer encoder to capture the syntactic phrase representation. The results in their papers are reported here. We also implement a Transformer-Base by using OpenNMT toolkit, which outperforms the original Transformer in (Vaswani et al. 2017).

Several SyntAligner variants with/out BS-SA mechanism are compared. " +SyntAligner-Base" is the SyntAligner without border-word self-attention to influence the BS-SA alignment distribution. Various BS-SA mechanisms are applied to generate: " +SyntAligner-Src" with self-attention of source syntactic span border-words; " +SyntAligner-Tgt" with self-attention of target syntactic span border-words; and " +SyntAligner-Bi" with self-attention of both source and target syntactic span border-words. These greatly lift the base BS-SA mechanism for small and large language pairs.

SyntAligner, its variants and all the baselines on top of the advanced Transformer model (Vaswani et al. 2017) are implemented by using the open-source toolkit OpenNMT (Klein et al. 2017). We follow the Transformer (base model) setting in (Vaswani et al. 2017) to train the models and reproduce their reported results on the En→De task. The hidden size is 512, filter size is 2,048, and the number of attention heads is 8. All models are trained on four NVIDIA TITAN Xp GPUs where each is allocated with a batch size of 4,096 tokens. We adopt a fine-tuning training strategy for all SyntAligner variants, and firstly pretrain about 30 epochs for all translation tasks with the cross-entropy training objective. Then, we fine-tune the NMT models for about $1 \sim 3$ epochs by using both the SyntAligner and cross-entropy training objectives. Specifically, we take turn to optimize the networks by self-supervised or supervised training objectives. Due to the introduction of the $\mathcal{J}_{MIM}$ objective, our training speed of fine-tuning is about $1.5\times$ slower than traditional NMT methods.

The byte-pair encoding (BPE) toolkit[2] (Sennrich et al. 2016) is used with 32K merge operations. The 4-gram NIST BLEU score (Papineni et al. 2002) is used as the evaluation metric. The Berkeley Neural Parser (Kitaev et al. 2018) generates the constituency spans for English, German and French languages. Besides, the statistical significance test method in (Collins et al. 2005) is taken.

Main Results

Tab. 1 shows the main results of three baseline Transformers and the Transformer-Base enabled by our proposed SyntAligner with multiple BS-SA mechanism variants on the WMT14 En→De dataset with 4.5M pairs, IWSLT14 De→En dataset with 160K pairs, and the WMT16 En→Fr dataset with 180K pairs.

The test against SyntAligner variants in the lower table shows the effectiveness of maximizing the mutual dependency between bilingual syntactic representations with the precisely-aligned span samples. The border-word self-attention lifts the SyntAligner, which together, i.e., SyntAligner-Tgt, substantially outperforms Transformer by +1.25 BLEU points on En→De, +1.50 BLEU points on De→En, and +0.91 BLEU points on En→Fr. These results demonstrate the efficacy and applicability of SyntAligner and the alignment accuracy of the BS-SA mechanism.

The upper table shows the results of three types of baselines: Transformer, LightConv and MG-SA. While both LightConv and MG-SA make an improvement over Transformer, the SyntAligner with multiple BS-SA-enabled (SyntAligner-Base, SyntAligner-Src, SyntAligner-Tgt, SyntAligner-Bi) NMT models substantially and consistently beat the standard Transformer and both LightConv and MG-SA. For example, our SyntAligner-Tgt on Transformer outperforms MG-SA by over 0.28 BLEU points on En→De and outperforms LightConv by over 0.33 BLEU points on De→En. This is owing to the BS-SA and SyntAligner design of aligning the bilingual syntactic representation and maximizing the syntax-consistency between the word hidden states of encoder and decoder. On the other

[2]https://github.com/rsennrich/subword-nmt

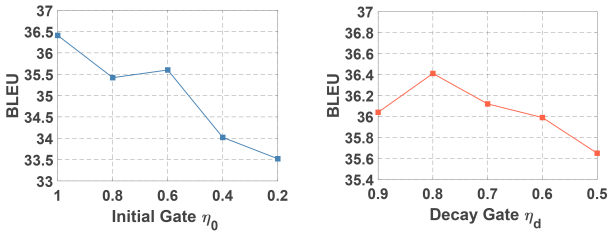

Figure 3: The Effect of Curriculum Learning Gates.

hand, the results of the SyntAligner-Tgt and SyntAligner-Bi models outperform the SyntAligner-Src model. It may be owing to that the target-side self-attention weights can bias the relative value of the span alignment weights more directly for the $softmax$ function than the former one. Meanwhile, the scale of parameters is well controlled in our models compared to all the baselines, because our models mostly transform the word hidden states into syntactic span vectors without using additional networks and thus do not incur too many training parameters.

Analysis

Here, we analyze the effectiveness and generalizability of different mechanisms in SyntAligner, the effect of BS-SA-enabled alignment, and the visualization of comparing mutual information loss function and cross-entropy loss function during training. Owing to space limitation, we only report testing results on the IWSLT14 De→En validation set.

Tuning the Curriculum Learning Gates A major challenge of this work is that no prior golden-alignment syntactic span pairs are available as the positive samples to maximize the syntactic mutual information in SyntAligner. We thus use the curriculum learning-based sampling to guarantee the alignment quality of positive samples at the cost of initial sampling scale.

Fig. 3 shows the performance of adjusting the values of initial gate and decay gate of the curriculum learning sampling strategy. The left figure fixes the decay gate ($\eta_d = 0.8$) to evaluate the effect of initial alignment quality on performance, where higher initial gate values lead to better results, proving that strictly controlling the alignment quality at initial epochs is critical for the mutual dependency. The right figure fixes the initial gate ($\eta_0 = 1.0$) to find a trade-off between the scale and quality of effectively-aligned syntactic pairs. The performance firstly increases as the decay gate decreases until $\eta_d = 0.8$ and then drops with the decrease of decay gate value. The results reveal that a too loose sampling strategy may not guarantee the effectiveness of SyntAligner training objective. For example, when $\eta_d = 0.5$, it means the curriculum gate $\eta = 0.075 \ll 1$ after 50k steps. Hence, it may introduce some noise because it almost samples the total syntactic span pairs.

Effect of Scaling Weight on BS-SA and SyntAligner Fig. 4 shows how the scaling weights W_s and W_t combining with border-word self-attention influences the alignment confidence for the De→En sentence pair: "*was macht man , wenn man solch eine unterbrechung im fluss hat ?*" , "*what do you do when you have this sort of disrupted flow ?*".

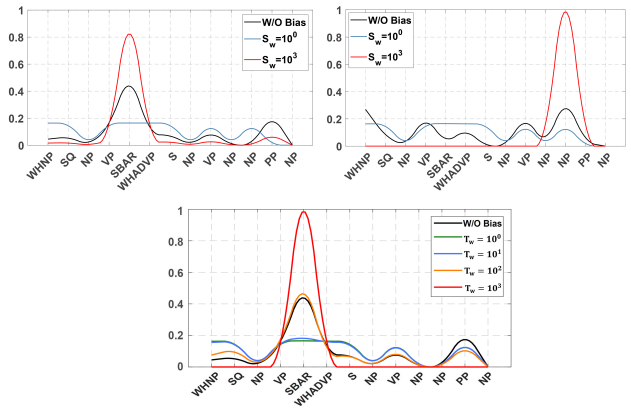

Figure 4: The Effect of Scaling Bias on BS-SA Alignment Distribution.

The top two figures in Fig. 4 show the source BS-SA alignment distributions for two source syntactic spans (*wenn*, *hat*, S-MO) and (*solch*, *eine*, AP-NK), respectively. We illustrate the alignment distributions of three models: BS-SA-Base (the black line), BS-SA-Src with $W_s = 10^0$ (the blue line) and BS-SA-Src with $W_s = 10^3$ (the red line). The scaled self-attention ($W_s \times S_{Conf}$) of border-word (*wenn*, *hat*) and (*solch*, *eine*) are 6.8 and 93.3. Both the source syntactic spans have the corresponding target syntactic spans (at the top of distribution) in top-left and top-right figures, which are Subordinate Clause [(*wenn*, *hat*, S-MO), (*when*, *hat*, SBAR)] and Noun Phrase [(*solch*, *eine*, AP-NK), (*this*, *sort*, NP)]. However, the improvement of the highest alignment attention weight (the gap between the peak values of red and black lines) in top-right figure is more significant than in top-left figure, owing to the border-words (*this*, *sort*) are neighbouring words with a tight connection, resulting in a higher confidence score $W_s \times S_{Conf} = 93.3$ for (*solch*, *eine*, AP-NK). This shows that the source confidence bias influences the alignment distribution in different rows at varying degrees. On the other hand, both top figures also show that an appropriate scaling weight W_s should be set for the source confidence bias S_{Conf}, otherwise incurring a more uniform alignment distribution like the blue line to make the sampling aligned syntactic pairs more difficult.

Further, the bottom figure in Fig. 4 compares the alignment distribution of different scaling weights W_t. Contrary to top figures in Fig. 4, the target confidence bias changes the related values between different columns according to the target border-word connection. When $W_t = 10^0$ or 10^1, the distributions are more uniform than the BS-SA-Base model. When $W_t = 10^3$, the different target confidence biases for each column widen the alignment attention weight gap between (*when*, *hat*, SBAR) (the $5th$ column) and (*of*, *flow*, PP) (the $12th$ column) from 0.31 to 1.0.

In addition, Tab. 2 shows that different scaling weights of both SyntAligner-Src and SyntAligner-Tgt lift the SyntAligner performance at varying degrees. This proves (1) ap-

Systems	W_s/W_t	BLEU	△
SyntAligner-Base	N/A	35.80	-
SyntAligner-Src	10^0	35.65	-0.15
	10^1	35.60	-0.20
	10^2	35.78	-0.02
	10^3	36.01	+0.21
SyntAligner-Tgt	10^0	35.70	-0.1
	10^1	35.85	+0.05
	10^2	35.99	+0.19
	10^3	36.41	+0.61

Table 2: The Effect of Scaling Weights W_s and W_t on SyntAligner-Src and SyntAligner-Tgt models.

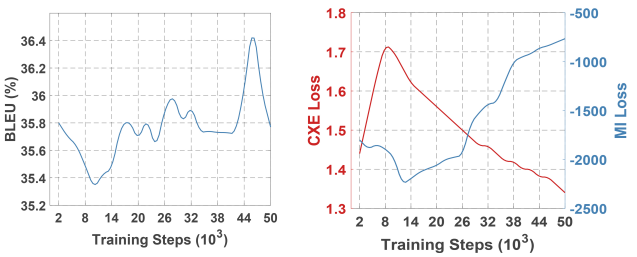

Figure 5: The translation performance on the validation set in the joint-objective-based training (the left). The Cross-Entropy Loss (the red line) vs. Mutual Information Loss (the blue line) in the joint-objective-based training (the right)

propriate scaling weights W_s and W_t can enhance the confidence of aligned bilingual syntactic span pairs and boost the performance of SyntAligner by +0.21 and +0.61 BLEU point, respectively; and (2) the influence of scaling weights plays a more important role on the SyntAligner-Tgt model by improving 0.61 BLEU point when $T_w = 10^3$. Together with the results in Fig. 4, we can conclude that the higher quality of aligned syntactic span pairs induces the better performance of SyntAligner-based NMT.

Effect of the Joint Objective-based Training The SyntAligner introduces a mutual information maximization-based self-supervised training objective to leverage the alignment of bilingual syntactic structures, combined with supervised cross-entropy minimization training objective, to jointly train it for NMT. Fig. 5 shows the performance of this joint objective-based training. Both the translation performance and the loss variation follow the same trend, which first descend at the initial several steps and then turns to increase at the late training steps. The Supplementary further illustrates more results about the similar trend across different language translations. In the right part of Fig. 5, the Cross-Entropy Minimization Loss (CE Loss) is disturbed by the newly-applied Mutual Information Maximization (MI Loss) Loss before 10k training steps. Correspondingly, the translation performance also decreases in left part of Fig. 5. It may be owing to the low alignment quality in the initial fine-tuning, which provides noisy positive samples for mutual information maximization. Meanwhile, the worse trend of CE Loss backlashes against the MI Loss, which leads to the decrease of bilingual syntactic mutual information. As training proceeds, both training objectives gradually adapt to the joint optimization. The final best translation performance achieves at 48k training steps in the left part of Fig. 5. This shows the necessity and challenge to find the balance between the self-supervised and supervised training objectives.

Related Work and Discussion

One popular extension to NMT is to improve the linguistic representation by integrating syntactic knowledge on either the source-side (Bugliarello et al. 2020; Eriguchi et al. 2016; Hao et al. 2019; Li et al. 2017; Zhang et al. 2020) or the target-side (Aharoni et al. 2017; Passban et al. 2018; Wu et al. 2017). For example, a String-to-Tree NMT model (Aharoni et al. 2017) translates into linearized, lexicalized constituency trees. However, the related work on the source-to-target integration is limited. (Wu et al. 2018) simply applies the dependence structures in both source and target languages without deeply analyzing the syntactic correspondence and divergence between source-target sentences.

Mutual information-based objective functions such as the InfoMax principle (Linsker 1988) were used in self-supervised representation learning in such domains as computer vision, audio processing, and reinforcement learning (Bachman et al. 2019; Belghazi et al. 2018; Hjelm et al. 2019; Löwe et al. 2019; van den Oord et al. 2018). Some of the related methods maximize a particular lower bound of mutual information, e.g. InfoNCE (van den Oord et al. 2018), also known as contrastive learning (Saunshi et al. 2019). While less work is reported in Natural Language Processing and NMT, (Kong et al. 2020) use the mutual information as training objective that unifies classical word embedding models (e.g., Skip-gram) and modern contextual embedding (e.g., BERT, XLNet). Our work makes a new attempt to adopt mutual information to optimize the bilingual syntactic alignment to boost the translation performance.

More work on learning hierarchical semantic, syntactic and linguistic couplings (Cao 2015; Cheng et al. 2013) within/between sources and targets.

Conclusion

While mono-lingual syntactic knowledge has been widely explored for NMT tasks, an open challenge is to characterize the dependency between source and target syntactic structures in the classic sequence-to-sequence modeling. This paper makes one step forward by not only aligning the bilingual syntactic structures but also introducing a self-supervised mutual information maximization-based training objective, combined with traditional supervised cross-entropy training objective, which enables bilingual syntax-consistency-based translation. The proposed SyntAligner effectively aligns the source and target syntactic structures across multiple languages.

Acknowledgements

We very appreciate the comments from anonymous reviewers which will help further improve our work. This work is supported by National Key R&D Plan(No.2018YFC0832104), National Natural Science Foundation of China (No.61732005).

References

Aharoni, R.; Goldberg, Y.; et al; and et al. 2017. Towards String-To-Tree Neural Machine Translation. In *ACL*, 132–140. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-2021. URL https://www.aclweb.org/anthology/P17-2021.

Ba, L. J.; Kiros, J. R.; Hinton, G. E.; and et al. 2016. Layer Normalization. *CoRR* abs/1607.06450. URL http://arxiv.org/abs/1607.06450.

Bachman, P.; Hjelm, R. D.; Buchwalter, W.; and et al. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 15509–15519. URL http://papers.nips.cc/paper/9686-learning-representations-by-maximizing-mutual-information-across-views.

Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; and Sima'an, K. 2017. Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 1957–1967. Association for Computational Linguistics. doi:10.18653/v1/d17-1209. URL https://doi.org/10.18653/v1/d17-1209.

Belghazi, I.; Rajeswar, S.; Baratin, A.; Hjelm, R. D.; and Courville, A. C. 2018. MINE: Mutual Information Neural Estimation. *CoRR* abs/1801.04062. URL http://arxiv.org/abs/1801.04062.

Bugliarello, E.; Okazaki, N.; et al; and et al. 2020. Enhancing Machine Translation with Dependency-Aware Self-Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 1618–1627. URL https://www.aclweb.org/anthology/2020.acl-main.147/.

Cao, L. 2015. Coupling learning of complex interactions. *Inf. Process. Manage.* 51(2): 167–186.

Cheng, X.; Miao, D.; Wang, C.; and Cao, L. 2013. Coupled term-term relation analysis for document clustering. In *IJCNN'2013*, 1–8.

Collins, M.; Koehn, P.; Kucerova, I.; and et al. 2005. Clause Restructuring for Statistical Machine Translation. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, 531–540. doi:10.3115/1219840.1219906. URL https://www.aclweb.org/anthology/P05-1066/.

Cross, J.; Huang, L.; et al; and et al. 2016. Span-Based Constituency Parsing with a Structure-Label System and Provably Optimal Dynamic Oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1–11. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1001. URL https://www.aclweb.org/anthology/D16-1001.

Eriguchi, A.; Hashimoto, K.; Tsuruoka, Y.; and TT. 2016. Tree-to-Sequence Attentional Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. doi:10.18653/v1/p16-1078. URL https://doi.org/10.18653/v1/p16-1078.

Gutmann, M.; Hyvärinen, A.; et al; and et al. 2012. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *J. Mach. Learn. Res.* 13: 307–361. URL http://dl.acm.org/citation.cfm?id=2188396.

Hao, J.; Wang, X.; Shi, S.; Zhang, J.; and Tu, Z. 2019. Multi-Granularity Self-Attention for 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 2019, Hong Kong, China, November 3-7, 2019*, 887–897. doi:10.18653/v1/D19-1082. URL https://doi.org/10.18653/v1/D19-1082.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. doi:10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. URL https://openreview.net/forum?id=Bklr3j0cKX.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Proceedings of the Neural Computation* 9(8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Kasami, T. 1965. An efficient recognition and syntax analysis algorithm for context-free languages. Technical Report AFCRL-65-758, Air Force Cambridge Research Laboratory, Bedford, MA†.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the EMNLP*, 1746–1751. URL http://aclweb.org/anthology/D/D14/D14-1181.pdf.

Kitaev, N.; Klein, D.; et al; and et al. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2676–2686. doi:10.18653/

v1/P18-1249. URL https://www.aclweb.org/anthology/P18-1249/.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. M. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the ACL*, 67–72. doi:10.18653/v1/P17-4012. URL https://doi.org/10.18653/v1/P17-4012.

Koehn, P.; Och, F. J.; Marcu, D.; and et al. 2003. Statistical Phrase-Based Translation. In *NAACL*. URL https://www.aclweb.org/anthology/N03-1017/.

Kong, L.; de Masson d'Autume, C.; Yu, L.; Ling, W.; Dai, Z.; and Yogatama, D. 2020. A Mutual Information Maximization Perspective of Language Representation Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL https://openreview.net/forum?id=Syx79eBKwr.

Li, J.; Xiong, D.; Tu, Z.; Zhu, M.; Zhang, M.; and Zhou, G. 2017. Modeling Source Syntax for Neural Machine Translation. In *ACL*, 688–697. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-1064. URL https://www.aclweb.org/anthology/P17-1064.

Linsker, R. 1988. Self-Organization in a Perceptual Network. *IEEE Computer* 21(3): 105–117. doi:10.1109/2.36. URL https://doi.org/10.1109/2.36.

Löwe, S.; O'Connor, P.; Veeling, B. S.; and et al. 2019. Greedy InfoMax for Biologically Plausible Self-Supervised Representation Learning. *CoRR* abs/1905.11786. URL http://arxiv.org/abs/1905.11786.

Mukherjee, A.; Ala, H.; Shrivastava, M.; and Sharma, D. M. 2020. MEE : An Automatic Metric for Evaluation Using Embeddings for Machine Translation. In *DSAA'2020*, 292–299.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL*, 311–318. URL http://www.aclweb.org/anthology/P02-1040.pdf.

Passban, P.; Liu, Q.; Way, A.; and et al. 2018. Improving Character-Based Decoding Using Target-Side Morphological Information for Neural Machine Translation. In *ACL*, 58–68. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1006. URL https://www.aclweb.org/anthology/N18-1006.

Piazza, N. 2020. Classification Between Machine Translated Text and Original Text By Part Of Speech Tagging Representation. In *DSAA'2020*, 739–740.

Saunshi, N.; Plevrakis, O.; Arora, S.; Khodak, M.; and Khandeparkar, H. 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 5628–5637. URL http://proceedings.mlr.press/v97/saunshi19a.html.

Sennrich, R.; Haddow, B.; Birch, A.; and et al. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the ACL*. URL https://www.aclweb.org/anthology/P16-1162/.

Stern, M.; Andreas, J.; Klein, D.; and et al. 2017. A Minimal Span-Based Neural Constituency Parser. 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*, 818–827. doi:10.18653/v1/P17-1076. URL https://doi.org/10.18653/v1/P17-1076.

van den Oord, A.; Li, Y.; Vinyals, O.; and et al. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748. URL http://arxiv.org/abs/1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proceedings of the NeurIPS*, 5998–6008. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y. N.; and Auli, M. 2019. Pay Less Attention with Lightweight and Dynamic Convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. URL https://openreview.net/forum?id=SkVhlh09tX.

Wu, S.; Zhang, D.; Yang, N.; Li, M.; and Zhou, M. 2017. Sequence-to-Dependency Neural Machine Translation. In *ACL*, 698–707. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-1065. URL https://www.aclweb.org/anthology/P17-1065.

Wu, S.; Zhang, D.; Zhang, Z.; Yang, N.; Li, M.; and Zhou, M. 2018. Dependency-to-Dependency Neural Machine Translation. *IEEE/ACM TASLP* 26(11): 2132–2141.

Zhang, T.; Huang, H.; Feng, C.; and Wei, X. 2020. Similarity-aware neural machine translation: reducing human translator efforts by leveraging high-potential sentences with translation memory. *Neural Comput. Appl.* 32(23): 17623–17635. doi:10.1007/s00521-020-04939-y. URL https://doi.org/10.1007/s00521-020-04939-y.