Neural Machine Translation with Synchronous Latent Phrase Structure

Shintaro Harada Taro Watanabe
Nara Institute of Science and Technology (NAIST), Nara, Japan
{harada.shintaro.hk4, taro}@is.naist.jp

Abstract

It is reported that grammatical information is useful for machine translation (MT) task. However, the annotation of grammatical information requires the highly human resources. Furthermore, it is not trivial to adapt grammatical information to MT since grammatical annotation usually adapts tokenization standards which might not be suitable to capture the relation of two languages, and the use of sub-word tokenization, e.g., Byte-Pair-Encoding, to alleviate out-of-vocabulary problem might not be compatible with those annotations. In this work, we propose two methods to explicitly incorporate grammatical information without supervising annotation; first, latent phrase structure is induced in an unsupervised fashion from a multi-head attention mechanism; second, the induced phrase structures in encoder and decoder are synchronized so that they are compatible with each other using constraints during training. We demonstrate that our approach produces better performance and explainability in two tasks, translation and alignment tasks without extra resources. Although we could not obtain the high quality phrase structure in constituency parsing when evaluated monolingually, we find that the induced phrase structures enhance the explainability of translation through the synchronization constraint.

1 Introduction

Although machine translation (MT) has achieved improved performance using neural machine translation (NMT), the translation qualities for distant languages are still poor (Johnson et al., 2017). As a way to tackle the problem, statistical MT (SMT) incorporates synchronous grammar to achieve more linguistically accurate translations, in which complex structural relations between source and target languages are expressed using phrase structure (Wong et al., 2005). The synchronous grammar expresses the complex relationships between source and target languages and incorporates phrase structure to enable more linguistically accurate translation. A similar idea could be employed for NMT to achieve improved performance on those distant language pairs. However, grammatical information annotation demands high human resources. In addition, such grammatical annotation is done on word-level granularities, which might not be the best tokenization for MT tasks due by language mismatch or out-of-vocabulary problem, and often sub-word tokenization, e.g., Byte-Pair-Encoding (BPE) (Sennrich et al., 2016), is employed to alleviate the problem. As a result, it is difficult to incorporate grammatical information into NMT that handle multiple languages simultaneously.

Recently, there have been researches on unsupervised learning of phrase structure without relying on human annotations. Although these phrase structures learned in an unsupervised fashion are very close to the human annotation (Shen et al., 2018a,c), there exists no model which incorporates phrase structures as latent information to improve the performance and explainability of translation.

In this work, we introduce an approach to incorporate the phrase structure explicitly into Transformer (Vaswani et al., 2017). The approach can split into two steps; first, latent phrase structures are induced in an unsupervised fashion for the source and target sides (Shen et al., 2018a); second, the two induced latent phrase structures are synchronously agreed with each other through an attention mechanism (Deguchi et al., 2021). Experiments on German-English and Japanese-English show that our synchronous latent structures have achieved better performance on translation and alignment tasks. We also show that the induced phrase structures and synchronous structures can enhance the explainability of translation through

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 321–330

our detailed analysis in word alignment task.

2 Related Work

2.1 NMT with Supervised Tree Structure

In the previous work, it is reported that supervised phrase structures (Eriguchi et al., 2017; Nguyen et al., 2020) and dependency structures (Ma et al., 2019; Deguchi et al., 2019) can help the performance of MT. However, these approaches require an annotated corpus of syntactic structures. In addition, such syntactic annotation is done on word-level granularities, which might not be the best tokenization for MT tasks due by language mismatch or out-of-vocabulary problem, and often BPE (Sennrich et al., 2016), is employed to alleviate the problem. However, the application of BPE to grammatical information might require a different approach for each language.

2.2 Latent Grammar Induction with Neural Machine Translation

Shen et al. (2018a) introduce the concept called "syntactic distance" which represents the syntactic relation of word pairs. Similarly, Shen et al. (2018c) introduce ordered neurons which allows to learn long-term or short-term information by a novel gating mechanism and activation function. Kim et al. (2019) apply amortized variational inference for recurrent neural network grammar to learn the phrase structures in an unsupervised fashion. Wang et al. (2019) add an extra constraint to the multi-head self-attention mechanism in order to encourage the attention heads to follow phrase structures. Shen et al. (2020) introduce the constrained multi-head self-attention mechanism that allows to induct phrase and dependency structure at the same time.

These works successfully learn to induce phrase structure from language modeling task without extra linguistic resources. It is described in (Htut et al., 2019) that translation task is a conditional language modeling task with many supervisory signals and is suitable for deriving phrase structure. Unfortunately, despite grammatical information helps the understanding model work, previous work has not explicitly used induced phrase structures.

2.3 Transformer NMT

We employ the Transformer (Vaswani et al., 2017) as our base model, which is an encoder-decoder model that relies on an attention mechanism for computing the contextual representations of source and target text. Both the encoder and decoder are composed of multiple layers, each of which includes a multi-head attention (MHA) and a feed-forward sub-layer. To compute the MHA output, three inputs, query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are projected into N different sub-spaces, namely heads, with each output computed in each subspace, then, projected back to the original space after aggregation:

$$\hat{\mathbf{Q}}^{1:N} = \mathbf{Q}\mathbf{W}_Q, \hat{\mathbf{K}}^{1:N} = \mathbf{K}\mathbf{W}_K, \hat{\mathbf{V}}^{1:N} = \mathbf{V}\mathbf{W}_V \tag{1}$$

$$\mathbf{H}^n = \mathbf{A}\hat{\mathbf{V}}^n = \mathrm{softmax}\left(\frac{\hat{\mathbf{Q}}^n \hat{\mathbf{K}^n}^\top}{\sqrt{d_h}}\right)\hat{\mathbf{V}}^n \tag{2}$$

$$\mathrm{MHA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \mathrm{concat}\left(\mathbf{H}^1, ..., \mathbf{H}^N\right)\mathbf{W}_O \tag{3}$$

where $\mathbf{W}_Q \in \mathbb{R}^{d_o \times d_h}$, $\mathbf{W}_K \in \mathbb{R}^{d_o \times d_h}$, $\mathbf{W}_V \in \mathbb{R}^{d_o \times d_h}$, $\mathbf{W}_O \in \mathbb{R}^{d_o \times d_h}$ are projection parameters. d_o is dimension of original space. $d_h = d_o/N$ is the dimension of subspace. The value $\mathbf{A}$ denotes the attention probability for the jth target token overall the ith source token, computed by nth head.

In the translation task, Transformer is frequently used for its translation accuracy and efficiency. Transformer decoder employs the autoregressive model which guesses the next token having read all the previous ones. Also, since attention represents relationship the between source and target tokens, it is used in the alignment task (Garg et al., 2019).

2.4 Synchronous syntactic attention

Deguchi et al. (2021) find that NMT performance can be improved by synchronizing the encoder attention to decoder attention, which is called "synchronous syntactic attention". The dependency information is embedded in these attention by supervised learning task. The encoder-decoder attention can be viewed as a soft word alignment, which is a weight that can project the source vector into the target vector space without additional model parameters. This work synchronize the source and target attentions that be embedded dependency information by supervision task. To match the attention of encoder and decoder, they project the encoder attention to the target one, and incorporate constraints such that the source and target attention agree with each other.

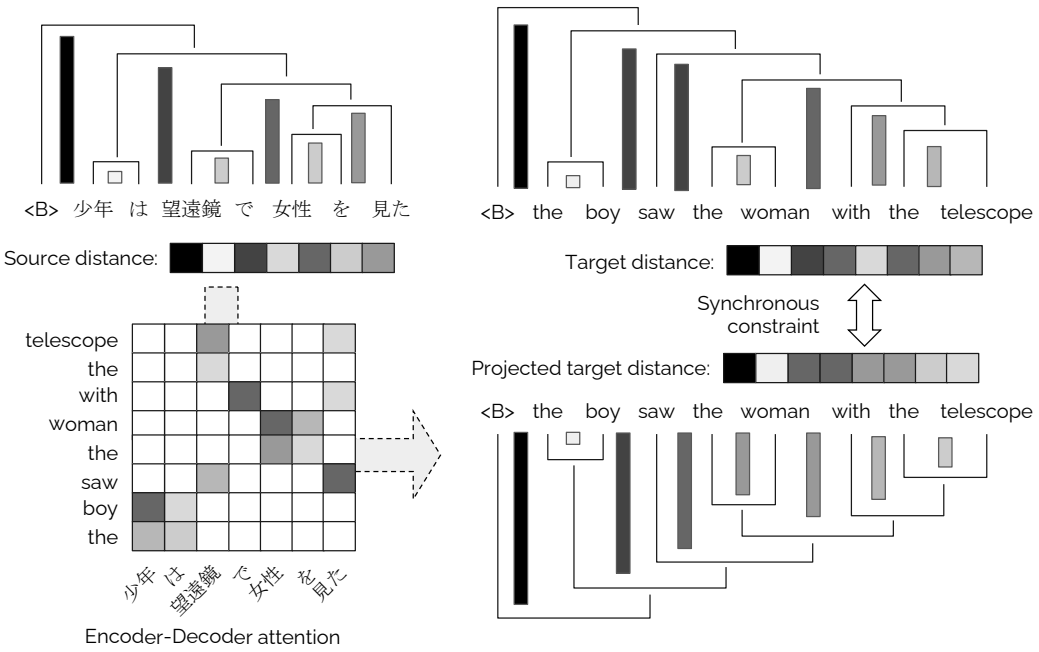


Figure 1: The example of relation between syntactic distances and synchronous constraint on Japanese to English translation task. Starting from induction of source and target syntactic distances, we project the source distance to the target one through encoder-decoder attention weight. By measuring the difference between the projected target syntactic distance and target one with the synchronous constraint. It can embed the syntactic correspondences of source and target language into the encoder-decoder attention weight.

3 Synchronous Latent Phrase Structure

In this section, we present the Synchronous Latent Phrase Structure. This proposed method is split into two steps. One is Latent Phrase Structure Induction (LPSI) and the other is Synchronous Constraint. Figure 1 shows the flow of synchronizing Japanese source and English target syntactic distances.

3.1 Latent Phrase Structure Induction

We employ syntactic distance (Shen et al., 2018a) as a way to induce phrase structure. Each syntactic distance d_i is associated with each span $(i, i+1)$ which indicates the relative order of hierarchically splitting a sentence into smaller components. For example, Figure 1 shows that the target syntactic distance between 'woman' and 'with' covers the phrase 'the woman with the telescope'. Mathematically, syntactic distance d_i is computed through the convolution-based network:

$$d_i = \tanh(\mathbf{W_D} \begin{bmatrix} \boldsymbol{k}_{i-M} \\ \boldsymbol{k}_{i-M+1} \\ \cdots \\ \boldsymbol{k}_i \end{bmatrix} + b_D) \quad (4)$$

where $\mathbf{W}_D$ and b_D are convolution kernel parameter, kernel size M represents a look-back range to calculate syntactic distance $\boldsymbol{d}$. $\boldsymbol{k}_i \in \hat{\mathbf{K}}^n$ is same as key used in MHA. The attention gate values are computed as follows:

$$g_{i,t} = P(b_t \leq i) = \prod_{j=i+1}^{t-1} \alpha_{j,t} \quad (5)$$

$$\alpha_{j,t} = \frac{\mathrm{hardtanh}((d_t - d_j) \cdot \tau) + 1}{2}$$

where t is the current time step. $\alpha_{j,t}$ is a probability value that represents the syntactic relationship of distance d_j and d_t, and $\mathrm{hardtanh}(x) = \max(-1, \max(1, x))$. τ is the temperature hyper parameter that controls the sensitivity of $\alpha_{j,t}$ to the differences between syntactic distances. b_t is a variable that indicates the position of break in the phrase structure. This α is sharper than softmax function, which allows to separate the constituents more easily. The phrase structured MHA is defined based on the gates:

$$\tilde{a}_{i,t} = \frac{g_{i,t} \cdot a_{i,t}}{\sum_i g_{i,t} \cdot a_{i,t}} \quad (6)$$

where a is an element of attention $\mathbf{A}$. The gate $g_{i,t}$ is a weight that constrains attention to only the

same hierarchy in the phrase structure. Here, $\tilde{a}$ is used in place of the elements of $\mathbf{A}$ in Equation 2.

3.2 Synchronous Constraint

In the MT model, encoder and decoder learn separate phrase structures, which are not necessarily synchronized in that two structures may not be compatible with each other in terms of vector representations. Therefore, synchronizing each phrase structure learned in encoder and decoder, inspired by synchronous grammar in SMT, may improve the performance of translation by the synchronous structure. Inspired by synchronous syntactic attention (Deguchi et al., 2021), we project the structure expressed by the encoder syntactic distance to the target one, and incorporate constraints such that the source and target syntactic distances agree with each other. In Figure 1, the source syntactic distance is projected to the target syntactic distance through the attention weight, and the syntactic correspondence between Japanese and English is learned from the target and projected syntactic distances of the phrase 'saw the woman with the telescope'.

The synchronous constraint can be represented by using the Mean Squared Error (MSE) of the syntactic distance between the source and target languages:

$$\mathcal{L}_{sync} = \sum_{l}^{L} \sum_{i} \left(d_i^{(l)} - \tilde{d}_i^{(l)} \right)^2 \tag{7}$$

$\boldsymbol{d}^{(l)}$ is projected syntactic distance in lth decoder layer and computed as:

$$\tilde{\boldsymbol{d}}^{(l)} = \mathbf{C}^{(l)} \boldsymbol{e}^{(l)} \tag{8}$$

where $\boldsymbol{e}^{(l)}$ is syntactic distance in lth encoder layer. $C^{(l)} \in \mathbf{R}^{J \times I}$ is the lth encoder-decoder attention weight, which represents the relationships of encoder and decoder representations, works just like MHA. Here, I and J are length of source and target sentence. The lth encoder-decoder attention weight is computed as:

$$\mathbf{C}^{(l)} = \mathrm{softmax}(\frac{\hat{\mathbf{Q}}_{dec}^{(l)} \hat{\mathbf{K}}_{enc}^{(l)\top}}{\sqrt{\delta_h}}) \tag{9}$$

where $\mathbf{Q}_{dec}^{(l)}$ and $\mathbf{K}_{enc}^{(l)}$ are lth decoder and encoder hidden weights.

The important element in phrase structure is the hierarchical positional relationship derived from syntactic distance. However, MSE over-penalizes the models, because it results in the exact distance prediction task. Therefore, we use the rank loss (Burges et al., 2005) as proposed by Shen et al. (2018b), which takes hierarchical positioning into account. Applying the rank loss to the synchronous constrict, we obtain the following:

$$\mathcal{L}_{sync} = \sum_{l}^{L} \sum_{i,j>i} \mathrm{hinge}\left(d_i^{(l)} - d_j^{(l)}, \tilde{d}_i^{(l)} - \tilde{d}_j^{(l)} \right) \tag{10}$$

where $\mathrm{hinge}(x_1, x_2) = \max\left(0, 1 - \mathrm{sign}(x_1) \cdot x_2\right)$ and $\mathrm{sign}(x)$ is sign function. Therefore, the overall objective $\mathcal{L}$ is represented by:

$$\mathcal{L} = \mathcal{L}_{trans} + \lambda \mathcal{L}_{sync} \tag{11}$$

where $\mathcal{L}_{trans} = -\sum_i^J \log p(y_i | \boldsymbol{x}, \boldsymbol{y}_{<i})$ where $\mathcal{L}_{trans}$ is the objective of machine translation task and $\lambda \geq 0$ is hyper parameter to control the degree of the synchronous constraint $\mathcal{L}_{sync}$. $\boldsymbol{x}$ and $\boldsymbol{y}$ are source and target sentences, respectively.

4 Experiments

We train our proposed models using the training objective in Equation 11 and evaluate them on three tasks: translation, constituency parsing, and word alignment. We implement models within the Fairseq sequence modeling toolkit (Ott et al., 2019).

4.1 Training Details

We employ the `transformer_iwslt_de_en_align` fairseq configuration for German-English dataset and the `transformer_align` fairseq configuration for Japanese-English dataset. We use two MHA layers from the bottom to induct the phrase structures, and two encoder-decoder MHA layers from the top to synchronize the encoder and decoder syntactic distances [1]. The hyper parameters are set as look back range $M = 5$ and temperature $\tau = 1.0$ [1]. The synchronous constrain hyper parameter is set by $\lambda = 0.01$ for MSE and Rank loss.

4.2 Tasks

4.2.1 Translation Task

We evaluated the effectiveness of the synchronous latent phrase structures for MT tasks on IWSLT'14 German-English and ASPEC Japanese-English

[1] We tried various settings in our preliminary experiments, and this setting achieved the best performance.

	Train	Valid	Test
IWSLT'14	160,239	7,283	6,750
ASPEC	1,255,372	1,790	1,812
Europarl v7	1,905,695	997	508

Table 1: Number of sentences in each dataset.

datasets. We train the translation models on the IWSLT'14 German-English and ASPEC Japanese-English (Nakazawa et al., 2016) datasets. We use the `prepare_iwslt14.sh` for IWSLT'14 German-English and follow the instruction of constructing the baseline system of WAT [2], but KyTea (Neubig et al., 2011) is used as the tokenizer for Japanese sentences. These datasets are applied BPE. Table 1 shows the detailed data statistics. To compare the effectiveness of synchronous latent phrase structure, we run additional baselines without latent phrase induction but with synchronous constraints applied to the attention weights. We run inference with a beam size of 5 and report the quality of translation of our models with BLEU (Papineni et al., 2002).

4.2.2 Constituency Parsing Task

In this experiment, we did not apply BPE and English data was parsed using `Stanford CoreNLP version 4.1.0` [3], and thus the number of tokens in each sentence is preserved.

The latent phrase structure is obtained by force decoding; we feed the gold target sentences from the test set into the word-wise trained MT models. We report unlabeled F-measure (UF) as the quality of English latent phrase structures, inducted from the bottom syntactic distances, with scoring script `Evalb` [4]. Here, UF is an F-measure that ignores constituency tags and evaluates only by bracketing.

4.2.3 Alignment Task

We also measure the impact of the alignment qualities represented by our synchronous grammar against other models including a statistical model FAST-ALIGN (Dyer et al., 2013) [5]. We use the same experimental setup as described in (Chen et al., 2020) and use the scripts [6] for pre-processing and evaluation. The scripts provide three different

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/baseline/dataPreparationJE.html

[3] https://stanfordnlp.github.io/CoreNLP/

[4] https://nlp.cs.nyu.edu/evalb/

[5] https://github.com/clab/fast_align

[6] https://github.com/lilt/alignment-scripts

	BLEU[%]	
	De→En	Ja→En
Transformer	34.42	29.48
w/. Synchronous Attn	34.54	29.56
Transformer + LPSI	34.83	29.44
w/. SynchMSE	34.79†	**29.79**†
w/. SynchRank	**35.05**†	29.62†

Table 2: Results on translation task in IWSLT'14 German to English (De→En) and ASPEC Japanese to English (Ja→En). Translation quality is reported in BLEU and its values in bold indicate the best performance. The numbers with † are significantly different from the Transformer baseline measured by approximate randomization test ($\alpha = 1\%$).

datasets, but we only use German-English Europarl v7 training data and the gold alignments [7] provided by (Vilar et al., 2006). Table 1 shows the detailed data statistics. We report the alignment quality in the penultimate layer following (Garg et al., 2019) with Alignment Error Rate (AER) introduced in (Vilar et al., 2006). In this task, the trained model is BPE-wise, but the reported AER is word-wise. Furthermore, we report the quality of symmetrized alignments that combined both unidirectional alignments. The combination method is employed the `grow-diagonal` heuristic (Koehn et al., 2005), in which alignments are greedily enlarged from the intersected alignments.

4.3 Results

4.3.1 Translation Task

Table 2 compares the performance of our methods against baselines. The NMT models with synchronous latent phrase structures have better translation performance. In IWSLT'14 German-English dataset, the NMT model with synchronous latent phrase structure by rank loss improves 0.63 BLEU point. In ASPEC Japanese-English dataset, the NMT model with synchronous latent phrase structure by MSE loss improves 0.31 BLEU point. These results show that the use of explicit latent phrase structures can be useful in MT tasks involving syntactically distant languages like Japanese-English.

However, the Rank synchronous constraint model performed worse than the MSE synchronous constraint model in the ASPEC Japanese-English dataset. This probably is because that the phrase

[7] https://www-i6.informatik.rwth-aachen.de/goldAlignment/

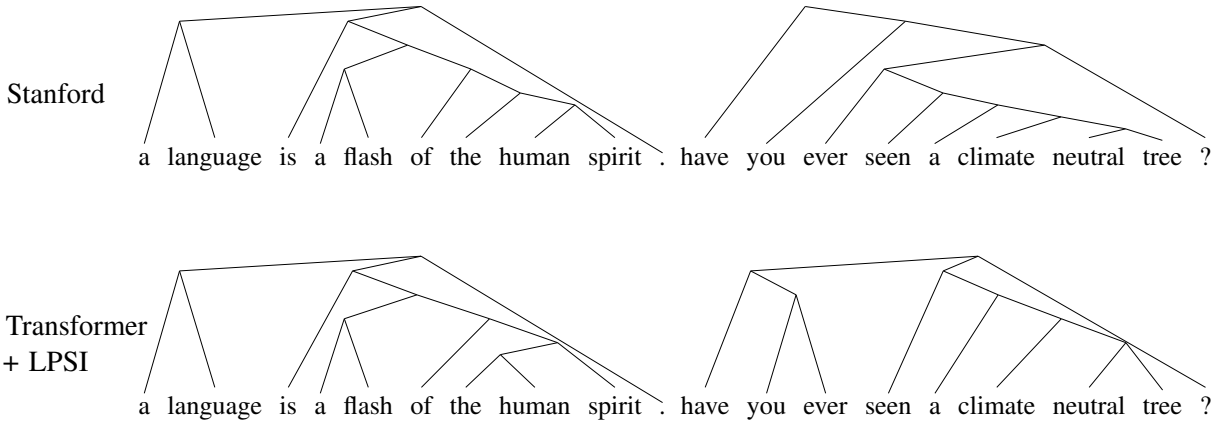


Figure 2: The top parse trees are obtained from the Stanford parser. The bottom parse trees are inducted from our transformer with LPSI (first layer) trained on IWSLT'14 German to English.

	UF[%]	BLEU[%]
	De→En	
(Hent et. al., 2019)	**56.1**	30.2
Transformer + LPSI	37.40	30.69
w/. SynchMSE	14.33	30.41
w/. SynchRank	33.75	**30.80**

Table 3: Results on constituency parsing task in IWSLT'14 German to English (De→En). Latent phrase structure quality is reported in UF and its values in bold indicate the best performance.

structure is not well inducted from the Japanese-English dataset and the advantage of Rank synchronous constraint is not utilized. The difficulty of induction phrase structure in the Japanese-English dataset can also be read from the results of Transformer with LPSI.

The synchronous syntactic attention model (Deguchi et al., 2021) also have good translation performance, but we can improve it further by incorporating the syntactic distance into the attention.

Although not shown in previous work (Htut et al., 2019), Table 2 shows that the use of explicit latent phrase structure is useful for the MT task. Interestingly, we found that the effective synchronous constrain differed between syntactically close, i.e., German-English, and distant languages, i.e., Japanese-English.

4.3.2 Constituency Parsing Task

Table 3 compares the performance of our methods against baselines. The results show that the synchronous constraint hurt the quality of latent phrase structures. Especially, in MSE synchronous constraint, UF is drooped 17.01 points from the result of Transformer with latent phrase structure induction. This is because the MSE synchronous constraints induct a synchronous grammar that is different from the phrase structure being evaluated. In other words, synchronous constrain hinders the derivation of the latent phrase structures. However, the decrease in UF by synchronous constrain by rank loss is small, whereas synchronous constrain by MSE greatly reduced UF. It suggests that synchronous constrain by MSE derives an exact synchronization grammar and synchronous constrain by rank loss derives a minimal synchronization grammar.

As with prior study (Htut et al., 2019), we did not find any correlation between the phrase structure qualities and translation qualities especially when two structures are synchronized in encoder and decoder. This indicates that our induced grammatical structures using synchronous constraints might capture bilingual correspondence better than non-constrained models.

Figure 2 shows examples of parse tree from Stanford Parser and our Transformer with LSPI. In the first example "a flash of the human spirit", our model almost correctly inducts phrase structure in comparison with Stanford Parser. The only mistake is grouping "the" and "human" first in the noun phrase "the human spirit". This mistake can be unique to concepts of syntactic distance, as it is the same as in the prior study (Htut et al., 2019). In the second example "have you ever seen a climate neutral tree ?", our model correctly inducts the verb phrase "ever seen a climate neutral tree", but fails to induct the phrase "have you ever" correctly.

	AER[%] (precision[%], recall[%])			BLEU[%]	
	De→En	**En→De**	**Symmetrized**	**De→En**	**En→De**
FAST-ALIGN	30.8 (68.2, 70.3)	32.4 (66.8, 68.4)	27.7 (81.4, 65.0)	-	-
Transformer	46.2 (51.0, 57.1)	47.5 (49.5, 56.1)	35.8 (84.9, 51.3)	33.62	26.59
Transformer + LPSI	43.4 (53.5, 60.1)	**45.9 (51.1, 57.6)**	34.3 (84.6, 53.4)	33.25	26.98
w/. SynchMSE	**42.4 (54.5, 61.2)**	46.3 (50.8, 57.2)	**34.1** (84.5, **53.8**)	33.96	26.61
w/. SynchRank	44.4 (52.7, 58.9)	50.1 (47.3, 52.9)	36.1 (**86.5**, 50.5)	**34.13**	**27.03**

Table 4: Results on the alignment and translation task in Europarl v7 German to English (De→En) and English to German (En→De). 'Symmetrized' indicates the alignments combined both unidirectional alignments De→En and En→De. Alignment quality is reported in AER, translation quality in BLEU and its values in bold indicate best performance.

	AER[%] (Precision[%], Recall[%])		
Layer	**Transformer**	**w/. SynchMSE**	**w/. SynchRank**
1	92.2 (62.7, 4.1)	95.0 (25.6, 2.7)	93.2 (26.0, 3.9)
2	92.1 (28.3, 4.5)	91.1 (34.9, 5.0)	90.8 (28.9, 5.4)
3	84.0 (42.7, 9.8)	88.6 (34.1, 6.8)	81.5 (37.2, 12.2)
4	49.3 (79.7, 37.0)	53.2 (75.2, 33.8)	40.7 (81.4, 46.4)
5	35.8 (84.9, 51.3)	**34.1** (84.5, **53.8**)	36.1 (**86.5**, 50.5)
6	47.2 (86.9, 37.7)	52.1 (87.7, 32.8)	56.9 (87.5, 28.4)

Table 5: Results of AER on each layer. The value in bold indicates the best performance.

	De→En	**Ja→En**
Transformer	34.42	29.48
w/o. Positional Embedding	17.01	15.40
Transformer + LPSI	34.83	29.44
w/o. Positional Embedding	33.94	28.89
Transformer + Local Attn	34.77	30.19
w/o. Positional Embedding	33.84	29.58

Table 6: Results on IWSLT'14 German to English (De→En) and ASPEC Japanese to English (Ja→En) for effectiveness of learning word order. 'w/o. Positional Embedding' indicates removing positional embedding from the models. The local attention mask is applied only to the encoder following a prior study (Cui et al., 2019).

4.3.3 Alignment Task

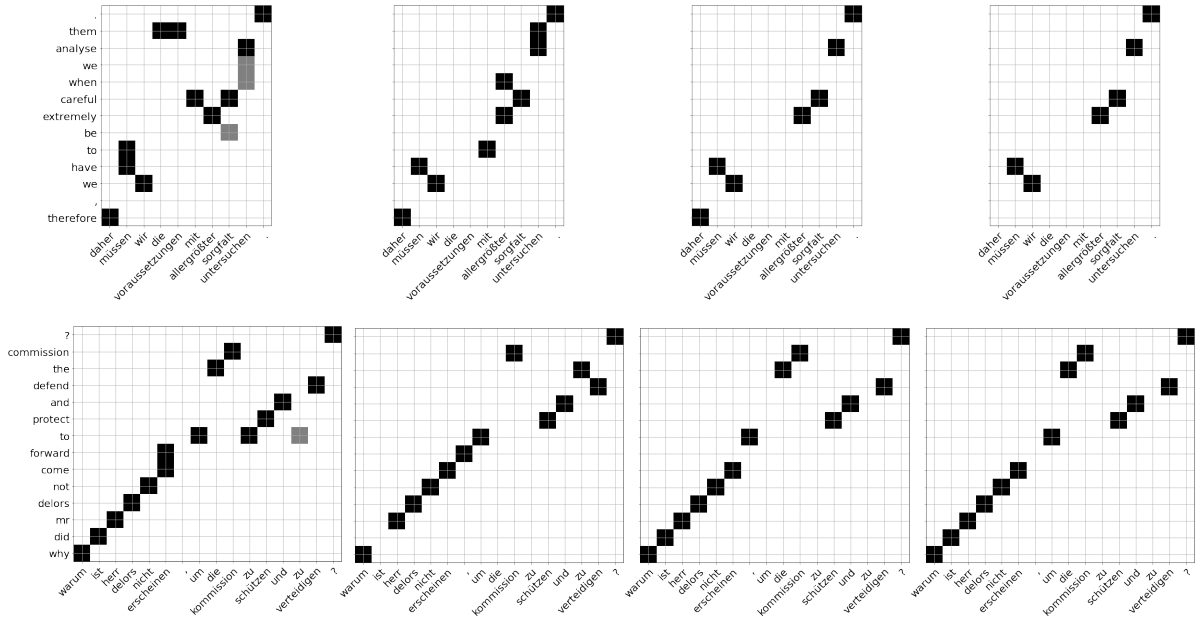
Table 4 compares the performance of our methods against statistic and neural baseline approaches. Compared with Transformer, the model with latent phrase structure show better translation performance and quality of alignments. Furthermore, synchronizing source and target latent phrase structure decreases the AER, which indicates that synchronous constrain improves the interpretability of translation. However, synchronous constrain by Rank loss resulted in a deterioration in AER, despite improving the translation performance BLEU. Therefore, the relationship between BLEU and AER does not seem to be significantly correlated.

Table 5 shows that the effectiveness of synchronous latent phrase structure for two layers from the top in terms of AER. In the penultimate layer, while synchronous constrain by MSE contributed to the improvement of AER, but synchronous constrain by rank loss conversely worsened AER. However, rank loss resulted in a significant improvement AER in the third and fourth layers. In the final layer, both synchronous constraints by MSE and rank loss result in the worse AER. It suggests that the quality of the latent phrase structure derived from the second layer from the bottom is poor and this may have affected the results adversely.

5 Analysis

5.1 Effectiveness of Attention Gate

We realize that our gated multi-head attention (GMHA), without synchronous constraint, is very similar to local attention within mixed multi-head attention (MMHA) (Cui et al., 2019). MMHA encourages each head to acquire different features by masking them differently and allows the model to be aware of the order of the sequence. Table 6 show that Transformer without position embedding decrease of 17.41 BLEU point in IWSLT'14



(a) Reference (b) FastAlign (c) SyncMSE (d) SyncRank

Figure 3: The symmetrized examples from the German-English alignment test set. Gold Alignment is shown in (a). Alignment in (b) show the output from FastAlign (BPE-wise trained), (c) from synchronized MSE model, and (d) from synchronized Rank model. Black squares and gray squares in the reference represent sure and possible alignments, respectively.

German-English and 14.08 BLEU point in ASPEC Japanese-English. In the Transformer with latent phrase structure induction (LPSI), the performance is only reduced by 0.89 BLEU point in IWSLT'14 German-English and 0.55 BLEU point in ASPEC Japanese-English without position embedding. For a fair comparison, we employ local attention with 2 window in the two bottom layers of encoder. Similarly, in the Transformer with local attention, the performance is only reduced by 0.93 BLEU point in IWSLT'14 German-English and 0.61 BLEU point in ASPEC Japanese-English without position embedding. It indicates that local constraints on attention mechanisms help learning the order of the sequence rather than latent phrase structure induction.

5.2 Effectiveness of Synchronous Latent Phase Structure

Figure 3 shows examples from the German-English alignment test set. In the first example, we find that there are no false alignments in our models with synchronous constraints. However, in rank loss, the alignment between 'Therefore' and 'Daher', which was captured by MSE, is lost. In the second example, duplicated our model correctly aligns them with 'um' compared with FastAlign. Therefore, The synchronous constraints by MSE and rank loss indicate that only alignments with high confidence are provided. Furthermore, as can be seen from the precision values in this Table 4, there are no false alignments in synchronous constrain by rank loss, and definite explainability of translation is achieved. In other words, the synchronization constraint favors precision over recall, which may make the AER worse, but it can provide a reliable explanation for human. The prior study (Jain and Wallace, 2019; Serrano and Smith, 2019) conclude that the attentions have not explainability. However, our attention is constrained by the syntactic distance, it can explain the relation between source and target sentence following the constituency tree. We will work it as the future works.

6 Conclusion

This paper introduces the approach to improve the performance and explainability of MT. In the MT task, our model improves the quality of translation even through distant language pairs. In the alignment task, we demonstrate that synchronous

constraint for syntactic distance can produce high precisional alignments to interpret MT hypothesis. Currently, our approach induces the poor latent phrase structure constructed with the previous work. To achieve the more high performance and explainability of MT, we would like to investigate other syntactic structures and a translation model which can induce better latent phrase structure.

References

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Hongyi Cui, Shohei Iida, Po-Hsuan Hung, Takehito Utsuro, and Masaaki Nagata. 2019. Mixed multi-head self-attention for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 206–214, Hong Kong. Association for Computational Linguistics.

Hiroyuki Deguchi, Akihiro Tamura, and Takashi Ninomiya. 2019. Dependency-based self-attention for transformer NMT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 239–246, Varna, Bulgaria. INCOMA Ltd.

Hiroyuki Deguchi, Akihiro Tamura, and Takashi Ninomiya. 2021. Synchronous syntactic attention for transformer nmt. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Phu Mon Htut, Kyunghyun Cho, and Samuel R Bowman. 2019. Inducing constituency trees through neural machine translation. *arXiv preprint arXiv:1909.10056*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and G'abor Melis. 2019. Unsupervised recurrent neural network grammars. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1105–1117, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Improving neural machine translation with neural syntactic distance. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2032–2037, Minneapolis, Minnesota. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Steven CH Hoi, and Richard Socher. 2020. Tree-structured attention with hierarchical accumulation. *arXiv preprint arXiv:2002.08046*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. 2018a. Neural language modeling by jointly learning syntax and lexicon. In *International Conference on Learning Representations*.

Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, and Yoshua Bengio. 2018b. Straight to the tree: Constituency parsing with neural syntactic distance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1180, Melbourne, Australia. Association for Computational Linguistics.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018c. Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv preprint arXiv:1810.09536*.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron Courville. 2020. Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

David Vilar, Maja Popović, and Hermann Ney. 2006. Aer: Do we need to "improve" our alignments? In *International Workshop on Spoken Language Translation (IWSLT) 2006*.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China. Association for Computational Linguistics.

Fai Wong, Dong-Cheng Hu, Yu-Hang Mao, Ming-Chui Dong, and Yi-Ping Li. 2005. Machine translation based on constraint-based synchronous grammar. In *International Conference on Natural Language Processing*, pages 612–623. Springer.