

# Leveraging Grammar Induction for Language Understanding and Generation

Jushi Kai Shengyuan Hou Yusheng Huang Zhouhan Lin\*

Shanghai Jiao Tong University

json.kai@sjtu.edu.cn lin.zhouhan@gmail.com

## Abstract

Grammar induction has made significant progress in recent years. However, it is not clear how the application of induced grammar could enhance practical performance in downstream tasks. In this work, we introduce an unsupervised grammar induction method for language understanding and generation. We construct a grammar parser to induce constituency structures and dependency relations, which is simultaneously trained on downstream tasks without additional syntax annotations. The induced grammar features are subsequently incorporated into Transformer as a syntactic mask to guide self-attention. We evaluate and apply our method to multiple machine translation tasks and natural language understanding tasks. Our method demonstrates superior performance compared to the original Transformer and other models enhanced with external parsers. Experimental results indicate that our method is effective in both from-scratch and pre-trained scenarios. Additionally, our research highlights the contribution of explicitly modeling the grammatical structure of texts to neural network models.<sup>1</sup>

## 1 Introduction

Neural network models, like Transformer (Vaswani et al., 2017), RoBERTa (Liu et al., 2019), and GPTs (Brown et al., 2020), have gained widespread adoption in various natural language processing tasks. These models can generate desired answers on different tasks and show strong language understanding ability on multiple datasets. However, they give up explicit parsing of the specific syntactic structure of the text data and cannot effectively establish structured and interpretable language understanding models (Ramakrishnan et al., 2018; Guo et al., 2019; Hewitt and Manning, 2019; Dai

et al., 2021). This limitation has emerged as one of the bottlenecks for neural network models to understand natural language deeply.

To this end, researchers attempt to take advantage of traditional syntax parsers (Brill et al., 1990; Magerman and Marcus, 1991) to identify grammatical components within textual data and utilize them in subsequent processing steps. These parsers are built upon established linguistic frameworks and regulations. There has been growing interest in investigating the impact of syntax on neural network models and improving them through the lens of grammar induction (Yang et al., 2020; Bai et al., 2021; Li et al., 2021; Hou et al., 2022).

While these techniques have exhibited enhanced efficacy, they still grapple with two primary challenges. Firstly, these approaches (Yang et al., 2020; Bai et al., 2021; Hou et al., 2022) usually depend on external parsers to obtain additional specialized annotations, which are expensive and not time-efficient. Their performance is greatly influenced by the choice of external parsers and may not be universally applicable across all languages. The other problem is that these approaches only focus on from-scratch scenarios (Ma et al., 2019; Yang et al., 2020) or pre-trained scenarios (Li et al., 2021; Tziafas et al., 2021; Chen et al., 2022)). They cannot share consistent improvements. This discrepancy between the two model types engenders a predicament where these methodologies are tailored exclusively to one of the scenarios.

In this paper, we introduce a novel method to induce grammar information for language understanding and generation, obviating the necessity for supplementary syntactic annotations. *The self-induced grammar features are integrated into the Transformer and simultaneously learned during the training of downstream tasks.* Additionally, we devise BPE (Byte Pair Encoding) embeddings and trade-off loss functions to facilitate grammar induction. Experiments demonstrate the compati-

\* Corresponding author

<sup>1</sup>We release our code at <https://github.com/0-KaiKai-0/Grammar-Induction>.

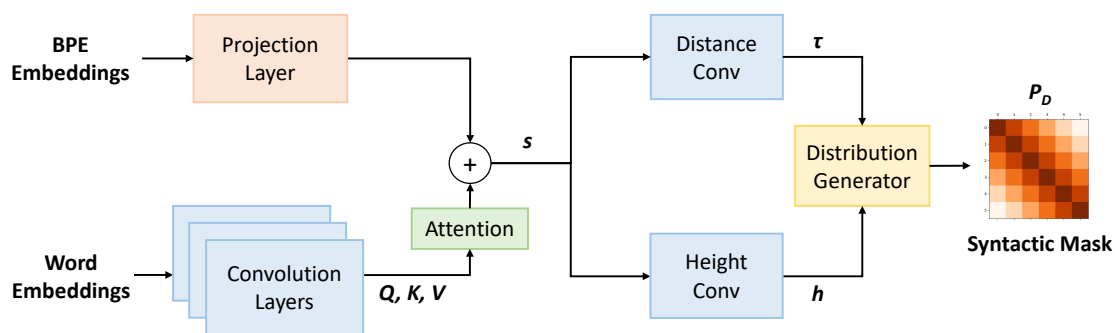


Figure 1: The pipeline for the construction of our syntactic mask. Word embeddings and BPE embeddings are utilized to induce the intermediate grammar features  $s$ , which are subsequently used to derive the syntactic distance  $\tau$  and height  $h$ . The two vectors are leveraged to estimate the dependency distribution for the sentence, and generate the syntactic mask  $P_D$ . The mask is then employed to guide the self-attention mechanism within the encoders. These parsing modules are integrated into the Transformer model and trained together in downstream tasks.

bility of our method in both the from-scratch and pre-trained scenarios. By strengthening the induction ability of the deep neural network model to the grammatical information, we can improve the model’s understanding of natural language. Our method surpasses other external-parser-enhanced methods in machine translation and language understanding tasks, showcasing persistent efficacy and applicability.

## 2 Preliminary

We use syntactic distance (Shen et al., 2018a) and height (Luo et al., 2019) to generate the dependency distribution among tokens. They are two feature vectors used to describe the constituency structure and dependency relations of a sequence of words.

### 2.1 Syntactic Distance and Height

**Syntactic distance** Syntactic distance is first proposed by Shen et al. (2018a) to model the syntactical proximity of adjacent constituents in a sentence. For a sentence  $(w_1, \dots, w_n)$ , syntactic distance  $\tau_i$  quantifies the height of the lowest common ancestor for two consecutive words  $w_i$  and  $w_{i+1}$ . The relative order of syntactic distances represents syntactic affinities between words. A smaller distance  $\tau_i$  signifies that it is easier to communicate between  $w_{1\dots i}$  and  $w_{i+1\dots n}$ .

**Syntactic height** Syntactic height is introduced by Luo et al. (2019) to depict the syntactic status of words. In the dependency graph of a sentence, the syntactic height  $h_i$  aims to capture the distance of the word  $w_i$  to the root. The word with a higher

syntactic height is more likely to stand closer to the root of the graph, containing more representative global information for the sentence.

### 2.2 Dependency Distribution

StructFormer (Shen et al., 2021) proposes a two-stage method to calculate the estimation of the dependency distribution among tokens. They use syntactic distance and height to identify the smallest legal constituent for each token and the parent of the constituent. The procedure goes through the sentence to estimate the syntactic dependency of each token on the others. We leverage their estimation method to generate the **dependency distribution** for a sentence. The estimation procedure can be referred to in Appendix A.

We introduce unsupervised grammar learning as inductive bias into Transformer for NLP downstream tasks. Our method achieves consistent improvement in both from-scratch and pre-trained scenarios.

## 3 Method

Alongside the data pipeline in the original Transformer, we construct a parser to derive grammar features and estimate a syntactic mask to guide the attention mechanism. The "parsing"<sup>2</sup> pipeline is delineated in Figure 1.

Grammar features  $s$  are induced with convolution layers and a self-attention module working on

<sup>2</sup>Despite no ground-truth parses involved in our method, the pipeline can be regarded as parsing to estimate dependency relations among tokens.

word embeddings. Specifically, convolution layers can capture localized details within segments, while the self-attention module can furnish global information of the entire sequence. BPE embeddings are introduced so that grammar status can be shared among congenetic subwords. Utilizing grammar features  $s$ , syntactic distance  $\tau$ , and height  $h$  can be derived through convolution layers. They are leveraged by the distribution generator to construct a syntactic mask  $P_D$  that provides the dependency distribution among tokens in the encoder layers of Transformer.

### 3.1 Syntactic Mask

Following previous works (Shen et al., 2018a; Luo et al., 2019; Shen et al., 2021), we quantify syntactic distance and height on the same scale. We use convolution layers and a self-attention module to induce grammar features  $s_i$ :

$$s_i = \text{Attn}(\text{Conv}(x_{i-l}, \dots, x_{i+l})) + \text{Proj}(x'_{i-l}, \dots, x'_{i+l}) \quad (1)$$

where  $x_i$  is the embedding of the  $i$ -th token, and  $2l + 1$  is the kernel size of the convolution module.  $x'_i$  is the BPE embedding, which will be introduced in Section 3.3.  $\text{Proj}(\cdot)$  is the projection layer composed of a Linear layer and a LayerNorm layer.

Then we use linear matrices  $W_1^\tau, W_1^h$  and convolution modules  $W_2^\tau, W_2^h$  to derive syntactic distance  $\tau_i$  and height  $h_i$ :

$$\tau_i = \mathbf{W}_1^\tau \tanh \left( \mathbf{W}_2^\tau \begin{bmatrix} s_i \\ s_{i+1} \end{bmatrix} \right) + b_1^\tau \quad (2)$$

$$h_i = \mathbf{W}_1^h \tanh \left( \mathbf{W}_2^h s_i + b_2^h \right) + b_1^h \quad (3)$$

Syntactic distance measures the syntactic proximity of adjacent words, while syntactic height measures the syntactic status of the word itself. Therefore, the kernel size of  $W_2^\tau$  and  $W_2^h$  is 2 and 1 respectively.

As introduced in Section 2.2, syntactic distance and height are put into the distribution generator to calculate the estimation of the dependency distribution. It serves as a mask to adjust the weights in the self-attention:

$$P_D = \mathcal{F}_P(\tau, h) \quad (4)$$

where the distribution generator is formalized as the estimation function  $\mathcal{F}_P(\cdot)$ .  $P_D$  is the derived syntactic mask containing the dependency distributions.

### 3.2 Syntax-guided Attention

The syntactic mask  $P_D$  provides the probability of information transfer among tokens. It is used to guide self-attention heads in the encoder layers of Transformer.

$$\text{Attention}(P_D, Q, K, V) = P_D \cdot \text{S} \left( \frac{QK^T}{\sqrt{d}} \right) \cdot V \quad (5)$$

where  $Q, K, V$  are query, key and value matrices and  $d$  is the hidden dimension.  $\text{S}(\cdot)$  is an activation function.

It should be noted that two different activation functions are employed in the scenarios of building from scratch and utilizing pre-trained models. In the from-scratch scenario,  $\text{S}(\cdot)$  is a sigmoid function instead of the original softmax function so that  $\text{S}(\frac{QK^T}{\sqrt{d_k}})$  can indicate an independent probability of each token's attendance on each other.

In the pre-trained setting, we adhere to the use of the softmax function. Given our utilization of the official pre-trained models RoBERTa (Liu et al., 2019) and the integration of our parser during fine-tuning, maintaining consistency in the activation function with the pre-training phase is more suitable and practical. In addition, the distribution weight is set to be  $P_D + 1$ , where the syntactic mask plays an auxiliary role in reweighting attention. This is because RoBERTa has been pre-trained devoid of grammar induction. Directly imposing new constraints and altering the model training paradigm may not align seamlessly with its existing framework.

### 3.3 BPE Embedding

When parsing, we compute dependency distributions for the tokens within sequences. Typically, the model receives subwords rather than complete words as input, owing to the data being preprocessed using BPE subword tokenization. Nonetheless, all subwords stemming from the same word should share an equivalent grammatical status within their respective sentences.

To make our parser aware of the grammar sharing among the congenetic subwords, we design BPE embeddings to represent the condition of word segmentation in the sequence. Tokens that remain intact are labeled as  $\mathbf{0}$ , while subwords resulting from word segmentation are assigned  $\mathbf{2}$  (with  $\mathbf{1}$  reserved for padding). For instance, in the sentence "How could pay arrangements be redesigned to address these problems?", the word "redesigned"

is segmented into "re" and "designed" during pre-processing. The two subwords should share grammatical information among the sentence. Consequently, we will assign them label **2** and the others label **0**.

The vector of BPE labels will be input into an embedding layer and a projection layer to obtain BPE embeddings. The BPE embeddings will be added to hidden states before they are transferred into the convolution modules to compute syntactic distances and heights. Concatenating is also considered to combine the two embeddings, but it proves ineffective. We do not introduce BPE embeddings into RoBERTa because of consistency and adaptability from the pre-training stage to fine-tuning.

### 3.4 Loss Function

We also investigate the device of loss function to facilitate grammar induction for the two scenarios studied in our paper.

Voita et al. (2019) ascertain that the flow of information through Transformer layers is contingent upon the choice of the learning objective. For machine translation (MT), the representation of the input sequence will be refined in the model and transferred from the source language to the target language. In contrast, for masked language modeling (MLM), the information about the context will be rebuilt during the encoding process.

In order to make maximum leverage of grammar induction, we trade off between the two loss functions of MLM and MT with a weighted parameter  $\lambda$ :

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{MLM} + (1 - \lambda) \cdot \mathcal{L}_{MT} \quad (6)$$

Where the  $\lambda$  will be searched for different translation tasks.

Our MLM loss is deployed using the output of the encoder module. We randomly mask part of the input with a special token and restore the masked tokens from the output. The mask rate is set to be 0.15.

In the fine-tuning phase of RoBERTa, we opt not to include the MLM loss. This decision stems from the fact that RoBERTa has already been pre-trained on a vast corpus of data, with MLM serving as its core learning objective. Consequently, the incorporation of MLM loss during fine-tuning does not yield additional benefits. More analyses about the effect of MLM loss are provided in Appendix C.

## 4 Machine Translation

We use Transformer (Vaswani et al., 2017) as our baseline and conduct experiments on six machine translation tasks of three datasets: IWSLT14-De/En, NC11-De/En, and ASPEC-Zh/Ja. Models are trained from scratch and evaluated on each task.

### 4.1 Datasets

**IWSLT14-De/En** The IWSLT14 (International Workshop on Spoken Language Translation) includes the MT track on TED Talks. We use the German (De) and English (En) corpus. We follow the standard pre-processing steps in fairseq (Ott et al., 2019). The sizes of training, validation and test sets are 160k, 7.3k and 6.8k.

**NC11-De/En** The NC11 dataset come from news commentary. We pre-process the dataset following the steps of Bugliarello and Okazaki (2020). The sizes of training, validation and test sets are 234k, 2.2k and 3.0k.

**ASPEC-Zh/Ja** The ASPEC (Asian Scientific Paper Excerpt Corpus; Nakazawa et al., 2016) dataset is a Chinese (Zh) - Japanese (Ja) scientific paper excerpt corpus. We use the official steps provided by WAT (Workshop on Asian Translation) to pre-process the dataset. The sizes of training, validation and test sets are 672k, 2.1k and 2.1k.

The three datasets we choose are of different scales. Considering the sizes of datasets, IWSLT14 and NC11 tasks could be used to simulate low-resource scenarios. ASPEC, on the other hand, is of large size and provides more source languages.

### 4.2 Experiment Settings

We use six layers of encoder-decoder architecture as the backbone. The number of the convolution layers in our parser is 3, and the dimension of the embedding layer for BPE is set to 256 before input into a projection layer by searching powers of 2 from 2 to 512. We use the syntactic mask in the first encoder layer for all the tasks.

We adopt the inverse square root learning rate scheduler, and the peak learning rate is set to  $5e-4$ ,  $1e-3$  and  $3e-4$  for IWSLT14, NC11 and ASPEC. The most appropriate values of the weighted parameter  $\lambda$  for the loss function are selected by doing a grid search over the range of 0.2 to 0.6 for each task.  $\lambda$  is chosen to be 0.47, 0.35, 0.3, 0.45 for IWSLT14, NC11, ASPEC Chinese to Japanese, and Japanese to Chinese respectively. More details of experiment settings are summarized in Appendix B.

Models	IWSLT14	
	De→En	En→De
Transformer	34.71	28.91
<i>Using external parsers</i>		
PASCAL	34.84	29.10
LISA*	34.97	29.06
ST-NMT*	35.24	-
Distance-Transformer*	35.74	29.28
<i>No external parsers</i>		
LPSI*	35.05	-
StructFormer	32.78	-
<b>Ours</b>	<b>35.89</b>	<b>29.31</b>

Table 1: BLEU scores on IWSLT14-De/En. While PASCAL, ST-NMT and Distance Transformer explicitly leverage external parsers to obtain syntactic information, LPSI, StructFormer and our method introduce self-induced grammar and do not require external parsers. "\*" means we report the results of ST-NMT, Distance Transformer and LPSI from their papers. The scores of LISA are the implementation by Bugliarello and Okazaki, 2020

Models	NC11	
	De→En	En→De
Transformer	27.04	25.23
PASCAL	27.65	25.52
LISA*	27.10	25.30
Distance-Transformer*	27.67	<b>26.19</b>
<b>Ours</b>	<b>27.69</b>	25.83

Table 2: BLEU scores on NC11-De/En.

### 4.3 Results

We take 5 random seeds for each task and average the last 5 checkpoints to be evaluated for each seed. We use the BLEU score to assess the performance of models on the test sets. Our experimental results on the three datasets are presented in Table 1, Table 2 and Table 3. We report the scores of Transformer we implement and other state-of-the-art (SOTA) models, including PASCAL (Bugliarello and Okazaki, 2020), LISA (Strubell et al., 2018), ST-NMT (Yang et al., 2020) and Distance Transformer (Hou et al., 2022), which are enhanced by external parsers, as well as LPSI (Harada and Watanabe, 2021), which unsupervisedly induces latent phrase structures and incorporate them into the attention mechanism. We also replace the Transformer encoder with StructFormer

Models	ASPEC	
	Zh→Ja	Ja→Zh
Transformer	47.66	34.36
Distance-Transformer*	48.34	-
<b>Ours</b>	<b>48.86</b>	<b>34.48</b>

Table 3: BLEU scores on ASPEC-Zh/Ja.

(Shen et al., 2021), which is an encoder-based model for grammar parsing, and train the model on IWSLT14 De→En.

Tables demonstrate that our method obtains SOTA results in most translation tasks. It consistently outperforms the vanilla Transformer across all six tasks, with particularly notable improvements observed in IWSLT14 German to English and ASPEC Chinese to Japanese tasks. It surpasses other external-parser-enhanced methods and the unsupervised method LPSI and StructFormer on IWSLT14-De/En.

It is worth noting that our method achieves scores comparable to the reported outcomes of Distance Transformer (Hou et al., 2022). They rely on the external parser to generate constituency grammar information so as to enhance the self-attention mechanism in Transformer. Moreover, when we reproduce their results, we find that their model benefits from attention dropout, and they use the best checkpoint at inference time. Compared with their supervised Distance Transformer, our unsupervised method learns to comprehend grammar structures without any additional knowledge or external tools.

In addition, it can be observed that all the other external-parser-enhanced methods are mainly applied in German and English since the parsers they leveraged are primarily developed for these two languages. Their models are affected by external parsers and limited in the diversity of languages. In contrast, our method internalizes grammar induction within Transformer for downstream tasks without the requirement for additional syntactic annotations. As a result, our approach can be applied across a wider range of languages.

### 4.4 Effect of BPE Embeddings

To assess the effect of BPE embeddings we have designed, we compare the performance of our method with and without the incorporation of BPE embeddings.

The average scores and standard deviations pre-

Models	IWSLT14		NC11		ASPEC	
	De→En	En→De	De→En	En→De	Zh→Ja	Ja→Zh
Transformer	34.71±0.12	28.91±0.14	27.04±0.10	25.23±0.17	47.66±0.25	34.36±0.09
Ours (-BPE)	35.83±0.07	29.19±0.12	27.39±0.11	25.27±0.63	48.78±0.23	<b>34.70±0.12</b>
<b>Ours (+BPE)</b>	<b>35.89±0.18</b>	<b>29.31±0.13</b>	<b>27.69±0.17</b>	<b>25.83±0.26</b>	<b>48.86±0.14</b>	34.48±0.27

Table 4: Effect of BPE embeddings. "-BPE" means using our method without BPE embeddings, while "+BPE" means using our method with BPE embeddings.

tasks	De→En		En→De	
	-BPE	+BPE	-BPE	+BPE
precision	35.55	<b>35.70</b>	25.63	<b>31.96</b>
recall	34.13	<b>34.27</b>	24.59	<b>30.66</b>
F1	34.83	<b>34.97</b>	25.10	<b>31.30</b>

Table 5: Parsing performance on IWSLT14.

sented in Table 4 indicate that while our BPE embeddings have a modest influence on machine translation, they do offer a positive contribution across all tasks except the ASPEC Japanese to Chinese task. It is observable that BPE embeddings are particularly beneficial in the context of English to German translations. This enhanced utility is likely attributed to the fact that the grammar status sharing among congenetic subwords is more compatible with English grammar. Regardless of the presence of BPE embeddings, our method shows consistent advancements over Transformer across machine translation tasks, showcasing the robustness and adaptability of our approach.

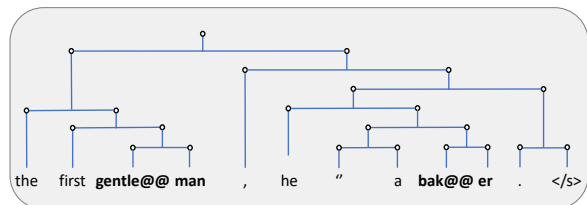
#### 4.5 Parsing Performance

Furthermore, we implement the distance-to-tree algorithm (Shen et al., 2018b) to reconstruct the hierarchical structure of a constituency tree, leveraging the syntactic distances generated during the parsing process. Following the definition of syntactic distance, the two words with the smallest distance will be merged into a tree first. The sequence of tokens, along with their syntactic distances, will be converted into a binary tree. The transformation algorithm can be referred to in Appendix D.

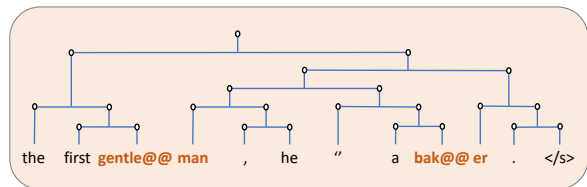
To assess the parsing performance of our method, we utilize the validation set of IWSLT14-De/En. The parser of Stanford CoreNLP<sup>3</sup> is employed to generate constituency trees as reference.<sup>4</sup> Preci-

<sup>3</sup><https://nlp.stanford.edu/software/segmenter.shtml>

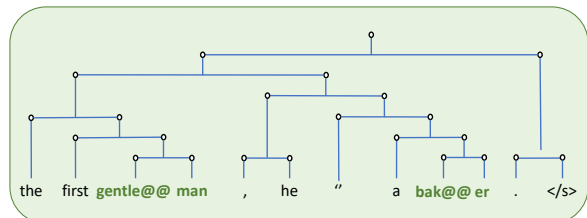
<sup>4</sup>Stanford CoreNLP was also leveraged by previous work (Htut et al., 2019) to generate standard parses for samples of



(a) The ground truth tree obtained by Stanford CoreNLP.



(b) The constituency tree obtained by using our method without BPE embeddings.



(c) The constituency tree obtained by using our method with BPE embeddings.

Figure 2: The constituency tree for the example sentence by using (a) Stanford CoreNLP, (b) our method without BPE embeddings, and (c) our method with BPE embeddings. "@@" in "gentle@@" and "bak@@" is the sign of BPE segmentation.

sion, recall, and F1 scores are computed to evaluate the trees generated by our method, with the results presented in Table 5.

The F1 scores of constituency trees generated by using our method achieve 34.97 and 31.30 for the source languages of German and English, re-

machine translation tasks. These generated parses served as references for evaluating parsing performance during translation.

Models	CoLA (mc)	SST-2 (acc)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)	Avg
RoBERTa	61.3	95.2	87.6/87.0	92.8	75.8	92.9	88.9	<b>90.9</b>	85.8
<i>Using external parsers</i>									
SLA*	60.0	93.3	-/-	91.4	67.8	-	-	89.9	-
Syntax-RoBERTa*	63.3	<b>96.1</b>	87.8/85.7	<b>94.3</b>	<b>81.2</b>	88.5	88.5	89.9	86.1
SynCLM*	<b>65.3</b>	95.1	87.2/-	93.0	80.1	<b>93.7</b>	88.9	90.8	-
<i>No external parsers</i>									
<b>Ours</b>	62.2 <sup>†</sup>	95.0	<b>87.7<sup>†</sup>/87.5<sup>†</sup></b>	93.3 <sup>†</sup>	79.8 <sup>†</sup>	92.8	<b>89.1<sup>†</sup></b>	<b>90.9</b>	<b>86.5<sup>†</sup></b>

Table 6: Scores on GLUE benchmark. The evaluation metrics are listed below the task names, where "mc" denotes Matthews correlation coefficient, "acc" denotes accuracy, "F1" denotes F1 score, and "pc" denotes Pearson correlation coefficients. The MNLI dataset has 2 versions of test sets represented as "m" and "mm". The scores with a superscript "†" denote our method outperforms the vanilla RoBERTa.

spectively. While the parsing performance may not match that of an expert-supervised grammar parser, with such moderate parsing accuracy, our method could achieve reasonable improvement in machine translation.

Additionally, it can be observed from Table 5 that the scores using BPE embeddings surpass those without BPE embeddings. In Figure 2, we illustrate constituency trees generated by the three different methods for an example sentence. Compared to the ground truth, the hierarchical structure of the constituency tree generated by using our method appears quite plausible. It suggests that BPE embeddings enhance the model’s ability to capture BPE segmentations, which proves that the grammar status sharing among congenetic subwords is practical and valid.

## 5 Language Understanding

We deploy our grammar induction method on RoBERTa (Liu et al., 2019) at the stage of fine-tuning, and evaluate its performance in the GLUE benchmark (Wang et al., 2019).

### 5.1 Datasets

The GLUE (General Language Understanding Evaluation) benchmark is a collection of datasets for evaluating the natural language understanding performance of models. It consists of (1) single-sentence classification tasks: CoLA (Corpus of Linguistic Acceptability) and SST-2 (Stanford Sentiment Treebank); (2) similarity and paraphrase tasks: MRPC (Microsoft Research Paraphrase Corpus), QQP (Quora Question Pairs) and STS-B (Semantic Textual Similarity Benchmark); (3) inference tasks: MNLI (Multi-Genre Natural Language

Inference Corpus), QNLI (Stanford Question Answering Dataset) and RTE (Recognizing Textual Entailment).

### 5.2 Experiment Settings

RoBERTa is a well and robustly pre-trained model with a strong ability for language understanding. It is essentially a Transformer model with only encoder layers. We use RoBERTa-base as our backbone. It has 12 encoder layers and 12 attention heads in each layer. We load official checkpoints of pre-trained RoBERTa. Details of the experimental settings are set following fairseq (Ott et al., 2019).

For our grammar-aware RoBERTa, the number of convolution layers in our parser is 2, and we use the syntactic mask in the first three encoder layers. We fine-tune and evaluate the model on each task of GLUE.

### 5.3 Results

The results on the GLUE benchmark are presented in Table 6. The best checkpoints are saved during training, and we use standard evaluation metrics for each task. We reproduce the vanilla RoBERTa and our grammar-enhanced RoBERTa in the same environment and report the scores in the table.

As shown in the table, our method outperforms the vanilla RoBERTa in the majority of GLUE tasks and achieves the highest average score. Remarkably, without relying on any additional syntactic annotations, our method has achieved improvements that are more consistent when compared to other syntax-enhanced models, including SLA (Li et al., 2021), Syntax-RoBERTa (Bai et al., 2021) and SynCLM (Zhang et al., 2022). To introduce syntactic information into pre-trained models, all of these

methods resort to external parsing models.

When integrated with our method for self-inducing grammatical knowledge, RoBERTa obtains improvements across six of the nine test sets in GLUE, with the scores on the remaining three test sets being nearly identical to those of the baseline RoBERTa. More specifically, it can be inferred from the table that pre-trained language models have exhibited such an impressive capability of language understanding that the integration of external parsers may not bring uniform enhancements across all tasks. They even hurt the performance in certain tasks, as evidenced by the performance of SLA on RTE, Syntax-RoBERTa on MRPC, and SynCLM on MNLI. Nonetheless, our approach performs comparably well in single-sentence classification tasks, and similarity and paraphrase tasks. It particularly excels in inference tasks, with a notable improvement observed in the RTE task.

## 6 Related Work

There are generally two ways for syntactic knowledge-based language comprehension.

**Explicit syntax enhancement** One way is to incorporate syntactic information explicitly by using a high-quality external parser. PASCAL (Bugliarello and Okazaki, 2020) builds up a syntax-guided localized attention mask where each token’s attention range is a Gaussian distribution centered by its dependency head. Distance Transformer (Hou et al., 2022) utilizes the relative magnitude of syntactic distance from the constituency tree to build up a syntactic local range. ST-NMT (Yang et al., 2020) slices the constituency tree at some depth and gets a list of constituents as labels. A new Transformer encoder is learned to predict this structure, and its output is integrated with the original encoder output. SEPREM (Xu et al., 2021) also focuses on pretraining architecture and defines a prior attention weight distribution by normalizing the inverse of the token’s distance in the dependency structure. However, all of these methods require sophisticated external parsers to generate dependency or constituency syntax trees.

**Implicit syntax enhancement** The other way is to implicitly induce the latent grammar structure by learning from data distribution. PRPN (Shen et al., 2018a) enhances RNN neural language modeling by simultaneously learning a CNN-based parsing network, which could induce latent constituency

structure represented by syntactic distance. Recent works have focused more on attention mechanisms. StructFormer (Shen et al., 2021) introduces a parser layer that can induce dependency and constituency syntax simultaneously and integrates induced dependency structure into self-attention in a differentiable way. LPSI (Harada and Watanabe, 2021) induces latent phrase structure in an unsupervised fashion and integrates them into the multi-head attention mechanism. SyncAttn (Deguchi et al., 2021) enables neural OpenIE to induce the latent syntactic structure and adopt multi-view learning to capture multiple relationships from constituency and dependency graphs. Nevertheless, very little work pays attention to leveraging implicit induction in downstream tasks.

**Syntax enhancement of pre-trained models** In recent years, people also studied to introduce grammar induction into pre-trained models like BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019) instead of training from scratch. POS-BERT (Sundararaman et al., 2019) leverages POS tag information apart from the syntactic structure. SLA (Li et al., 2021) prohibits two tokens from attending to each other once their distance in the dependency tree exceeds some boundary, and it improves the fine-tuning performance on several GLUE tasks. Syntax-Roberta (Bai et al., 2021) develops a framework that can be easily plugged into an arbitrary pre-trained checkpoint, which automatically underlines the most relevant syntactic knowledge for each downstream task. SynCLM (Zhang et al., 2022) proposes a syntax-guided contrastive learning method where phrase-guided and tree-guided contrastive objectives based on constituency and dependency structures are optimized in the pre-training stage to help the language model capture rich syntactic knowledge in its representations. However, these methods resort to external parsers to obtain syntactic information, which could not work well with pre-trained models across downstream tasks.

## 7 Conclusion

In this paper, we study the utilization of grammar induction for language understanding and generation. We force the model to be trained with the self-induced grammar without external parsers or annotations, which is a more general approach. Moreover, our method is efficient in both the from-scratch and pre-trained scenarios. Through assess-



ments across various machine translation and language comprehension tasks, we validate the efficacy of our grammar induction technique, showcasing substantial and consistent enhancements. These results underscore the tangible benefits that can be achieved through the incorporation of grammar induction techniques. Our research offers promising insights into the practical application of induced grammar in downstream tasks.

## Limitations

The performance of the grammar-aware method may be constrained in large-scale datasets and well-trained models. As shown in our paper, none of the grammar-aware methods can achieve consistent improvements on all the tasks of GLUE. This phenomenon could be explained by the adequacy of model training. Since RoBERTa has been pre-trained over a tremendous amount of data, it has already done well in language understanding. But it occurs to us whether pre-trained models or large language models have some form of syntax parsing modules inside them.

The parsing performance of our parser is a little limited. We found that the quality of constituency trees generated from the syntactic distance in our parser is far from that of well-designed and trained supervised constituency parsers. However, our method does make neural network models more interpretable. More appropriate optimization methods could be explored to make grammar induction more effective.

## Ethics Statements

Our work pertains to neural machine translation and fine-tuning of pre-trained language models to introduce grammar induction into Transformer. In this work, we use only publicly available data and artifacts.

## Acknowledgement

This work was sponsored by the National Natural Science Foundation of China (NSFC) grant (No.62106143), and National Key Research and Development Program of China (No. 2023ZD0121402).

## References

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. [Syntaxbert: Improving pre-trained transformers with syntax](#)

[trees](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3011–3020. Association for Computational Linguistics.

Eric Brill, David M. Magerman, Mitchell Marcus, and Beatrice Santorini. 1990. [Deducing linguistic structure from the statistics of large corpora](#). In *Next Decade in Information Technology: Proceedings of the 5th Jerusalem Conference on Information Technology 1990, Jerusalem, October 22-25, 1990*, pages 380–389. IEEE Computer Society.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Emanuele Bugliarelli and Naoaki Okazaki. 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*, pages 1618–1627. Association for Computational Linguistics.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. [Discrete opinion tree induction for aspect-based sentiment analysis](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2051–2064. Association for Computational Linguistics.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. [Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1816–1829. Association for Computational Linguistics.

Hiroyuki Deguchi, Akihiro Tamura, and Takashi Nomiya. 2021. [Synchronous syntactic attention for transformer neural machine translation](#). In *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, July 5-10, 2021*, pages 348–355. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of](#)

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan S. Kankanhalli. 2019. **Quantifying and alleviating the language prior problem in visual question answering**. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 75–84. ACM.
- Shintaro Harada and Taro Watanabe. 2021. Neural machine translation with synchronous latent phrase structure. In *ACL (student)*, pages 321–330. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. **A structural probe for finding syntax in word representations**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.
- Shengyuan Hou, Jushi Kai, Haotian Xue, Bingyu Zhu, Bo Yuan, Longtao Huang, Xinbing Wang, and Zhouhan Lin. 2022. **Syntax-guided localized self-attention by constituency syntactic distance**. *CoRR*, abs/2210.11759.
- Phu Mon Htut, Kyunghyun Cho, and Samuel R. Bowman. 2019. **Inducing constituency trees through neural machine translation**. *CoRR*, abs/1909.10056.
- Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. **Improving BERT with syntax-aware local attention**. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 645–653. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **Roberta: A robustly optimized BERT pretraining approach**. *CoRR*, abs/1907.11692.
- Hongyin Luo, Lan Jiang, Yonatan Belinkov, and Jim Glass. 2019. **Improving neural language models by segmenting, attending, and predicting the future**. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1483–1493. Association for Computational Linguistics.
- 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2032–2037. Association for Computational Linguistics.
- David M. Magerman and Mitchell P. Marcus. 1991. **Pearl: A probabilistic chart parser**. In *EACL 1991, 5th Conference of the European Chapter of the Association for Computational Linguistics, April 9-11, 1991, Congress Hall, Alexanderplatz, Berlin, Germany*, pages 15–20. The Association for Computer 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).
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. **fairseq: A fast, extensible toolkit for sequence modeling**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.
- Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. **Overcoming language priors in visual question answering with adversarial regularization**. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1548–1558.
- Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron C. Courville. 2018a. **Neural language modeling by jointly learning syntax and lexicon**. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordani, Aaron C. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1171–1180. Association for Computational Linguistics.
- Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron C. Courville. 2021. **Structformer:**

- Joint unsupervised induction of dependency and constituency structure from masked language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7196–7209. Association for Computational Linguistics.
- Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. [Linguistically-informed self-attention for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.
- Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. [Syntax-infused transformer and BERT models for machine translation and natural language understanding](#). *CoRR*, abs/1911.06156.
- Giorgos Tziafas, Konstantinos Kogkalidis, Gijs Wijnholds, and Michael Moortgat. 2021. [Improving BERT pretraining with syntactic supervision](#). *CoRR*, abs/2104.10516.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. [The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives](#). 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*, pages 4395–4405. Association for Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. [Syntax-enhanced pre-trained model](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5412–5422. Association for Computational Linguistics.
- Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou. 2020. [Improving neural machine translation with soft template prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5979–5989. Association for Computational Linguistics.
- Shuai Zhang, Lijie Wang, Xinyan Xiao, and Hua Wu. 2022. [Syntax-guided contrastive learning for pre-trained language model](#). In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2430–2440. Association for Computational Linguistics.

## A Estimation of Dependency Distribution

The estimation function  $\mathcal{F}_{\mathcal{P}}(\cdot)$  in Equation 4 leverages the two-stage method by Shen et al. (2021) to calculate the estimation of the dependency distribution among tokens. We utilize it to generate the syntactic mask with syntactic distance and height. We describe the calculation procedure in detail.

To estimate the probability  $p_{\mathcal{D}}(w_j|w_i)$  that the  $j$ -th token  $w_j$  is the parent of the  $i$ -th token  $w_i$ , they decompose  $p_{\mathcal{D}}(w_j|w_i)$  into two factors:

$$p_{\mathcal{D}}(j|i) = \sum_{[l,r]} p_{\mathcal{P}\mathbf{r}}(j|[l,r])p_{\mathcal{C}}([l,r]|i) \quad (7)$$

where  $p_{\mathcal{P}\mathbf{r}}(j|[l,r])$  denotes the probability that  $w_j$  is the root of  $w_{l\dots r}$ . It can be parameterized by syntactic height:

$$p_{\mathcal{P}\mathbf{r}}(j|[l,r]) = \frac{\exp(h_j)}{\sum_{l \leq k \leq r} \exp(h_k)} \quad (8)$$

And  $p_{\mathcal{C}}([l,r]|i)$  denotes the probability that  $w_{l\dots r}$  is the smallest legal constituent  $C(w_i)$  for  $w_i$ :

$$p_{\mathcal{C}}([l,r]|i) = p_{\mathcal{L}}(l|i)p_{\mathcal{R}}(r|i) \quad (9)$$

where  $p_{\mathcal{L}}(l|i)$  denotes the probability that  $l$  is the left margin of  $C(w_i)$  and  $p_{\mathcal{R}}(r|i)$  stands for the right margin.

The distribution is parameterized that the  $l$ -th token  $w_l$  is inside  $C(w_i)$  with the probability  $p(l \in C(w_i))$  that  $h_i$  is larger than the maximum distance  $\tau$  between  $l$  and  $i$ :

$$p(l \in \mathcal{C}(w_i)) = \sigma((h_i - \max(\tau_l, \dots, \tau_{i-1})) / \mu) \quad (10)$$

Models	IWSLT14		NC11	
	De→En	En→De	De→En	En→De
Transformer	34.71	28.91	27.04	25.23
Ours (-MLM)	34.41	28.44	26.24	25.26
<b>Ours (+MLM)</b>	<b>35.89</b>	<b>29.31</b>	<b>27.69</b>	<b>25.83</b>

Table 7: Effect of MLM loss on machine translation. "-MLM" means using our method without MLM loss, while "+MLM" means using our method with MLM loss.

Models	CoLA (mc)	SST-2 (acc)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)	Avg
RoBERTa	61.3	<b>95.2</b>	87.6/87.0	92.8	75.8	<b>92.9</b>	88.9	<b>90.9</b>	85.8
<b>Ours (-MLM)</b>	<b>62.2</b>	95.0	87.7/87.5	<b>93.3</b>	<b>79.8</b>	92.8	<b>89.1</b>	<b>90.9</b>	<b>86.5</b>
Ours (+MLM)	62.1	95.0	<b>88.0/87.5</b>	93.1	75.8	92.7	88.7	89.4	85.8

Table 8: Effect of MLM loss on GLUE.

where  $\sigma$  is the sigmoid function and  $\mu$  is a learnable temperature term.

The event  $l \in C(w_i)$  consists of 2 cases:  $l \in C(w_i) \cap l-1 \in C(w_i)$  and  $l \in C(w_i) \cap l-1 \notin C(w_i)$ .

For the first case, if  $l-1 \in C(w_i)$ , we must have  $l \in C(w_i)$ . So the probability can be measured with  $p(l-1 \in C(w_i))$ . For the second case, if  $l \in C(w_i)$  and  $l-1 \notin C(w_i)$ , it means that  $l$  is the left margin of  $C(w_i)$ .

Therefore, the probability  $p_L(l | i)$  that  $l$  is the left margin of  $C(w_i)$  can be derived by:

$$p_L(l | i) = p(l \in C(w_i)) - p(l-1 \in C(w_i)) \quad (11)$$

And  $p_R(r | i)$  can be derived by:

$$p_R(r | i) = p(r-1 \in C(w_i)) - p(r \in C(w_i)) \quad (12)$$

The procedure goes through the sentence to estimate the syntactic dependency of each token on the others.

## B Experiment Details

We perform all the training on 2 RTX3090 GPUs for IWSLT14 and NC11, 2 80G-A100 for ASPEC, and 1 40G-A100 for GLUE. For our method with BPE embeddings, each training took about 2.5 hours for IWSLT14, 3 hours for NC11, and 50 hours for ASPEC. For our method without BPE embeddings, each training took about 2 hours for IWSLT14, 2.5 hours for NC11, and 29 hours for ASPEC. We use Adam optimizer with  $\beta_1 = 0.9, \beta_2 = 0.98$ . Cross entropy is

the default loss function with label smoothing of 0.1 and weight decay of 0.0001. The batch size ( $GPU\_num \times Update\_num \times Max\_tokens$ ) is  $2 \times 1 \times 4096$ ,  $2 \times 2 \times 8192$  and  $2 \times 1 \times 8192$  respectively.

The numbers of attention heads in each layer and dimensions of the fully connected layer are 4 and 1024 for IWSLT14 and NC11, 8 and 2048 for ASPEC.

We employ beam search at inference time. Following the settings of previous work [Hou et al., 2022](#), the beam size and length penalty are set to 5 and 1.0 for IWSLT14 and ASPEC, and set to 4 and 0.6 for NC11.

The number of model parameters is about 49M for IWSLT14, 57M for NC11, and 108M for ASPEC, with an increase of 9M. For grammar-aware RoBERTa on GLUE, the fine-tuning took about 7 hours and the number of model parameters is about 135M with an increase of 10M.

We also see increased numbers of parameters in other syntax-enhanced models like SLA ([Li et al., 2021](#)), ST-NMT ([Yang et al., 2020](#)), LPSI ([Harada and Watanabe, 2021](#)) and Syntax-Roberta ([Bai et al., 2021](#)), while they do not report statistics in their papers.

## C Effect of MLM loss

We conduct experiments to study the effect of MLM loss on our method. The results on machine translation tasks and GLUE are shown in [Table 7](#) and [Table 8](#).

Experiment results demonstrate that MLM loss

---

**Algorithm 1** Syntactic Distance to Constituency Tree

---

**Input:** Syntactic distance  $\tau_{1\dots n-1}$ , sentence  $w_{1\dots n}$

**Output:** Constituency Tree  $T$

**Function** Tree( $\tau, w$ )

```
1: if  $\tau = []$  then  
2:    $T \leftarrow \text{Leaf}(w)$   
3: else  
4:    $i \leftarrow \text{argmax}_i(\tau)$   
5:    $\text{node}_l \leftarrow \text{Tree}(\tau_{1\dots i-1}, w_{1\dots i})$   
6:    $\text{node}_r \leftarrow \text{Tree}(\tau_{i+1\dots n-1}, w_{i+1\dots n})$   
7:    $T \leftarrow \text{Node}(\text{node}_l, \text{node}_r)$   
8: end if  
9: return  $T$ 
```

---

has a different effect on models' grammar integration in the two scenarios. As discussed in Section 3.4, the language modeling objective contributes to the hierarchical generalization of Transformers. Consequently, our grammar induction works well with MLM loss to train Transformers from scratch on machine translation tasks. Nevertheless, it does not benefit from MLM loss when fine-tuning a pre-trained model RoBERTa. One possible reason is that learning to induce grammar with MLM loss might influence hierarchical generalization inside the model, which has been developed by MLM loss in the pre-training of RoBERTa. It deserves further work to study the relation between language modeling and grammar induction during different stages of training.

## D Transformation Algorithm

The algorithm we implemented in Section 4.5 is to transform the syntactic distances of a sentence into a constituency tree.

The transformation algorithm is designed with a top-down merging method. It divides the input sequence  $w$  into two subsequences at the split point  $i$  whose syntactic distance  $\tau_i$  is the largest in the sequence. Then the subtrees of the two subsequences are merged into a whole.