

# SG-Net: Syntax Guided Transformer for Language Representation

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao\*, Rui Wang

**Abstract**—Understanding human language is one of the key themes of artificial intelligence. For language representation, the capacity of effectively modeling the linguistic knowledge from the detail-riddled and lengthy texts and getting rid of the noises is essential to improve its performance. Traditional attentive models attend to all words without explicit constraint, which results in inaccurate concentration on some dispensable words. In this work, we propose using syntax to guide the text modeling by incorporating explicit syntactic constraints into attention mechanisms for better linguistically motivated word representations. In detail, for self-attention network (SAN) sponsored Transformer-based encoder, we introduce syntactic dependency of interest (SDOI) design into the SAN to form an SDOI-SAN with syntax-guided self-attention. Syntax-guided network (SG-Net) is then composed of this extra SDOI-SAN and the SAN from the original Transformer encoder through a dual contextual architecture for better linguistics inspired representation. The proposed SG-Net is applied to typical Transformer encoders. Extensive experiments on popular benchmark tasks, including machine reading comprehension, natural language inference, and neural machine translation show the effectiveness of the proposed SG-Net design.

**Index Terms**—Artificial Intelligence, Natural Language Processing, Transformer, Language Representation, Reading Comprehension, Machine Translation.



## 1 INTRODUCTION

TEACHING machines to read and comprehend human languages is a long-standing goal of artificial intelligence, where the fundamental is language representation. Recently, much progress has been made in general-purpose language representation that can be used across a wide range of tasks [2, 3, 4, 5, 6]. Understanding the meaning of texts is a prerequisite to solve many natural language understanding (NLU) problems, such as machine reading comprehension (MRC) based question answering [7], natural language inference (NLI) [8], and neural machine translation (NMT) [9], all of which require a rich and accurate representation of the meaning of a sentence.

For language representation, the first step is to encode the raw texts in vector space using an encoder. The dominant language encoder has evolved from recurrent neural

networks (RNN) to Transformer [10] architectures. With stronger feature extraction ability than RNNs, Transformer has been widely used for encoding deep contextualized representations [2, 3, 11]. After encoding the texts, the next important target is to model and capture the text meanings. Although a variety of attentive models [9, 12] have been proposed to implicitly pay attention to key parts of texts, most of them, especially global attention methods [9] equally tackle each word and attend to all words in a sentence without explicit pruning and prior focus, which would result in inaccurate concentration on some dispensable words [13].

Taking machine reading comprehension based question answering (QA) task [7, 14] (Figure 1) as an example, we observe that the accuracy of MRC models decreases when answering long questions (shown in Section 7.1). Generally, if the text is particularly lengthy and detailed-riddled, it would be quite difficult for a deep learning model to understand as it suffers from noise and pays vague attention to the text components, let alone accurately answering questions [15, 16, 17]. In contrast, existing studies have shown that machines could read sentences efficiently by taking a sequence of fixation and saccades after a quick first glance [18] or referring to compressed texts [19], which inspire us to integrate structured information into a language representation model to obtain a hierarchical and focused representation. The most common source for structure information is syntactic parsing [20].

Incorporating human knowledge into neural models is one of the major research interests of artificial intelligence. Recent Transformer-based deep contextual language representation models have been widely used for learning universal language representations from large amounts of unlabeled data, achieving dominant results in a series of NLU benchmarks [2, 3, 11, 21, 22, 23]. However, they only learn from plain context-sensitive features such as character

- This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU long term AI project, Cutting-edge Machine Reading Comprehension and Language Model (Corresponding author: Hai Zhao).
- Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, R. Wang are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, and also with Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, and also with MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University.  
E-mail: {zhangzs, will8821, zhoujunru, 1140339019dsf}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn, wangrui.nlp@gmail.com.
- Part of this work was finished when Z. Zhang visited National Institute of Information and Communications Technology (NICT) and R. Wang was with NICT.
- Part of this study has been accepted as “SG-Net: Syntax-Guided Machine Reading Comprehension” [1] in the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020). This paper extends the previous syntax-guided attention method to natural language comprehension, inference, and generation tasks. We further conduct comprehensive experiments, to verify the effectiveness, as well as generalization ability on different benchmarks with thorough case studies.

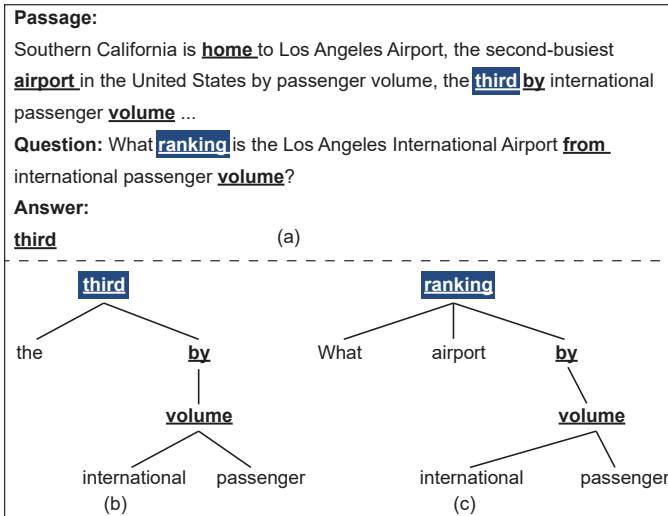


Fig. 1. (a) Example of syntax-guided span-based QA. The SDOI of each word consists of all its ancestor words. (b-c) The dependency parsing tree of the given passage sentence and question.

or word embeddings, with little consideration of explicitly extracting the hierarchical dependency structures that are entailed in human languages, which can provide rich dependency hints for language representation.

Besides, as a common phenomenon in language representation, an input sequence always consists of multiple sentences. Nearly all of the current attentive methods and language representation models regard the input sequence as a whole, e.g., a passage, with no consideration of the inner linguistic structure inside each sentence. This would result in process bias caused by much noise and a lack of associated spans for each concerned word.

All these factors motivate us to seek for an informative method that can selectively pick out important words by only considering the related subset of words of syntactic importance inside each input sentence explicitly. With the guidance of syntactic structure clues, the syntax-guided method could give more accurate attentive signals and reduce the impact of the noise brought about by lengthy sentences.

So far, we have two types of broadly adopted contextualized encoders for building sentence-level representation, i.e., RNN-based and Transformer-based [10]. The latter has shown its superiority, which is empowered by a self-attention network (SAN) design. In this paper, we extend the self-attention mechanism with syntax-guided constraint, to capture syntax related parts with each concerned word. Specifically, we adopt pre-trained dependency syntactic parse tree structure to produce the related nodes for each word in a sentence, namely syntactic dependency of interest (SDOI), by regarding each word as a child node and the SDOI consists all its ancestor nodes and itself in the dependency parsing tree. An example is shown in Figure 1.

To effectively accommodate such SDIOI information, we propose a novel syntax-guided network (SG-Net), which fuses the original SAN and SDIOI-SAN, to provide more linguistically inspired representation to comprehend language hierarchically. Our method allows the Transformer to learn

more interpretable attention weights that better explain how the model processes and comprehends the natural language.

The contribution of this paper is three-fold:

1) We propose a novel syntax-guided network (SG-Net) that induces explicit syntactic constraints to Transformer architectures for learning structured representation. To our best knowledge, we are the first to integrate syntactic relationships as attentive guidance for enhancing state-of-the-art SAN in the Transformer encoder.

2) Our model boosts the baseline Transformer significantly on three typical advanced natural language understanding tasks. The attention learning shows better interpretability since the attention weights are constrained to follow the induced syntactic structures. By visualizing the self-attention matrices, our model provides the information that better matches the human intuition about hierarchical structures than the original Transformer.

3) Our proposed method is lightweight and easy to implement. We can easily apply syntax-guided constraints in an encoder by adding an extra layer with the SDIOI mask, which makes it easy enough to be applicable to other systems.

## 2 BACKGROUND

### 2.1 Language Representation

Language representation is the foundation of deep learning methods for natural language processing and understanding. As the basic unit of language representation, learning word representations has been an active research area, and aroused great research interests for decades, including non-neural [24, 25, 26] and neural methods [27, 28].

Distributed representations have been widely used as a standard part of natural language processing (NLP) models due to the ability to capture the local co-occurrence of words from large scale unlabeled text [27]. However, when using the learned embedding for NLP tasks, these approaches for word vectors only involve a single, context independent representation for each word with a word-vectors lookup table, with little consideration of contextual encoding in sentence level. Thus recently introduced contextual language representation models including ELMo [11], GPT [2] and BERT [3] fill the gap by strengthening the contextual sentence modeling for better representation,<sup>1</sup> showing powerful to boost NLU tasks to reach new high performance. Two stages of training are adopted in these models: firstly, pre-train a model using language model objectives on a large-scale text corpus, and then fine-tune the model (as an pre-trained encoder with simple linear layers) in specific downstream NLP tasks (e.g., text classification, question answering, natural language inference). Among these contextualized language representation models, BERT uses a different pre-training objective, masked language model, which allows capturing both sides of context, left and right. Besides, BERT also introduces a *next sentence prediction* task that jointly pre-trains text-pair representations. The latest evaluation shows

1. Strictly speaking, the pre-trained models are not traditionally-defined language models when in use, as the former are mainly used as encoders to vectorize texts, which we call *language (representation) model* following the formulation in previous works [3].

that BERT is powerful and convenient for downstream NLU tasks.

The major technical improvement over traditional embeddings of these newly proposed language representation models is that they focus on extracting context-sensitive features from language models. When integrating these contextual word embeddings with existing task-specific architectures, ELMo helps boost several major NLP benchmarks [11] including question answering on SQuAD, sentiment analysis [29], and named entity recognition [30], while BERT especially shows effective on language understanding tasks on GLUE [31], MultiNLI [32] and SQuAD [3]. In this work, we follow this line of extracting context-sensitive features and take pre-trained BERT as our backbone encoder for jointly learning explicit context semantics.

However, the pre-trained language representation models only exploit plain context-sensitive features such as character or word embeddings. They rarely consider incorporating explicit, structured syntactic information, which can provide rich dependency information for language representation. To promote the ability to model the structured dependency of words in a sentence, we are inspired to incorporate extra syntactic constraints into the multi-head attention mechanism in the dominant language representation models. The latest evaluation shows that BERT is powerful and convenient for downstream tasks. Following this line, we extract context-sensitive syntactic features and take pre-trained BERT as our backbone encoder to verify the effectiveness of our proposed SG-Net.

## 2.2 Syntactic Structures

Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies [33]. For parsing dependency structures, the salient resources are Treebanks, which are collections of sentences that have been manually annotated with a correct syntactic analysis and part-of-speech (PoS) tags. Among them, the Penn Treebank (PTB) is an annotated corpus consisting of 50k sentences with over 4.5 million words of American English [34].

Generally, three annotation formats are used for dependency parsing [35], including two classic representations for dependency parsing, namely, *Stanford Basic* (SB) and *CoNLL Syntactic Dependencies* (CD), and bilexical dependencies from the *head-driven phrase structure grammar* (HPSG) English Resource Grammar (ERG), so-called *DELPH-IN Syntactic Derivation Tree* (DT). The annotation comparison is shown in Figure 2. HPSG is a highly lexicalized, constraint-based grammar developed by Pollard and Sag (1994) [36], which enjoys a uniform formalism representing rich contextual syntactic and even semantic meanings. HPSG divides language symbols into categories of different types, such as vocabulary, phrases, etc. The complete language symbol which is a complex type feature structure represented by attribute value matrices (AVMs) includes phonological, syntactic, and semantic properties, the valence of the word and interrelationship between various components of the phrase structure. Based on the above advantages of modeling word interrelationships of HPSG, we use the HPSG format to obtain the syntactic dependencies in this work.

There are three major classes of parsing models [37], *transition-based* [38, 39], *graph-based* [40, 41], and *grammar-*

*based models* [42, 43]. Recently, dependency syntactic parsing has been further developed with neural network and attained new state-of-the-art results [44, 45, 46, 47]. Benefiting from the highly accurate parser, neural network models could enjoy even higher accuracy gains by leveraging syntactic information rather than ignoring it [48, 49, 50, 51, 52].

Syntactic dependency parse tree provides a form that is capable of indicating the existence and type of linguistic dependency relation among words, which has been shown generally beneficial in various natural language understanding tasks [53]. To effectively exploit syntactic clue, our early work [50] extended local attention with syntax-distance constraint in RNN encoder by focusing on syntactically related words with the predicted target words for neural machine translation; Kasai et al. (2019) [54] absorbed parse tree information by transforming dependency labels into vectors and simply concatenated the label embedding with word representation. However, such simplified and straightforward processing would result in higher dimensions of the joint word and label embeddings and is too coarse to capture contextual interactions between the associated labels and the mutual connections between labels and words. This inspires us to seek for an attentive way to enrich the contextual representation from the syntactic source. A related work is from Strubell et al. (2018) [55], which proposed to incorporate syntax with multi-task learning for semantic role labeling. However, their syntax is incorporated by training one extra attention head to attend to syntactic ancestors for each token while we use all the existing heads rather than adding an extra one. Besides, this work is based on the remarkable representation capacity of recent language representation models such as BERT, which have been suggested to be endowed with some syntax to an extent [56]. Therefore, we are motivated to apply syntactic constraints through syntax guided method to prune the self-attention instead of purely adding dependency features.

In this work, we form a general approach to benefit from syntax-guided representations, which is the first attempt for the SAN architecture improvement in Transformer encoder to our best knowledge. The idea of updating the representation of a word with information from its neighbors in the dependency tree, which benefits from explicit syntactic constraints, is well linguistically motivated.

Our work differentiates from previous studies by both sides of technique and motivation:

1) Existing work [50, 57] only aimed to improve the dependency modeling in RNN while this is the pioneering work to integrate the parsing structure into pre-trained Transformer with a new methodology as SDOI mask. Our practice is much more difficult as Transformer is a stronger baseline with some ability to capture dependency itself;

2) Previous work majorly focused on similar fundamental linguistic tasks [55]. In contrast, we are motivated to apply linguistics (e.g., syntax) to more advanced NLU tasks (e.g., MRC), which are more complex and closer to AI. For NLU, pre-trained Transformers-based language representation models (e.g., BERT) have been widely used. It would be more beneficial for downstream tasks by taking them as the backbone, instead of focusing on RNNs that are less important for NLU recently. The advance would be potential to inspire the applications in other NLP and NLU tasks.

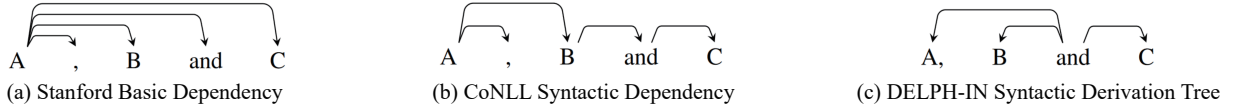


Fig. 2. Example of dependency formats.

TABLE 1  
Our notations and symbols.

Symbol	Meaning
$T$	Input text sequence
$X$	Embedding of the input sequence
$\mathbf{H}$	Contextual representation from the encoder
$\mathcal{M}$	SDOI mask
$H'$	Output of the syntax-guided layer
$\tilde{H}$	Output of the context aggregation
FFN	Feedforward Network
MultiHead	Multihead attention layer

For the application, our method is lightweight and easy to cooperate with other neural models. Without the need for architecture modifications in existing models, we can easily apply syntax-guided constraints in an encoder by adding an extra layer fed by the output representation of the encoder and the SDOI mask, which makes it easy enough to be applicable to other systems.

### 3 SYNTAX-GUIDED NETWORK

Our goal is to design an effective neural network model that makes use of structured linguistic information as effectively as possible. We first present the general syntax-guided attentive architecture, building upon the recent advanced Transformer-based encoder as task-agnostic architecture and then fit with downstream task-specific layers for downstream language comprehension tasks.<sup>2</sup>

Figure 3 depicts the whole architecture of our model. Our model first directly takes the output representations from a SAN-empowered Transformer-based encoder, then builds a syntax-guided SAN from the SAN representations. At last, the syntax-enhanced representations are fused from the syntax-guided SAN and the original SAN and passed to task-specific layers for final predictions. Table 1 summarizes the notations and symbols used in this article.

#### 3.1 Transformer Encoder

The raw text sequences are firstly represented as embedding vectors to feed an encoder (e.g., a pre-trained language representation model such as BERT [3]). The input sentence is first tokenized to word pieces (subword tokens). Let  $T = \{t_1, \dots, t_n\}$  denote a sequence of subword tokens of length  $n$ . For each token, the input embedding is the sum of its token embedding, position embedding, and token-type embedding.<sup>3</sup> Let  $X = \{x_1, \dots, x_n\}$  be the embedding

2. Note that our method is not limited to cooperate with BERT in our actual use, but any encoder with a self-attention network (SAN) architecture.

3. For BERT-like models [3, 23], they often take sequence pairs (e.g., passage + question) which are packed into a single sequence as input. Token type here is used to indicate whether each token belongs to the first sequence, or the second one.

of the sequence, which are features of encoding sentence words of length  $n$ . The input embeddings are then fed into the multi-head attention layer to obtain the contextual representations.

The embedding sequence  $X$  is processed to a multi-layer bidirectional Transformer for learning contextual representations, which is defined as

$$\mathbf{H} = \text{FFN}(\text{MultiHead}(K, Q, V)), \quad (1)$$

where  $K, Q, V$  are packed from the input sequence representation  $X$ . As the common practice, we set  $K = Q = V$  in the implementation.

In detail, the function of Eq. (1) is defined as follows. Let  $X^l = \{x_1^l, \dots, x_n^l\}$  be the features of the  $l$ -th layer. In the  $(l+1)$ -th layer, the corresponding features  $x^{l+1}$  are computed by

$$\tilde{h}_i^{l+1} = \sum_{m=1}^M W_m^{l+1} \left\{ \sum_{j=1}^N A_{i,j}^m \cdot V_m^{l+1} x_j^l \right\}, \quad (2)$$

$$h_i^{l+1} = \text{LayerNorm}(x_i^l + \tilde{h}_i^{l+1}), \quad (3)$$

$$\tilde{x}_i^{l+1} = W_2^{l+1} \cdot \text{GELU}(W_1^{l+1} h_i^{l+1} + b_1^{l+1}) + b_2^{l+1}, \quad (4)$$

$$x_i^{l+1} = \text{LayerNorm}(h_i^{l+1} + \tilde{x}_i^{l+1}), \quad (5)$$

where  $m$  is the index of the attention heads, and  $A_{i,j}^m \propto \exp[(Q_m^{l+1} x_i^l)^\top (K_m^{l+1} x_j^l)]$  denotes the attention weights between elements  $i$  and  $j$  in the  $m$ -th head, which is normalized by  $\sum_{j=1}^N A_{i,j}^m = 1$ .  $W_1^{l+1}, Q_1^{l+1}, K_1^{l+1}$  and  $V_m^{l+1}$  are learnable weights for the  $m$ -th attention head,  $W_1^{l+1}, W_2^{l+1}$  and  $b_1^{l+1}, b_2^{l+1}$  are learnable weights and biases, respectively.

For the following part, we use  $\mathbf{H} = \{h_1, \dots, h_n\}$  to denote the last-layer hidden states of the input sequence.

#### 3.2 Syntax-Guided self-attention Layer

Our syntax-guided representation is obtained through two steps. Firstly, we pass the encoded representation from the Transformer encoder to a syntax-guided self-attention layer. Secondly, the corresponding output is aggregated with the original encoder output to form a syntax-enhanced representation. It is designed to incorporate the syntactic tree structure information inside a multi-head attention mechanism to indicate the token relationships of each sentence, which will be demonstrated as follows.

In this work, we first pre-train a syntactic dependency parser to annotate the dependency structures for every sentence, which are then fed to SG-Net as guidance of token-aware attention. Details of the pre-training process of the parser are reported in Section 5.

To use the relationship between head word and dependent words provided by the syntactic dependency tree of sentence, we restrain the scope of attention only between

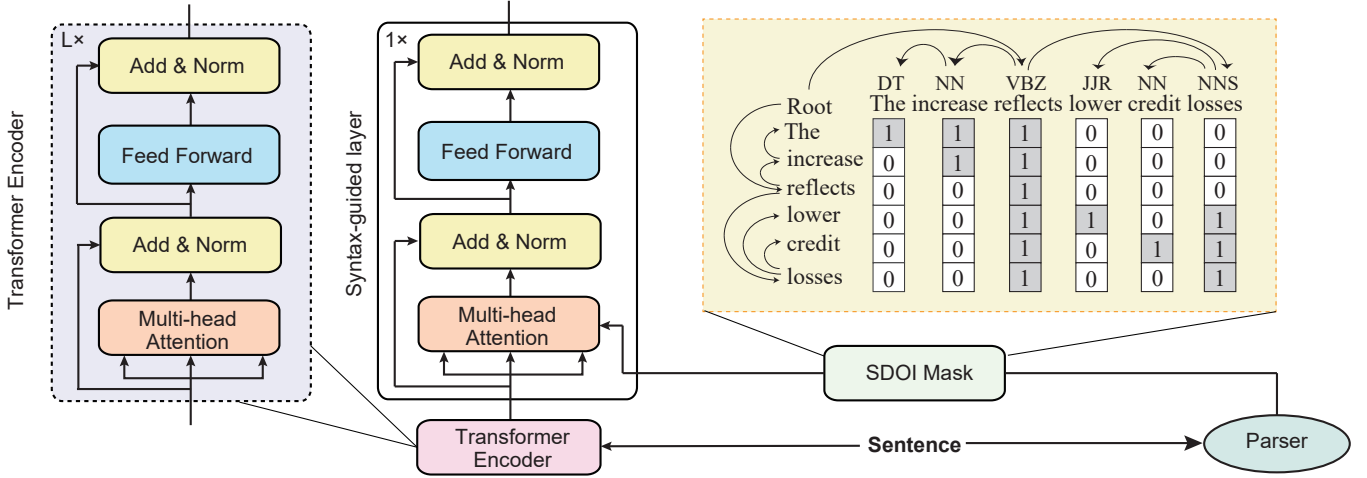


Fig. 3. Overview of the syntax-guided network.

word and all of its ancestor head words.<sup>4</sup> In other words, we would like to have each word only attend to words of syntactic importance in a sentence, the ancestor head words in the view of the child word. As the SDOI mask shown in Figure 3, instead of taking attention with each word in whole passage, the word *credit* only makes attention with its ancestor head words *reflects* and *losses* and itself in this sentence, which means that the SDOI of *credit* contains *reflects*, *losses* along with itself. Language representation models usually use special tokens, such as [CLS], [SEP], and [PAD] used in BERT, and <PAD>, </s> tokens used in Transformer for NMT. The SDOI of these special tokens is themselves alone in our implementation, which means these tokens will only attend to themselves in the syntax-guided self-attention layer.

Specifically, given input token sequence  $S = \{s_1, s_2, \dots, s_n\}$  where  $n$  denotes the sequence length, we first use syntactic parser to generate a dependency tree. Then, we derive the ancestor node set  $P_i$  for each word  $s_i$  according to the dependency tree. Finally, we learn a sequence of SDOI mask  $\mathcal{M}$ , organized as  $n * n$  matrix, and elements in each row denote the dependency mask of all words to the row-index word.

$$\mathcal{M}[i, j] = \begin{cases} 1, & \text{if } j \in P_i \text{ or } j = i \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Obviously, if  $\mathcal{M}[i, j] = 1$ , it means that token  $s_i$  is the ancestor node of token  $s_j$ . As the example shown in Figure 3, the ancestors of *credit* ( $i=4$ ) are *reflects* ( $j=2$ ), *losses* ( $j=5$ ) along with itself ( $j=4$ ); therefore,  $\mathcal{M}[4, (2, 4, 5)] = 1$  and  $\mathcal{M}[4, (0, 1, 3)] = 0$ .

We then project the last layer output  $H$  from the vanilla Transformer into the distinct key, value, and query representations of dimensions  $L \times d_k$ ,  $L \times d_v$ , and  $L \times d_q$ , respectively, denoted  $K'_i$ ,  $Q'_i$  and  $V'_i$  for each head  $i$ .<sup>5</sup> Then we perform a

dot product to score key-query pairs with the dependency of interest mask to obtain attention weights of dimension  $L \times L$ , denoted  $A'_i$ :

$$A'_i = \text{Softmax} \left( \frac{\mathcal{M} \cdot (Q'_i K'_i T)}{\sqrt{d_k}} \right). \quad (7)$$

We then multiply attention weight  $A'_i$  by  $V'_i$  to obtain the syntax-guided token representations:

$$W'_i = A'_i V'_i. \quad (8)$$

Then  $W'_i$  for all heads are concatenated and passed through a feed-forward layer followed by GeLU activations [58]. After passing through another feed-forward layer, we apply a layer normalization to the sum of output and initial representation to obtain the final representation, denoted as  $H' = \{h'_1, h'_2, \dots, h'_n\}$ .

### 3.3 Dual Context Aggregation

Considering that we have two representations now, one is  $H = \{h_1, h_2, \dots, h_n\}$  from the Transformer encoder, the other is  $H' = \{h'_1, h'_2, \dots, h'_n\}$  from syntax-guided layer from the above part. Formally, the final model output of our SG-Net  $\bar{H} = \{\bar{h}_1, \bar{h}_2, \dots, \bar{h}_n\}$  is computed by:

$$\bar{h}_i = \alpha h_i + (1 - \alpha) h'_i. \quad (9)$$

### 3.4 Task-specific Adaptation

#### 3.4.1 Machine Reading Comprehension

We focus on two types of reading comprehension tasks, i.e., *span-based* and *multi-choice* style which can be described as a tuple  $\langle P, Q, A \rangle$  or  $\langle P, Q, C, A \rangle$  respectively, where  $P$  is a passage (context), and  $Q$  is a query over the contents of  $P$ , in which a span or choice  $C$  is the right answer  $A$ . For the span-based one, we implemented our model on SQuAD 2.0 task that contains unanswerable questions. Our system is supposed to not only predict the start and end position in the passage  $P$  and extract span as answer  $A$  but also return a null string when the question is unanswerable. For the multi-choice style, the model is implemented on the

4. We extend the idea of using parent in Strubell et al. (2018) [55] to ancestor for a wider receptive range.

5. In SG-Net, there is only one extra attention layer compared with multi-layer Transformer architecture. The total numbers of our model and the baseline Transformer are very close, e.g., taking BERT as the baseline backbone, the parameters are 347M (+SG layer) vs 335M.

RACE dataset which is requested to choose the right answer from a set of candidate ones according to given passage and question.

Here, we formulate our model for both of the two tasks and feed the output from the syntax-guided network to task layers according to the specific task. Given the passage  $P$ , the question  $Q$ , and the choice  $C$  specially for RACE, we organize the input  $X$  for the encoder as the following two sequences.

Span:	[CLS]	P	[SEP]	Q	[SEP]
Choice:	[CLS]	P    Q	[SEP]	C	[SEP]

where || denotes concatenation.

In this work, pre-trained BERT is adopted as our detailed implementation of the Transformer encoder. Thus the sequence is fed to the BERT encoder mentioned above to obtain the contextualized representation  $H$ , which is then passed to our proposed syntax-guided self-attention layer and aggregation layer to obtain the final syntax-enhanced representation  $\bar{H}$ . To keep simplicity, the downstream task-specific layer basically follows the implementation of BERT. We outline below to keep the integrity of our model architecture. For the span-based task, we feed  $\bar{H}$  to a linear layer and obtain the probability distributions over the start and end positions through a softmax. For the multi-choice task, we feed it into the classifier to predict the choice label for the multi-choice model.

•**SQuAD 2.0** For SQuAD 2.0, our aim is a span of answer text; thus we employ a linear layer with SoftMax operation and feed  $\bar{H}$  as the input to obtain the start and end probabilities,  $s$  and  $e$ :

$$s, e = \text{SoftMax}(\text{Linear}(\bar{H})). \quad (10)$$

The training objective of our SQuAD model is defined as cross entropy loss for the start and end predictions,

$$\mathcal{L}_{has} = y_s \log s + y_e \log e. \quad (11)$$

For prediction, given output start and end probabilities  $s$  and  $e$ , we calculate the has-answer score  $score_{has}$  and the no-answer score  $score_{null}$ :

$$\begin{aligned} score_{has} &= \max(s_k + e_l), 1 < k \leq l \leq n, \\ score_{null} &= s_1 + e_1, \end{aligned} \quad (12)$$

We obtain a difference score  $score_{diff}$  between  $score_{has}$  and  $score_{null}$  as the final score. A threshold  $\delta$  is set to determine whether the question is answerable, which is heuristically computed with dynamic programming according to the development set. The model predicts the answer span that gives the has-answer score if the final score is above the threshold, and null string otherwise.

We find that adding an extra answer verifier module could yield a better result, which is trained in parallel only to determine whether question is answerable or not. The logits of the verifier are weighted with  $score_{null}$  to give the final predictions. In detail, we employ a parallel MRC model (i.e., SG-Net) whose training objective is to measure the answerability of the given questions. The verifier’s function

can be implemented as a cross-entropy loss. The pooled first token (the special symbol, [CLS]) representation  $\bar{h}_1 \in \bar{H}$ , as the overall representation of the sequence, is passed to a fully connection layer to get classification logits  $\hat{y}_i$  composed of answerable ( $logit_{ans}$ ) and unanswerable ( $logit_{na}$ ) elements. We use cross entropy as training objective:

$$\mathcal{L}_{ans} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)], \quad (13)$$

where  $\hat{y}_i \propto \text{SoftMax}(\text{Linear}(\bar{h}_1))$  denotes the prediction and  $y_i$  is the target indicating whether the question is answerable or not.  $N$  is the number of examples. We calculate the difference as the new verification score:  $score_{ext} = logit_{na} - logit_{ans}$ . The no-answer score is calculated as:

$$score_{final} = \beta_1 score_{diff} + \beta_2 score_{ext}, \quad (14)$$

where  $\beta_1$  and  $\beta_2$  are weights. Our model predicts the answer span if  $score_{final} > \delta$ , and null string otherwise.

•**RACE** As discussed in Devlin et al. (2019) [3], the pooled representation explicitly includes classification information during the pre-training stage of BERT. We expect the pooled to be an overall representation of the input. Thus, the first token representation  $\bar{h}_0$  in  $\bar{H}$  is picked out and is passed to a feed-forward layer to give the prediction  $p$ . For each instance with  $n$  choice candidates, we update model parameters according to cross-entropy loss during training and choose the one with the highest probability as the prediction when testing. The training objectives of our RACE model is defined as,

$$L(\theta) = -\frac{1}{N} \sum_i y_i \log p_i, \quad (15)$$

where  $p_i$  denotes the prediction,  $y_i$  is the target, and  $i$  denotes the data index.

### 3.4.2 Natural Language Inference

Since the natural language inference task is modeled as  $n$ -label classification problem, the model implementation is similar to RACE. The first token representation  $\bar{h}_0$  in  $\bar{H}$  is picked out and is passed to a feed-forward layer to give the prediction  $p$ . For each example with  $n$  labels (e.g., *entailment*, *neutral* and *contradiction*), we update model parameters according to cross-entropy loss during training and choose the one with highest probability as the prediction when testing.

### 3.4.3 Neural Machine Translation

Intuitively, NMT aims to produce a target word sequence with the same meaning as the source sentence, which is a natural language generation task. A Transformer NMT model consists of an encoder and a decoder, which fully rely on self-attention networks (SANs), to translate a sentence in one language into another language with equivalent meaning.

We introduce the syntax-guided self-attention layer in the encoder in the same way as the other tasks. The SAN of decoder then uses both  $\bar{H}_i$  and target context hidden state  $H_{tgt}$  to learn the context vector  $o_i$  by “encoder-decoder attention”:

$$\begin{aligned} c_i &= \text{FFN}(\text{MultiHead}(H_{tgt}, H_{tgt}, \bar{H}_i)), \\ o_i &= c_i + H_{tgt}. \end{aligned} \quad (16)$$

Finally, the context vector  $o_i$  is used to compute translation probabilities of the next target word  $y_i$  by a linear, potentially multi-layered function:

$$P(y_i | y_{<i}, x) \propto \text{SoftMax}(L_0 \text{GELU}(L_w o_i)), \quad (17)$$

where  $L_0$  and  $L_w$  are projection matrices.

## 4 TASK SETUP

### 4.1 Reading Comprehension

Our experiments and analysis are carried on two data sets, involving span-based and multi-choice MRC and we use the fine-tuned cased BERT (whole word masking) as the baseline.

- **Span-based MRC.** As a widely used MRC benchmark dataset, SQuAD 2.0 [7] combines the 100,000 questions in SQuAD 1.1 [14] with over 50,000 new, unanswerable questions that are written adversarially by crowdworkers to look similar to answerable ones. For the SQuAD 2.0 challenge, systems must not only answer questions when possible, but also abstain from answering when no answer is supported by the paragraph. Two official metrics are selected to evaluate the model performance: Exact Match (EM) and a softer metric F1 score, which measures the weighted average of the precision and recall rate at a character level.

- **Multi-choice MRC.** Our multi-choice MRC is evaluated on Large-scale ReAding Comprehension Dataset From Examinations (RACE) dataset [59], which consists of two subsets: RACE-M and RACE-H corresponding to middle school and high school difficulty levels. RACE contains 27,933 passages and 97,687 questions in total, which is recognized as one of the largest and most difficult datasets in multi-choice MRC. The official evaluation metric is accuracy.

### 4.2 Natural Language Inference

Natural language inference (NLI) is proposed to serve as a benchmark for natural language understanding and inference, which is also known as recognizing textual entailment (RTE). In this task, a model is presented with a pair of sentences and asked to judge the relationship between their meanings, including entailment, neutral, and contradiction. Bowman et al. (2015) [8] released Stanford Natural language Inference (SNLI) dataset, which is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels *entailment*, *contradiction*, and *neutral*. As a high-quality and large-scale benchmark, the SNLI corpus has inspired various significant work [4, 60, 61, 62, 63, 64].

### 4.3 Neural Machine Translation

The proposed NMT model was evaluated on the WMT14 English-to-German (EN-DE) translation task, which is a standard large-scale corpus for NMT evaluation. For the translation task, 4.43M bilingual sentence pairs of the WMT14 dataset were used as training data, including Common Crawl, News Commentary, and Europarl v7. The *newstest2013* and *newstest2014* datasets were used as the dev set and test set, respectively. The evaluation metric is BLEU score.

## 5 IMPLEMENTATION

### 5.1 Syntactic Parser

For the syntactic parser, we adopt the dependency parser from Zhou and Zhao (2019) [65] by joint learning of constituent parsing [66] using BERT as sole input which achieves very high accuracy: 97.00% UAS and 95.43% LAS on the English dataset Penn Treebank (PTB) [67] test set.<sup>6</sup> Due to the fact that the parsing corpus is annotated in word-level, the parser can only produce the word-level tree structures. Therefore, previous works that employ syntactic information conduct experiments in word-level [50, 68]. However, the word-level setting will harm the baseline performance as the recent dominant models commonly benefit from subword tokens as input due to the efficiency and effectiveness as the minimal language unit [69, 70, 71].

In SG-Net, to avoid the harm to the advanced subword-based strong Transformer baseline, we transform the word-level parsing structures into subword-level to ensure that the SDOI mask is in the same shape as the subword sequence shape. The transformation is based on the criterion that if the word is segmented into subwords, the subwords share the same tag as the original word. Alternatively, we tried to use the other method: If a word is segmented into  $n$  subwords, the [1:n] subwords are regarded as the children of the first subword. We find that taking both of the criteria yields very similar final results, thus only report the former one.

Note the above parsing work for sentences is done in data preprocessing, and our parser is not updated with the following downstream models. For passages in the MRC tasks, the paragraphs are parsed sentence by sentence.

### 5.2 Downstream-task Models

For MRC model implementation, we adopt the Whole Word Masking BERT and ALBERT as the baselines.<sup>7</sup> The initial learning rate was set in  $\{8e-6, 1e-5, 2e-5, 3e-5\}$  with a warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size was selected in  $\{16, 20, 32\}$ . The maximum number of epochs was set to 3 or 10 depending on tasks. The weight of  $\alpha$  in the dual context aggregation was 0.5. All the texts were tokenized using wordpieces, and the maximum input length was set to 384 for both of SQuAD and RACE. The configuration for multi-head self-attention was the same as that for BERT.

For the NLU task, the baseline is also the Whole Word Masking BERT. We used the pre-trained weights of BERT and followed the same fine-tuning procedure as BERT without any modification. The initial learning rate was set in the range  $\{2e-5, 3e-5\}$  with a warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size was selected from  $\{16, 24, 32\}$ . The maximum number of epochs was set ranging from 2 to 5. Texts were tokenized using wordpieces, with a maximum length of 128.

6. We report the results without punctuation of the labeled and unlabeled attachment scores (LAS, UAS).

7. <https://github.com/huggingface/transformers>. Only for MRC tasks, the BERT baseline was further improved as strong baseline by synthetic self training following <https://nlp.stanford.edu/seminar/details/jdevlin.pdf>.

TABLE 2

Exact Match (EM) and F1 scores (%) on SQuAD 2.0 dataset for single models. Our model is in boldface. † refers to unpublished works. Our final model is significantly better than the baseline BERT with p-value < 0.01.

Model	Dev		Test	
	EM	F1	EM	F1
<i>Regular Track</i>				
Joint SAN	69.3	72.2	68.7	71.4
U-Net	70.3	74.0	69.2	72.6
RMR + ELMo + Verifier	72.3	74.8	71.7	74.2
<i>BERT Track</i>				
Human	-	-	86.8	89.5
BERT [3]	-	-	82.1	84.8
BERT + MMFT + ADA†	-	-	83.0	85.9
Insight-baseline-BERT†	-	-	84.8	87.6
SemBERT [4]	-	-	84.8	87.9
BERT + CLSTM + MTL + V†	-	-	84.9	88.2
BERT + NGM + SST†	-	-	85.2	87.7
BERT + DAE + AoA†	-	-	85.9	88.6
XLNet [21]	87.9	90.6	87.9	90.7
ALBERT [23]	87.4	90.2	88.1	90.9
<i>Our implementation</i>				
BERT Baseline	84.1	86.8	-	-
<b>SG-Net on BERT</b>	85.1	87.9	-	-
<b>+Verifier</b>	85.6	88.3	85.2	87.9
ALBERT Baseline	87.0	90.1	-	-
<b>SG-Net on ALBERT</b>	87.4	90.5	-	-
<b>+Verifier</b>	88.0	90.8	-	-

For the NMT task, our baseline is Transformer [72]. We used six layers for the encoder and the decoder. The number of dimensions of all input and output layers was set to 512. The inner feed-forward neural network layer was set to 2048. The heads of all multi-head modules were set to eight in both encoder and decoder layers. The byte pair encoding algorithm [69] was adopted, and the size of the vocabulary was set to 40,000. In each training batch, a set of sentence pairs contained approximately  $4096 \times 4$  source tokens and  $4096 \times 4$  target tokens. During training, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were  $p = 0.1$ . The Adam optimizer [73] was used to tune the parameters of the model. The learning rate was varied under a warm-up strategy with 8,000 steps. For evaluation, we validated the model with an interval of 1,000 batches on the dev set. Following the training of 300,000 batches, the model with the highest BLEU score of the dev set was selected to evaluate the test sets. During the decoding, the beam size was set to five.

For span-based MRC, we conducted the significance test using McNemar’s test [74] following Zhang et al. (2021). For multi-choice MRC and NLI tasks, which are modeled as classification tasks, we performed statistical significance tests using paired t-test [76]. For NMT, multi-bleu.perl was used to compute case-sensitive 4-gram BLEU scores for NMT evaluations.<sup>8</sup> The signtest [77] is a standard statistical-significance test. All models were trained and evaluated on a single V100 GPU.

## 6 EXPERIMENTS

To focus on the evaluation of syntactic advance and keep simplicity, we only compare with single models instead of

8. <https://github.com/moses-smt/mosesdecoder/tree/RELEASE-4.0/scripts/generic/multi-bleu.perl>

TABLE 3

Accuracy (%) on RACE test set for single models. Our model is significantly better than the baseline BERT with p-value < 0.01.

Model	RACE-M	RACE-H	RACE
<i>Human Performance</i>			
Turkers	85.1	69.4	73.3
Ceiling	95.4	94.2	94.5
<i>Existing systems</i>			
GPT [2]	62.9	57.4	59.0
RSM [78]	69.2	61.5	63.8
BERT [79]	75.6	64.7	67.9
OCN [79]	76.7	69.6	71.7
DCMN [80]	77.6	70.1	72.3
XLNet[21]	85.5	80.2	81.8
ALBERT[23]	89.0	85.5	86.5
BERT Baseline	78.4	70.4	72.6
<b>SG-Net on BERT</b>	78.8	72.2	74.2
ALBERT Baseline	88.7	85.6	86.5
<b>SG-Net on ALBERT</b>	89.2	86.1	87.0

TABLE 4

Accuracy on the SNLI test set. Previous state-of-the-art models are marked by †. Our model is significantly better than the baseline BERT with p-value < 0.05.

Model	Acc
<i>Existing systems</i>	
GPT [2]	89.9
DRCN [81]	90.1
MT-DNN† [63]	91.6
SemBERT† [4]	91.6
<i>Our implementation</i>	
BERT Baseline	91.5
<b>SG-Net</b>	91.8
ALBERT Baseline	92.6
<b>SG-Net on ALBERT</b>	92.9

ensemble ones.

### 6.1 Reading Comprehension

Table 2 shows the result on SQuAD 2.0.<sup>9</sup> Various state of the art models from the official leaderboard are also listed for reference. We can see that the performance of BERT is very strong. However, our model is more powerful, boosting the BERT baseline essentially. It also outperforms all the published works and achieves the 2nd place on the leaderboard when submitting SG-NET. We also find that adding an extra answer verifier module could yield a better result. With the more powerful ALBERT, SG-Net still obtains improvements and also outperforms all the published works.

For RACE, we compare our model with the following latest baselines: Dual Co-Matching Network (DCMN) [80], Option Comparison Network (OCN) [79], Reading Strategies Model (RSM) [78], and Generative Pre-Training (GPT) [2]. Table 3 shows the result.<sup>10</sup> Turkers is the performance of Amazon Turkers on a random subset of the RACE test set. Ceiling is the percentage of unambiguous questions

9. Besides published works, we also list competing systems on the SQuAD leaderboard at the time of submitting SG-Net (May 14, 2019). Our model is significantly better than the baseline BERT with p-value < 0.01.

10. Our concatenation order of  $P$  and  $Q$  is slightly different from the original BERT. Therefore, the result of our BERT baseline is higher than the public one on the leaderboard, thus our improved BERT implementation is used as the stronger baseline for our evaluation.



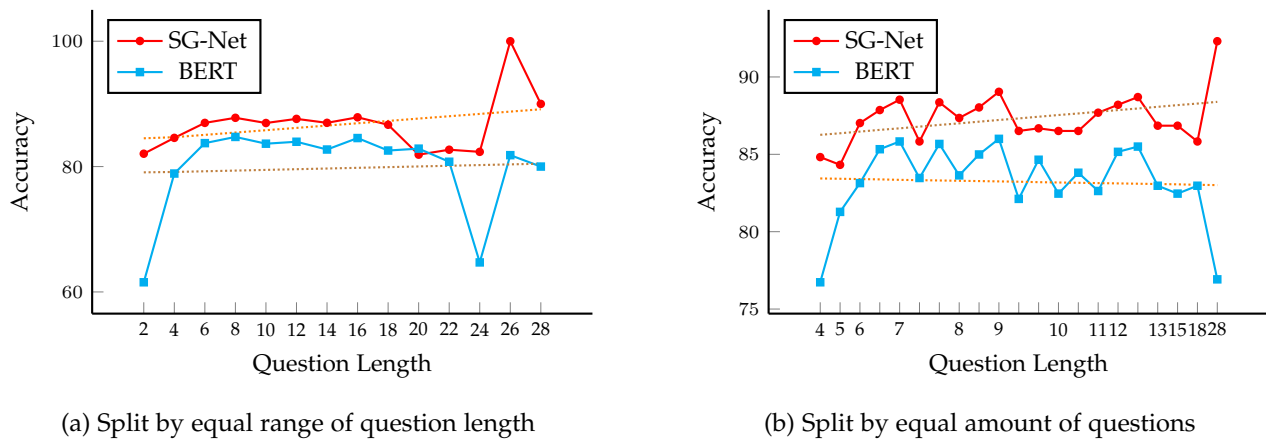


Fig. 4. Accuracy for different question length. Each data point means the accuracy for the questions in the same length range (a) or of the same number (b) and the horizontal axis in (b) shows that most of questions are of length 7-8 and 9-10.

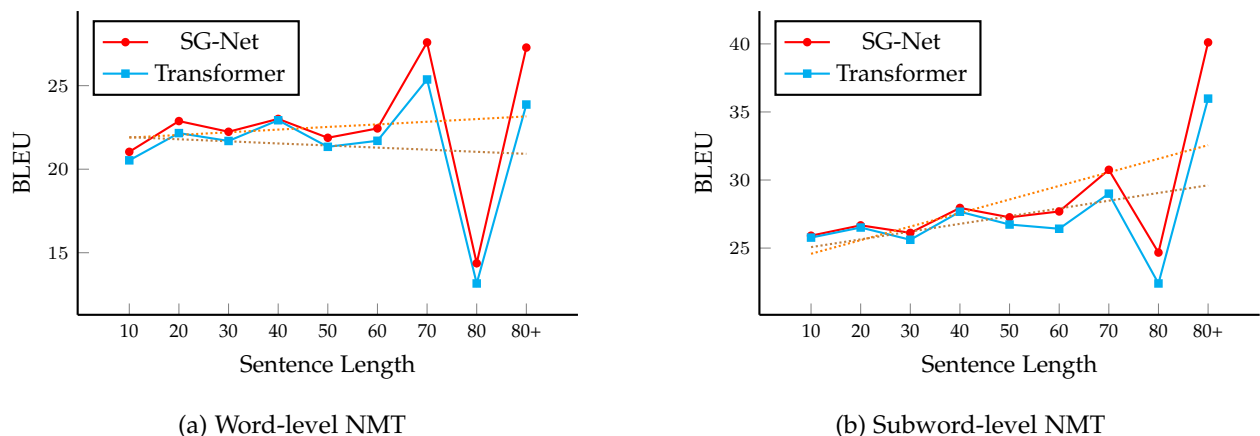


Fig. 5. Translation qualities of different sentence lengths. Each data point means the BLEU score in (a) word-level or (b) subword-level.

TABLE 5

BLEU scores on EN-DE dataset for the NMT tasks. “+/++” after the score indicates that the proposed method was significantly better than the baseline at significance level  $p < 0.05/0.01$ .

Model	Word	Subword
<i>Existing systems</i>		
Transformer [72]	-	27.3
SDAtt [50]	20.36	-
<i>Our implementation</i>		
Transformer	22.06	27.31
<b>SG-Net</b>	<b>23.02++</b>	<b>27.68+</b>

in the test set. From the comparison, we can observe that our model outperforms all baselines, which verifies the effectiveness of our proposed syntax enhancement.

## 6.2 Natural Language Inference

Table 4 shows SG-Net also boosts the baseline BERT, and achieves a new state-of-the-art on SNLI benchmark. The result shows that using our proposed method with induced syntactic structure information can also benefit natural language inference.

## 6.3 Neural Machine Translation

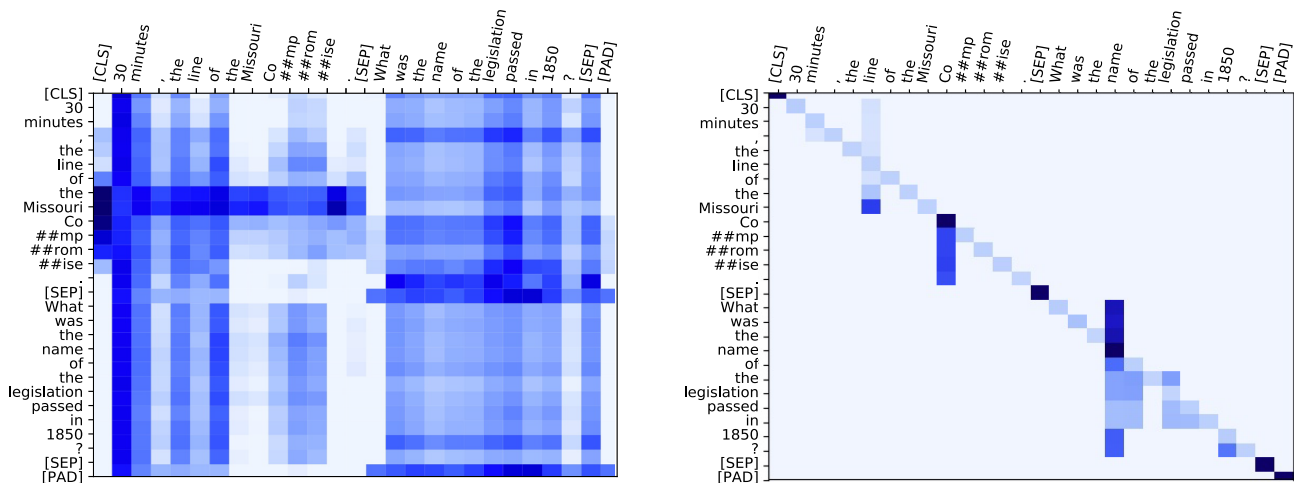
Table 5 shows the translation results. Our method significantly outperformed the baseline Transformer. The results showed that our method was not only useful for the NLU task but also more advanced translation tasks.

## 7 DISCUSSIONS

### 7.1 Effect of Answering Long Questions

Inspired by previous works that introducing syntactic structures might be beneficial for the modeling of long sentences [50], we are interested in whether the advance still holds when it comes to Transformer architectures. We grouped questions/sentences of similar lengths for both of the passage-level MRC and sentence-level NMT tasks, to investigate the model performance. For example, sentence length “10” indicates that the length of questions is between 8 and 10.

In detail, we sort the questions from the SQuAD 2.0 dev set according to the length and then group them into 20 subsets, split by equal length and equal amount of questions. Then we calculate the exact match accuracy of the baseline and SG-Net per group, as shown in Figure 4.



Passage (extract):...30 minutes, the line of the Missouri Compromise... Question:What was the name of the legislation passed in 1850? Answer:the Missouri Compromise

Fig. 6. Visualization of the vanilla BERT attention (left) and syntax-guided self-attention (right). Weights of attention are selected from first head of the last attention layer. For the syntax-guided self-attention, the columns with weights represent the SDOI for each word in the row. For example, the SDOI of *passed* contains {*name*, *of*, *legislation*, *passed*}. Weights are normalized by SoftMax for each row.

Since the number of questions varies in length, the gap could be biased by the distribution. We depict the two aspects to show the discovery comprehensively and we would like to clarify that the trends (dashed) in Figure 4 matter most, giving the intuition that the proposed one might not be as sensitive to long sentences as the baseline. We observe that the performance of the baseline drops heavily when encountered with long questions, especially for those longer than 20 words while our proposed SG-Net works robustly, even showing a positive correlation between accuracy and length. This shows that with syntax-enhanced representation, our model is better at dealing with lengthy questions compared with baseline.

For sentence-level NMT, we illustrate the comparison for both the word-level and subword-level NMT settings. According to the illustrations in Figure 5, we see similar trends with the passage-level MRC. We observe that our method can still boost the baseline substantially though the Transformer baseline has good ability of long-dependency modeling. We also find that SG-Net shows relatively greater advance for the passage-level task than that of sentence-level, in which there are always hundreds of words in an input sentence for passage-level MRC, which would require the stronger capacity of long-dependency, as well as structure modeling.

## 7.2 Visualization

To have an insight that how syntax-guided attention works, we draw attention distributions of the vanilla attention of the last layer of BERT and our proposed syntax-guided self-attention,<sup>11</sup> as shown in Figure 6. With the guidance of syntax, the keywords *name*, *legislation* and *1850* in the question

11. Since special symbols such as [PAD] and [CLS] are not considered in the dependency parsing tree, we confine the SDOI of these tokens to themselves. So these special tokens will have a value of 1 as weights over themselves in syntax-guided self-attention, and we will mask these weights in the following aggregation layer.

TABLE 6  
Ablation study on potential components and aggregation methods on SQuAD 2.0 dev set.

Model	EM	F1
baseline	84.1	86.8
+ Vanilla attention only	84.2	86.9
+ Syntax-guided attention only	84.4	87.2
+ Dual contextual attention	<b>85.1</b>	<b>87.9</b>
- Concatenation	84.5	87.6
Bi-attention	84.9	87.8

are highlighted, and *(the) Missouri*, and *Compromise* in the passage are also paid great attention, which is exactly the right answer. The visualization verifies that benefiting from the syntax-guided attention layer; our model is effective at selecting the vital parts, guiding the downstream layer to collect more relevant pieces to make predictions.

## 7.3 Dual Context Mechanism Evaluation

In SG-Net, we integrate the representations from syntax-guided attention layer and the vanilla self-attention layer in dual context layer. To unveil the contribution of each potential component, we conduct comparisons on the baseline with:

- 1) *Vanilla attention only* that adds an extra vanilla BERT attention layer after the BERT output.
- 2) *Syntax-guided attention only* that adds an extra syntax-guided layer after the BERT output.
- 3) *Dual contextual attention* that is finally adopted in SG-Net as described in Section 3.2.

Table 6 shows the results. We observe that dual contextual attention yields the best performance. Adding extra vanilla attention gives no advance, indicating that introducing more parameters would not promote the strong baseline. It is reasonable that syntax-guided attention only is also trivial since it only considers the syntax related parts

TABLE 7  
Eight probing tasks [82] to study what kind of properties are captured by the encoders.

Probing Tasks	Content	
Syntactic	TrDep	Checking whether an encoder infers the hierarchical structure of sentence
	ToCo	Sentences should be classified in terms of the sequence of top constituents immediately below the sentence node
	BShif	Testing whether two consecutive tokens within the sentence have been inverted
Semantic	Tense	Asking for the tense of the main clause verb
	SubN	Focusing on the number of the main clause’s subject
	ObjN	Testing for the number of the direct object of the main clause
	SoMo	Some sentences are modified by replacing a random noun or verb with another one and the classifier should tell whether a sentence has been modified
	CoIn	Containing sentences made of two coordinate clauses

TABLE 8  
Accuracy on eight probing tasks of evaluating linguistics embedded in the encoder outputs. “+”/“+” after the accuracy score indicate that the score is statistically significant at level  $p < 0.01/0.05$ .

Model	Syntactic			Semantic				
	TrDep	ToCo	BShif	Tense	SubN	ObjN	SoMo	CoIn
Baseline	28.34	58.33	76.34	80.66	77.28	76.32	64.42	67.51
SG-Net	30.02++	60.42++	76.63	80.52	77.72+	76.83+	64.38	67.23

TABLE 9  
The comparison of answers from baseline and our model. In these examples, answers from SG-Net are the same as the ground truth.

Question	Baseline	SG-Net
When did Herve serve as a Byzantine general?	105	1050s
What is Sky+ HD material broadcast using?	MP	MPEG-4
A problem set that that is hard for the expression NP can also be stated how?	the set of NP-hard problems	NP-hard

when calculating the attention, which is complementary to traditional attention mechanism with noisy but more diverse information and finally motivates the design of dual contextual layer.

Actually, there are other operations for merging representations in dual context layer besides the weighted dual aggregation, such as *concatenation* and *Bi-attention* [83], which are also involved in our comparison, and our experiments show that using dual contextual attention produces the best result.

## 7.4 Linguistic Analysis

We are interested in what kind of knowledge learned in the universal representations. In this section, we selected eight widely-used language probing tasks [82] (see Table 7) to study what syntactic and semantic properties are captured by the encoders. Specifically, we use the BERT and ALBERT models trained on the SNLI task to generate the sentence representation (embedding) of input to evaluate what linguistic properties are encoded in it. The results are shown in Table 8. Regarding the syntactic properties, our model achieves statistically significant improvement on the *TrDep* and *ToCo* tasks, which demonstrates that the syntax-guided model could help identify and represent the sentence structure and constituents, which discloses that SG-Net might learn better syntax-aware information with the syntax-constrained SAN design.

## 7.5 Model Prediction

To have an intuitive observation of the predictions of SG-Net, we show a list of prediction examples on SQuAD 2.0

from baseline BERT and SG-Net in Table 9. The comparison indicates that our model could extract more syntactically accurate answer, yielding more exact match answers while those from the baseline BERT model are often semantically incomplete. This shows that utilizing explicit syntax is potential to guide the model to produce meaningful predictions. Intuitively, the advance would attribute to better awareness of syntactic dependency of words, which guides the model to learn the syntax relationships explicitly.

## 7.6 Influence of Parsing Performance

Since our SDOI mask is derived from the syntactic parsing, the parsing performance would influence the overall model performance. To investigate the influence of the accuracy of the parser, we first use a relatively weaker parser with 95.62% UAS and 94.10% LAS accuracy, using other suboptimal hyper-parameters. We use the NLI task on SNLI for benchmark, the accuracy is 91.72% compared with the best result of 91.81% reported in Table 4. To evaluate in more extreme cases, we then degrade our parser by randomly turning specific proportion [0, 20%, 40%, 60%, 80%, 100%] of labels into random error ones as cascading errors. The SNLI accuracies are respectively [91.81%, 91.68%, 91.56%, 91.42%, 91.41%, 91.38%]. This stable performance can be attributed to dual attention design, i.e., the concatenation operation of BERT hidden states and pruned syntax-guided layer representation, in which the downgraded parsing representation (even noisy) would not affect the former one intensely. This result indicates that the LM can not only benefit from high-accuracy parser but also keep robust against noisy labels.

## 7.7 Model Efficiency

There are two aspects that contribute to the overall model computation: 1) preprocessing: the syntactic parsing on our task datasets; and 2) model training: training SG-Net for our tasks. Our calculation analysis is based on one 32G V100 GPU. For the processing, the annotation speed is rapid, which is around 10.3 batches (batch size=256 sentences, max sequence length=256) per second. For model training, the number of extra trainable parameters in our overall model is very small (only +12M, compared with 335M of BERT, for example) as we only add one extra attention layer. The dual aggregation is the tensor calculation on the existing outputs of SG layer and BERT, without any additional parameter. The training computation speeds of SG-Net and the baseline are basically the same (1.18 vs. 1.17 batches per second, batch size=32, max sequence length=128).

## 8 DISCUSSION OF PROS AND CONS

There are a variety of methods proposed to introduce tree or graph-based syntax into RNN-based NLP models. The main idea of the traditional solutions is transforming the data form by either 1) applying a tree-based network [84], or 2) linearizing syntax [85, 86], which converts the relationship of the nodes in dependency tree into discrete labels and integrates the label-formed features with word embeddings. The difference lies with the efficiency and the processing part in the neural models. Among the two methods, the advantage of the former is that it can preserve the original structured characteristics of the data, i.e., the attributes of words and relationships between different words, and the tree network can fully display the tree structure in the network. In contrast, the latter’s benefits are simplicity and high efficiency, where the syntactic features can be easily integrated with existing word embeddings. In terms of the processing part, the former will perform in the sentence-level encoding phase, while the latter focuses on the fundamental embedding layer before contextual encoding.

For the recent Transformer architecture, the self-attention mechanism ensures that the model can fully obtain the dependency between each token of the input sequence, which goes beyond the sequential processing of each successive token. Such a relationship is richer and more diverse than RNN models. However, Transformer could not score the relationship between tokens explicitly. Although the relationship might be obtained to some extent after training, the adjustment could be chaotic, redundant and unexplained. Syntactic information, which can reflect the relationship between tokens, is obviously a natural solution as a kind of guidance. Compared with the traditional methods, SG-Net has the following advantages:

1. High precision. The intermediate product of the attention matrix generated by the self-attention can be regarded as a set of relations among each couple of tokens in size of  $n \times n$  where  $n$  denotes the sequence length, that is, a complete graph as well. SG-Net can well save the tree or graph form of features to the greatest extent, and it is intuitive and interpretable.

2. High efficiency. Compared with the traditional processing of tree-based networks, SG-Net keeps the parallelism of Transformer with minor revisions in the archi-

ture. Besides, the data processing is finished before the model training, instead of putting the data processing in the model training as in traditional methods.

3. Light-weight. SG-Net makes full use of the structure of Transformer, and the change has little damage to Transformer and can be fully integrated into Transformer.

However, there still exist some disadvantages in theory. First, SG-Net can only be used in Transformer architectures that rely on self-attention mechanisms. Second, the type of relationship is considered. Although it can reflect the relationship between tokens, it can not reflect what kind of relationship it is. Third, compared with the traditional method, SG-Net requires high-precision syntax, and the wrong syntax will cause great damage. The reason is that this method directly modifies the strength of the relationship between tokens, which can suffer from error propagation. However, as discussed in Section 7.6, our dual attention design can well alleviate such a situation when using a weak parser and ensure the model focus more on the original output.

## 9 CONCLUSION

This paper presents a novel syntax-guided framework for enhancing strong Transformer-based encoders. We explore to adopt syntax to guide the text modeling by incorporating syntactic constraints into attention mechanism for better linguistically motivated word representations. Thus, we adopt a dual contextual architecture called syntax-guided network (SG-Net) which fuses both the original SAN representations and syntax-guided SAN representations. Taking pre-trained BERT as our Transformer encoder implementation, experiments on three typical advanced natural language understanding tasks, including machine reading comprehension, natural language inference, and neural machine translation show that our model can yield significant improvements in all the challenging tasks. This work empirically discloses the effectiveness of syntactic structural information for text modeling. The proposed attention mechanism also verifies the practicability of using linguistic information to guide attention learning and can be easily adapted with other tree-structured annotations.

For the application, our method is lightweight and easy to cooperate with other neural models. Without the need for architecture modifications in existing models, we can easily apply syntax-guided constraints in an encoder by adding an extra layer fed by the output representation of the encoder and the SDOI mask, which makes it easy enough to be applicable to other systems. The design is especially important for recent dominant pre-trained Transformer-based language representation models, such as BERT [3], XLNet [21], RoBERTa [22], ALBERT [23], etc. As a plugin, we do not need to train the models from scratch but directly fine-tune for downstream tasks.

Incorporating human expertise and knowledge to machine is a key inspiration of AI researches, which stimulates lots of studies that consider involving linguistic information to neural models in NLP scenarios. In this work, we discovered an effective masking strategy to incorporate extra structured knowledge into the state-of-the-art Transformer architectures. Besides syntactic information, this method is also compatible with a variety of structured knowledge as

explicit sentence-level constraints to improve the representation ability of Transformer. For example, it is potential to model the concept relationships by introducing extra structured knowledge graphs, including ConceptNet [87], DBpedia [88]. We hope this work can facilitate related research and shed lights on future studies of knowledge aggregation in the community.

## REFERENCES

- [1] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, "SG-Net: Syntax-guided machine reading comprehension," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, 2020.
- [2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," URL [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language_understanding_paper.pdf), 2018.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [4] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, 2020.
- [5] J. Zhou, Z. Zhang, H. Zhao, and S. Zhang, "LIMIT-BERT : Linguistics informed multi-task BERT," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4450–4461.
- [6] Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li, and H. Zhao, "Neural machine translation with universal visual representation," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=Byl8hhNYPS>
- [7] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 784–789, 2018.
- [8] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 632–642, 2015.
- [9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *The 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5998–6008.
- [11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pp. 2227–2237, 2018.
- [12] S. Wang, J. Zhang, and C. Zong, "Learning sentence representation with guidance of human attention," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 2017, pp. 4137–4143.
- [13] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, "Did the model understand the question?" *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 1896–1906, 2018.
- [14] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 2383–2392, 2016.
- [15] Z. Zhang and H. Zhao, "One-shot learning for question-answering in gaokao history challenge," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 449–461.
- [16] Z. Zhang, Y. Huang, and H. Zhao, "Subword-augmented embedding for cloze reading comprehension," in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2018, pp. 1802–1814.
- [17] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu, "Modeling multi-turn conversation with deep utterance aggregation," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3740–3752.
- [18] A. W. Yu, H. Lee, and Q. Le, "Learning to skim text," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, 2017, pp. 1880–1890.
- [19] Z. Li, R. Wang, K. Chen, M. Utiyama, E. Sumita, Z. Zhang, and H. Zhao, "Explicit sentence compression for neural machine translation," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, 2020.
- [20] Y. Shen, Z. Lin, C. wei Huang, and A. Courville, "Neural language modeling by jointly learning syntax and lexicon," in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=rkgOLb-0W>
- [21] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5754–5764.
- [22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.
- [24] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp.

- 467–479, 1992.
- [25] R. K. Ando and T. Zhang, “A framework for learning predictive structures from multiple tasks and unlabeled data,” *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.
- [26] J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with structural correspondence learning,” in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 120–128.
- [27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems (NIPS 2013)*, 2013, pp. 3111–3119.
- [28] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014)*, 2014, pp. 1532–1543.
- [29] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP 2013)*, 2013, pp. 1631–1642.
- [30] E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the conll-2003 shared task: language-independent named entity recognition,” in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 142–147.
- [31] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 353–355.
- [32] N. Nangia, A. Williams, A. Lazaridou, and S. R. Bowman, “The repeval 2017 shared task: Multi-genre natural language inference with sentence representations,” in *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, 2017, pp. 1–10.
- [33] J. Nivre, “Inductive dependency parsing of natural language text,” Ph.D. dissertation, School of Mathematics and Systems Engineering, Växjö University, 2005.
- [34] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of english: The penn treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.
- [35] A. Ivanova, S. Oepen, and L. Øvrelid, “Survey on parsing three dependency representations for english,” in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 2013, pp. 31–37.
- [36] C. Pollard and I. Sag, “Head-driven phrase structure grammar. university of chicago, press,” *Chicago, IL*, 1994.
- [37] S. Kübler, R. McDonald, and J. Nivre, “Dependency parsing,” *Synthesis lectures on human language technologies*, vol. 1, no. 1, pp. 1–127, 2009.
- [38] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, “Transition-based dependency parsing with stack long short-term memory,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 334–343.
- [39] Y. Zhang and J. Nivre, “Transition-based dependency parsing with rich non-local features,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 188–193.
- [40] W. Wang and B. Chang, “Graph-based dependency parsing with bidirectional lstm,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2306–2315.
- [41] Z. Zhang, H. Zhao, and L. Qin, “Probabilistic graph-based dependency parsing with convolutional neural network,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1382–1392.
- [42] P. Blunsom and T. Cohn, “Unsupervised induction of tree substitution grammars for dependency parsing,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1204–1213.
- [43] E. Metheniti, P. Park, K. Kolesova, and G. Neumann, “Identifying grammar rules for language education with dependency parsing in german,” in *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 2019, pp. 100–111.
- [44] Z. Zhang, H. Zhao, and L. Qin, “Probabilistic graph-based dependency parsing with convolutional neural network,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1382–1392.
- [45] Z. Li, J. Cai, S. He, and H. Zhao, “Seq2seq dependency parsing,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3203–3214.
- [46] X. Ma, Z. Hu, J. Liu, N. Peng, G. Neubig, and E. Hovy, “Stack-Pointer Networks for Dependency Parsing,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1403–1414.
- [47] Z. Li, H. Zhao, and K. Parnow, “Global greedy dependency parsing,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8319–8326.
- [48] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao, “Neural machine translation with source dependency representation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2846–2852.
- [49] K. Chen, T. Zhao, M. Yang, L. Liu, A. Tamura, R. Wang, M. Utiyama, and E. Sumita, “A neural approach to source dependence based context model for statistical machine translation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 266–280, 2017.
- [50] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, “Syntax-directed attention for neural machine translation,” in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018, pp. 4792–4799.
- [51] Y. Wang, H.-Y. Lee, and Y.-N. Chen, “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)*, 2019, pp. 1060–1070.
- [52] S. Duan, H. Zhao, J. Zhou, and R. Wang, “Syntax-aware transformer encoder for neural machine translation,” in *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019, pp. 396–401.
- [53] S. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts, “A fast unified model for parsing and sentence understanding,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1466–1477.
- [54] J. Kasai, D. Friedman, R. Frank, D. Radev, and O. Rambow, “Syntax-aware neural semantic role labeling with supertags,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, 2019, pp. 701–709.
- [55] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, “Linguistically-informed self-attention for semantic role labeling,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5027–5038.
- [56] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does bert look at? an analysis of bert’s attention,” in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 276–286.
- [57] W. Wu, H. Wang, T. Liu, and S. Ma, “Phrase-level self-attention networks for universal sentence encoding,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3729–3738.
- [58] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [59] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, “Race: Large-scale reading comprehension dataset from examinations,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 785–794.
- [60] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, “Natural language inference by tree-based convolution and heuristic matching,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, p. 130–136, 2016.
- [61] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, “Recurrent neural network-based sentence encoder with gated attention for natural language inference,” in *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, 2017, pp. 36–40.
- [62] R. Ghaeini, S. A. Hasan, V. Datla, J. Liu, K. Lee, A. Qadir, Y. Ling, A. Prakash, X. Fern, and O. Farri, “Dr-bilstm: Dependent reading bidirectional lstm for natural language inference,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1460–1469.
- [63] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” *arXiv preprint arXiv:1901.11504*, 2019.
- [64] Z. Zhang, Y. Wu, Z. Li, and H. Zhao, “Explicit contextual semantics for text comprehension,” in *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)*, 2019.
- [65] J. Zhou and H. Zhao, “Head-driven phrase structure grammar parsing on penn treebank,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2396–2408.
- [66] N. Kitaev and D. Klein, “Constituency parsing with a self-attentive encoder,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2676–2686.
- [67] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a Large Annotated Corpus of English: The Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, 1993.
- [68] X. Wang, G. Xu, J. Zhang, X. Sun, L. Wang, and T. Huang, “Syntax-directed hybrid attention network for aspect-level sentiment analysis,” *IEEE Access*, vol. 7, pp. 5014–5025, 2018.
- [69] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.
- [70] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [71] Z. Zhang, H. Zhao, K. Ling, J. Li, Z. Li, S. He, and G. Fu, “Effective subword segmentation for text comprehension,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1664–1674, 2019.
- [72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.
- [73] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [74] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [75] Z. Zhang, J. Yang, and H. Zhao, “Retrospective reader for machine reading comprehension,” in *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.
- [76] H. Hsu and P. A. Lachenbruch, “Paired t test,” *Encyclopedia of Biostatistics*, vol. 6, 2005.
- [77] M. Collins, P. Koehn, and I. Kucerova, “Clause restructuring for statistical machine translation,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 2005, pp. 531–540.
- [78] K. Sun, D. Yu, D. Yu, and C. Cardie, “Improving machine reading comprehension with general reading strategies,” 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)*, 2019, pp. 2633–2643.

- [79] Q. Ran, P. Li, W. Hu, and J. Zhou, "Option comparison network for multiple-choice reading comprehension," *arXiv preprint arXiv:1903.03033*, 2019.
- [80] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, and X. Zhou, "Dcmn+: Dual co-matching network for multi-choice reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9563–9570.
- [81] S. Kim, I. Kang, and N. Kwak, "Semantic sentence matching with densely-connected recurrent and co-attentive information," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 6586–6593.
- [82] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single  $\&\#\&\#^*$  vector: Probing sentence embeddings for linguistic properties," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2126–2136.
- [83] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *International Conference on Learning Representations*, 2016.
- [84] C. Ma, A. Tamura, M. Utiyama, T. Zhao, and E. Sumita, "Forest-based neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1253–1263.
- [85] J. Li, D. Xiong, Z. Tu, M. Zhu, M. Zhang, and G. Zhou, "Modeling source syntax for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 688–697.
- [86] A. Eriguchi, K. Hashimoto, and Y. Tsuruoka, "Tree-to-sequence attentional neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 823–833.
- [87] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: an open multilingual graph of general knowledge," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4444–4451.
- [88] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.



**Zhusheng Zhang** received his Bachelor's degree in internet of things from Wuhan University in 2016, his M.S. degree in computer science from Shanghai Jiao Tong University in 2020. He is working towards the Ph.D. degree in computer science with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University. He was an internship research fellow at NICT from 2019-2020. His research interests include natural language processing, machine reading comprehension, dialogue systems, and

machine translation.



**Yuwei Wu** received his Bachelor's degree in Computer Science from Shanghai Jiao Tong University in 2020, where he was a member of ACM Class, part of Zhiyuan College in SJTU. He is working towards the Ph.D. degree in Department of Computer Science, Shanghai Jiao Tong University. During his graduate study, he took research internships at Georgia Tech in 2019. His research interests lie in natural language processing, machine reading comprehension, dialogue systems, and multimodal.



**Junru Zhou** received the B.S. degree from South China University of Technology, in 2018. Since then, he has been a master student in Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research focuses on natural language processing, especially in syntactic and semantic parsing, pre-training language model.



**Sufeng Duan** received his Bachelor degree in Spatial Information and Digital Technology from Wuhan University and Master degree in Computer Science and Technology from Shanghai Jiao Tong University in 2014 and 2017. Since 2018, he has been a Ph.D. student with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University, Shanghai, China. His research focuses on natural language processing, especially Chinese word segmentation and machine translation.



**Hai Zhao** received the BEng degree in sensor and instrument engineering, and the MPhil degree in control theory and engineering from Yanshan University in 1999 and 2000, respectively, and the PhD degree in computer science from Shanghai Jiao Tong University, China in 2005. He is currently a full professor at department of computer science and engineering, Shanghai Jiao Tong University after he joined the university in 2009. He was a research fellow at the City University of Hong Kong from 2006 to 2009, a visiting scholar in Microsoft Research Asia in 2011, a visiting expert in NICT, Japan in 2012. He is an ACM professional member, and served as area co-chair in ACL 2017 on Tagging, Chunking, Syntax and Parsing, (senior) area chairs in ACL 2018, 2019 on Phonology, Morphology and Word Segmentation. His research interests include natural language processing and related machine learning, data mining and artificial intelligence.



**Rui Wang** is an associate professor at Shanghai Jiao Tong University since 2021. Before that, he was a researcher (tenured in 2020) at Japan National Institute of Information and Communications Technology (NICT) from 2016 to 2020. He received his B.S. degree from Harbin Institute of Technology in 2009, his M.S. degree from the Chinese Academy of Sciences in 2012, and his Ph.D. degree from Shanghai Jiao Tong University in 2016, all of which are in computer science. He was a joint Ph.D. at Centre National de

la Recherche Scientifique, France in 2014. His research interests are machine translation and natural language processing.