Article

Learning Machine Translation with Linguistic Interpretation

Kayal Alkmim * , Rodolfo Patel , Dolcetti Dave

Posted Date: 13 December 2023

doi: 10.20944/preprints202312.0921.v1

Keywords: Machine Translation; Linguistic Interpretation; Attention Models

Article

Learning Machine Translation with Linguistic Interpretation

Kayal Alkmim *, Rodolfo Patel and Dolcetti Dave

Briar Cliff University; rodolfopatel5@gmail.com; dave@briarcliff.edu
* Correspondence: alkmim@briarcliff.edu

Abstract: The Transformer architecture, while adept at capturing context through self-attention, falls short in encapsulating complex syntactic structures effectively. Addressing this gap, we introduce the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) approach in Machine Translation (MT) frameworks. Combining the strengths of Graph Attention Network (GAT) and BERT, LSGIB intricately captures syntactic dependencies as explicit knowledge from the source language. This enhances the source language representation and aids in more accurate target language generation. Our empirical analysis leverages gold-standard syntax-annotated sentences and employs a Quality Estimation (QE) model. This approach enables us to assess translation improvements in terms of syntactic accuracy, extending beyond traditional BLEU score metrics. The LSGIB model demonstrates superior translation quality across diverse MT tasks, maintaining robust BLEU scores. Our study delves into the optimal sentence lengths benefiting from LSGIB and identifies which syntactic dependencies are more precisely captured. We observe that GAT's ability to learn specific dependency relations directly influences the translation quality of sentences with those relations. Additionally, we discover that incorporating syntactic structure into BERT's intermediate and lower layers offers a novel approach to modeling linguistic structure in source sentences.

Keywords: machine translation; linguistic interpretation; attention models

1. Introduction

Neural Machine Translation (NMT) has significantly evolved, offering more fluent and coherent translations than its statistical predecessors. However, despite these advancements, NMT systems often grapple with syntactic complexities, leading to translations that are syntactically inconsistent. This is particularly evident in scenarios involving limited bilingual training resources. The Transformer model, a notable development in this field, employs a self-attention mechanism that, while effective in context capturing, still struggles with intricate syntactic nuances [1,2]. This limitation is not just a technical shortfall but also a barrier to achieving truly natural and accurate machine translations.

BERT, introduced by [3], builds upon the Transformer model, but with a significant enhancement - pre-training. This process involves unsupervised learning on a vast corpus, equipping BERT with an extensive understanding of language nuances. The model not only preserves the structural strengths of the Transformer but also brings in rich, implicit linguistic knowledge, making it a valuable asset in Machine Translation (MT) tasks [2,4,5]. BERT's ability to capture deep linguistic features has opened new avenues in the realm of NMT, allowing for more accurate and contextually rich translations.

The role of explicit linguistic knowledge, such as syntax, is paramount in refining NMT outputs. Traditional linear models like RNNs and even the Transformer, while capable of processing syntactic information to an extent, are inherently limited in their representation of non-linear syntactic structures and relationships [8,9]. This limitation often leads to translations that, although grammatically correct, miss the subtleties of linguistic structure and meaning.

Enter the Graph Attention Network (GAT) [10], a novel approach that represents syntactic structures and relationships more explicitly through a graph-based topology. GAT's ability to encapsulate complex inter-word dependencies in a non-linear manner offers a more accurate representation of syntactic phenomena. This explicit representation not only enhances the model's performance but also improves its interpretability, a crucial aspect in understanding and improving NMT systems [11,12]. The integration of GAT into NMT poses a compelling question: Can the fusion of explicit syntactic knowledge via GAT with the deep, implicit linguistic understanding of BERT lead to a significant leap in translation quality?

To address this, we introduce the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) model. LSGIB is an innovative approach that synergizes the syntactic data processing capabilities of GAT with the deep linguistic comprehension of BERT. This combination aims to tackle the inherent limitations of current Transformer-based NMT systems by integrating both syntactic structure and deep learning insights. LSGIB leverages multi-head attention mechanisms on graphs to utilize source-side syntactic dependencies explicitly. This not only enhances the source language representation but also provides a more informed basis for the target-side decoder.

Our research methodology involves comprehensive experiments across translation tasks from Chinese (Zh), German (De), and Russian (Ru) to English (En). These tasks are designed to rigorously test and demonstrate the effectiveness of the LSGIB approach. Our primary contributions through this research are manifold:

- LSGIB marks a pioneering effort in demonstrating the efficacy of combining graph-based syntactic knowledge with BERT in MT tasks. This model circumvents the need for training from scratch by being fine-tuned for specific applications.
- We conduct a thorough evaluation of the translation quality, focusing on syntactic accuracy and Quality Estimation (QE) scores. Our findings show that LSGIB not only maintains robust BLEU scores but also enhances translation quality, particularly for shorter and medium-length source sentences. We also identify specific syntactic dependencies that are more effectively captured by the model, leading to improved translations.
- Our study delves deep into the interpretability of translation quality improvements, particularly from a syntactic knowledge standpoint. The learning and representation of syntactic relationships through GAT, coupled with the deep linguistic processing capabilities of BERT, lead to a novel approach in modeling source sentences. This synergy results in significant enhancements in translation quality, attributable to both the explicit syntactic knowledge on the graph and the nuanced features reconstructed by BERT.

2. Related Work

The emergence of pre-trained models has revolutionized the field of Natural Language Processing (NLP), with the Transformer architecture being a cornerstone for many of these advancements [3,17–19]. Among these, BERT stands out as a seminal pre-trained model that leverages two innovative pre-training objectives: the Masked Language Model (MLM) and the Next Sentence Prediction (NSP). The MLM approach involves predicting masked words in a sentence using contextual clues, while NSP assesses whether two sentences are sequential. These objectives enable BERT to assimilate a vast array of implicit linguistic knowledge, which can be fine-tuned for various downstream tasks. Recognizing BERT's linguistic prowess, researchers have explored its integration into Neural Machine Translation (NMT) as both an encoder and decoder to enhance sentence modeling capabilities and overall translation performance. Notably, [20] utilized BERT directly as an encoder in MT systems, employing a two-stage optimization process that showed promise in low-resource language learning. [22] focused on mitigating catastrophic forgetting in MT tasks through a concerted training framework. Additionally, [4] fused BERT's output features with the encoder and decoder of the MT model using an attention module, enabling the model to fully exploit BERT's knowledge for adaptive learning.

In MT, the role of syntactic dependency is pivotal, as it aids in the analysis of grammatical structures and their representation in an intuitive tree format. This explicit structural information reduces sentence ambiguity and enhances the MT model's understanding of sentence context. Various studies have underscored the value of incorporating syntactic information into NMT. For instance, [25] explored the linearization and injection of syntactic information from the source language into the

Transformer model, examining its impact on low-resource translation tasks. [26] integrated specific syntactic dependencies into the attention mechanism, combined with the Transformer model, to achieve a linguistics-inspired representation. [27] introduced various masks to guide the attention mechanisms based on syntactic knowledge, allowing attention heads to select and learn from multiple syntactic patterns. However, most approaches to modeling syntactic information have been linear, with a lack of comprehensive exploration into topological representations of syntactic knowledge. Moreover, the integration of syntactic information in Transformer models and scenarios involving BERT in MT models remains an underexplored area.

Graph neural networks represent another frontier in feature integration, where nodes symbolize words in a sentence, and edges delineate the connections between these words. The pre-definition of a graph's structure for a sentence, which combines prior knowledge and explicit features, is crucial for designing an effective graph neural network. Recently, the Graph Attention Network (GAT) has been proposed as a potent tool for representing data in non-Euclidean spaces. GAT combines an attention mechanism to allocate varying weights to nodes on the graph, independent of the network's specific structure. With its capability for learning on graphs and supporting a multi-headed attention mechanism, GAT has been increasingly used in conjunction with BERT to represent linguistic knowledge in downstream tasks [11,28–33]. Despite these advancements, most studies have singularly focused on syntactic knowledge and BERT in MT scenarios, leaving a gap in understanding how the integration of explicit syntactic knowledge via GAT and BERT can enhance translation quality. Furthermore, there is a need for more interpretability from a linguistic perspective to elucidate the changes in translation quality brought about by these integrations. In this paper, we propose the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) model. LSGIB seeks to bridge these gaps by combining the explicit syntactic knowledge representation capabilities of GAT with the deep linguistic comprehension of BERT. This innovative approach aims to redefine the paradigms of syntactic modeling in NMT, providing a more nuanced and interpretable framework for understanding and improving translation quality.

3. Methodology

This section elaborates on the architecture of the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) model. The LSGIB model is composed of several layers: the encoding layer, the graph attention layer, and the fusion and output layer, each playing a crucial role in the overall translation process.

3.1. Encoding

The LSGIB model is tested on translations from three source languages to English: Chinese to English (Zh→En), Russian to English (Ru→En), and German to English (De→En). For a given source sentence $S = [w_1, w_2, w_3, \dots w_i]$, where i is the number of tokens, the sentence is first tokenized into subwords and processed by BERT: $\tilde{S} = [[CLS], w_1^1, w_1^{1\#1}, w_2, w_3^3, w_3^{3\#3}, \dots w_n, [SEP]]$. Here, $w^{n\#n}$ denotes subwords of w_n, and [CLS] and [SEP] are BERT's special tokens.

For each language, a specific BERT variant is utilized as the encoder: Chinese is represented by chinese-bert-wwm-ext[1], Russian by rubert-base[2], and German by bert-base-german[3]. These models, while structurally similar, differ in their pre-training methodologies, offering unique insights into each language's linguistic nuances.

The embedding sequence generated from the last layer of BERT, $h_B = BERT(\tilde{S})$, encapsulates the representation of each subword token. To extract syntactic dependency information from the source

[1] https://huggingface.co/hfl/chinese-bert-wwm-ext
[2] https://huggingface.co/DeepPavlov/rubert-base-cased
[3] https://huggingface.co/bert-base-german-cased

sentence $\tilde{S}$, a Universal Dependencies-based parser[4] is employed for tokenizing and parsing. The parsing results facilitate the construction of a node adjacency matrix, representing the graph structure of the sentence. Each token corresponds to a node on the graph, with node embeddings derived from BERT's word representations. Considering subword segmentation, these embeddings are averaged to represent each node.

3.2. Graph Attention

The sentence structure, encompassing words and their syntactic relationships, is modeled as a graph. In this graph, nodes represent words, and edges, the syntactic dependencies between them. The Graph Attention Network (GAT) [10] is employed to integrate this graph-structured data with node features. The node features input to a GAT layer are $\tilde{S} = [x_1, x_2, \ldots x_i, \ldots x_n], x_i \in \mathbb{R}^F$, where n is the total number of nodes and F the feature size for each node. The GAT's functioning is summarized in Equations (1) and (2):

$$h_i^{out} = \overset{K}{\underset{k=1}{\Vert}} \sigma\left(\sum_{j \in N_i} \alpha_{ij}^k W^k x_j\right) \tag{1}$$

$$\alpha_{ij}^k = \frac{exp(LeakyReLU(a^T[Wx_i \Vert Wx_j]))}{\sum_{v \in N_i} exp(LeakyReLU(a^T[Wx_i \Vert Wx_v]))} \tag{2}$$

where $j \in N_i$ are the 1-hop neighbors attended by node i, $\overset{K}{\underset{k=1}{\Vert}}$ denotes the concatenation of K multi-head attention outputs, and h_i^{out} is the representation of node i in the given layer. α_{ij}^k is the attention coefficient, W^k a linear transformation matrix, and a the weight vector for attention calculation. $LeakyReLU$ serves as the activation function. Simplified, the GAT layer's feature computation is $h_G = GAT(X, A; \Theta^l)$, with $X \in \mathbb{R}^{n \times F}$ being the input, $h_G \in \mathbb{R}^{n \times F'}$ the output, $A \in \mathbb{R}^{n \times n}$ the adjacency matrix, and Θ^l the trainable parameters.

3.3. Fusion and Output

Two approaches are proposed for integrating syntactic knowledge in the LSGIB model. The first, termed LSGIB Concatenation (LSGIBC), combines the graph's syntactic knowledge with BERT for the encoder, as delineated in Equations (3) and (4):

$$H_e^l = concat(h_B, h_G) \tag{3}$$

$$\tilde{h}_d^l = attn_D(h_d^l, H_e^l, H_e^l) \tag{4}$$

Here, $attn_D$ represents the encoder-decoder attention in MT engines, with l being the l-th layer's output and d the decoder-side token representation. H_e^l incorporates features from BERT (h_B) and GAT (h_G) and feeds them into the encoder-decoder attention module in the decoder. The attention features are then processed by a feed-forward network with a residual connection, similar to the vanilla Transformer model.

The second approach, LSGIB with Decoder (LSGIBD), applies syntactic knowledge from the graph not only to the encoder but also guides the decoder through syntax-decoder attention, as shown in Equations (5), (6), and (7):

$$\tilde{h}_d^l = attn_D(h_d^l, H_e^l, H_e^l) \tag{5}$$

$$\tilde{h}_s^l = attn_S(h_d^l, h_g^l, h_g^l) \tag{6}$$

$$\tilde{h}_t^l = concat(\tilde{h}_d^l, \tilde{h}_s^l) \tag{7}$$

In this setup, $attn_D$ and $attn_S$ signify encoder-decoder and syntax-decoder attention, respectively. h_g^l is GAT's output, containing syntactic dependency features, and $\tilde{h}_t^l$ is the final attention feature obtained

[4] https://github.com/hankcs/HanLP

by concatenating $attn_D$ and $attn_S$. The predicted word is generated following a feed-forward network with a residual connection and softmax function, akin to the original Transformer model.

4. Experiments

To validate the effectiveness of the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) model, we conducted extensive experiments. These experiments were based on BLEU score assessments on two major datasets: the United Nations Parallel Corpus (UNPC)[5] and the Europarl Corpus[6]. The datasets involved were UNPC Chinese-English (Zh→En) and Russian-English (Ru→En), as well as Europarl German-English (De→En). For each language pair, we selected 1M sentence pairs for the training set, and 6K and 5K sentence pairs for the validation and test sets, respectively. Additionally, to simulate low-resource language scenarios and limited training set conditions, we progressively reduced the training set size.

The MT engine's encoder is a single BERT variant for each source language, serving as our baseline model (Baseline). The Baseline and the LSGIB engines were trained under similar conditions for fair comparison. The decoders are derived from the vanilla Transformer model, with each source language having its unique BERT variant. These decoders consist of 6 layers and 8 attention heads, while other parameters are kept constant. The Graph Attention Network (GAT) within the LSGIB engines has 2 layers and 6 attention heads for Zh, and 4 attention heads for Ru and De. All engines are trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a learning rate of 2e-5, word embedding size of 768, and cross-entropy as the loss function. The experiments were executed on NVIDIA RTX 3080 and 3090 GPUs.

As depicted in Table 1, the LSGIB engines show promising performance across various source languages, achieving comparable or superior BLEU scores relative to the baseline models. Furthermore, the LSGIB engines demonstrate improvements over the baseline in scenarios with small training samples. This indicates potential benefits for other low-resource languages or limited training set scenarios (detailed analysis in Appendix Sec **??**). The explicit syntactic knowledge represented by graph attention and BERT proves to be beneficial for learning linguistic structures in MT models. Following the insights from [34], we also employed the COMET QE model to reassess the performance of the engines. COMET provides a QE score ranging from 0 to 100, considering the relationship between the source sentence, its translation, and the reference. The LSGIB engines consistently outperform in terms of both BLEU and QE scores. However, in actual translation tasks where a reference may be missing, and translations both within and outside the domain are required, the QE model emerges as a more robust metric than BLEU for addressing such challenging situations.

Table 1. Performance of the LSGIB model compared with the Baseline on BLEU and COMET QE scores for three language pairs.

Training Data Size	**Zh→En**	**Baseline**	**LSGIBC**	**LSGIBD**
1M	BLEU	47.15	**47.23**	47.17
	COMET	82.20	83.69	**84.78**
	Ru→En	Baseline	LSGIBC	LSGIBD
	BLEU	47.22	**47.36**	47.27
	COMET	80.93	81.34	**82.56**
	De→En	Baseline	LSGIBC	LSGIBD
	BLEU	37.59	**37.67**	37.63
	COMET	78.02	78.66	**79.37**

[5] https://opus.nlpl.eu/UNPC.php

[6] https://opus.nlpl.eu/Europarl.php

5. Experiments

In exploring the impact of the LSGIB approach on translation quality, we focus on both the linguistic nuances and human interpretability of translations, areas where BLEU scores may not provide sufficient insight [35,36]. To this end, we utilize a gold syntactic annotation corpus alongside a Quality Estimation (QE) model. This methodology allows for a comprehensive assessment of the LSGIB model's ability to retain source sentence semantics, ensure coherence in translation semantics, and maintain rationality in word order.

5.1. Overall Translation Quality Evaluation

For a thorough evaluation, we processed the PUD corpus in Chinese[7], Russian[8], and German[9], translating each language to English using both the baseline and LSGIB engines. The state-of-the-art QE model[10] was employed to score the translations, with scores ranging from 0 to 1 indicating translation quality. To statistically validate the improvements, we used a paired t-test and box plots to analyze the changes in translation quality pre- and post-implementation of LSGIB, with a significance level set at 0.05.

Table 2 demonstrates that, in the case of the Chinese-to-English translations, the LSGIB model outperforms the baseline in terms of QE scores, indicating a significant improvement in translation quality. This pattern is consistent across Russian and German translations, as evident from the t-test results and p-values. These findings suggest that the integration of syntactic knowledge via graph representation and BERT within the LSGIB framework leads to a noticeable enhancement in the quality of MT outputs. Notably, the LSGIB model showcases its superiority in scenarios involving small training samples, suggesting potential benefits for low-resource languages and limited training set conditions.

Table 2. Analysis of translation quality improvements in PUD corpus translations across three languages using paired t-tests between baseline and LSGIB models.

Language	Sample Size	Models		$\bar{x}_d$	S_d	t	P-value
Zh	1000	Baseline	LSGIBC	0.024	0.109	7.18	$p < 0.001$
			LSGIBD	0.032	0.111	9.12	$p < 0.001$
Ru	1000	Baseline	LSGIBC	0.024	0.042	18.38	$p < 0.001$
			LSGIBD	0.034	0.045	23.67	$p < 0.001$
De	1000	Baseline	LSGIBC	0.007	0.113	2.162	$p = 0.030$
			LSGIBD	0.012	0.110	3.617	$p < 0.001$

5.2. Influence of Sentence Length on Translation Quality

To further understand the impact of LSGIB, we investigated the relationship between the length of source sentences and the improvements in translation quality. After translating the PUD corpus with the baseline engines and assessing them using the QE model, we identified the bottom 30% of translations in terms of quality. These translations were then categorized based on the length of their source sentences, with classifications for short (S), medium (M), and long (L) sentences. This categorization took into account the linguistic differences in character and word length across Chinese, Russian, and German.

As illustrated in Table 3, the LSGIB model showed improvements in translation quality across all sentence lengths and language pairs. Interestingly, the LSGIB Concatenation (LSGIBC) variant

[7] https://github.com/UniversalDependencies/UD_Chinese-PUD
[8] https://github.com/UniversalDependencies/UD_Russian-PUD
[9] https://github.com/UniversalDependencies/UD_German-PUD
[10] https://github.com/TharinduDR/TransQuest

exhibited a notable efficacy for longer sentences, while the LSGIB with Decoder (LSGIBD) variant was particularly effective in enhancing translations of shorter and medium-length sentences. This indicates the nuanced capability of LSGIB in adapting to different sentence structures and lengths, offering significant improvements especially in scenarios where traditional metrics like BLEU may not fully capture translation nuances.

Table 3. Analysis of QE scores for low-quality translations across different sentence lengths, comparing Baseline and LSGIB models.

		Zh		
Sen Length	Samples	Baseline	LSGIBC	LSGIBD
Long	93	0.425	**0.512**	0.508
Medium	142	0.423	0.500	**0.517**
Short	65	0.434	0.543	**0.560**
		Ru		
Sen Length	Samples	Baseline	LSGIBC	LSGIBD
Long	32	0.719	**0.751**	0.745
Medium	155	0.698	0.746	**0.750**
Short	113	0.686	**0.752**	0.747
		De		
Sen Length	Samples	Baseline	LSGIBC	LSGIBD
Long	57	0.513	**0.554**	0.549
Medium	150	0.512	0.561	**0.586**
Short	93	0.482	0.574	**0.578**

5.3. Impact of Syntactic Relations on Translation Quality

We next focused on how specific syntactic relations in source sentences are influenced by the LSGIB model. By grouping low-quality translations based on their syntactic relations, we were able to determine which types of dependencies benefited the most from our approach. Each group of sentences containing a specific dependency relation was analyzed to measure the average improvement in QE scores post-LSGIB application.

Table 4 presents a comparative analysis of syntactic relations improvement for each language under the LSGIB model. The study reveals diverse degrees of enhancement in translation quality based on dependency relations across different languages. It's noteworthy that while both SGBC and SGBD variants of LSGIB incorporate graph syntactic knowledge, their dependency learning patterns vary. For instance, in the Chinese language, the "flat" dependency is more effectively handled by SGBC compared to SGBD. This suggests a nuanced differentiation in how each variant processes syntactic relations, and how this influences the overall translation quality. Despite these differences, certain syntactic relations consistently show improvement across both SGBC and SGBD models, highlighting the LSGIB's capability to explicitly enhance understanding of specific common dependencies.

In assessing the influence of the LSGIB model on translation quality, a key question arises: how does the explicit syntactic knowledge encoded in graphs interplay with the decision-making process of BERT? To delve into this, we embark on an exploration of the interpretability of our model with respect to syntax, focusing on syntactic prediction tests using Graph Attention Network (GAT) and representation similarity analysis with BERT.

Table 4. Enhanced QE scores for specific syntactic dependencies in source sentences across three languages using the LSGIB model.

	Baseline	LSGIB-C		Baseline	LSGIB-D
Zh					
obl:agent	0.379	0.576	obl:agent	0.379	0.597
discourse:sp	0.388	0.502	iobj	0.387	0.511
flat	0.387	0.494	nsubj:pass	0.423	0.545
flat:name	0.415	0.518	appos	0.404	0.518
mark:prt	0.435	0.532	discourse:sp	0.388	0.501
Ru					
	Baseline	LSGIB-C		Baseline	LSGIB-D
orphan	0.608	0.768	orphan	0.608	0.719
aux	0.700	0.764	aux	0.700	0.777
ccomp	0.681	0.745	ccomp	0.681	0.747
flat:name	0.703	0.761	discourse	0.614	0.676
fixed	0.688	0.742	fixed	0.688	0.750
De					
	Baseline	LSGIB-C		Baseline	LSGIB-D
csubj	0.449	0.566	flat	0.442	0.625
flat	0.442	0.553	csubj	0.449	0.554
expl	0.486	0.573	expl	0.486	0.589
compound:prt	0.493	0.579	compound:prt	0.493	0.595
compound	0.495	0.577	cop	0.502	0.586

5.4. Syntactic Predictions in GAT

A crucial determinant of translation quality improvement is whether GAT can effectively comprehend syntactic structures. We investigate this by designing a syntactic dependency prediction task for GAT. This experiment aims to ascertain the correlation between syntactic knowledge represented on graphs and the resultant translation quality. The Parallel Universal Dependencies (PUD) corpus is utilized as the training, validation, and test sets for each language, divided into 800, 100, and 100 sentences, respectively. Within these sentences, words and their syntactic dependencies are modeled as nodes and edges on the graph. GAT's task is to learn node associations to predict various dependency relations, with the F1-score as the evaluation metric.

Table 5 showcases the training efficiency of GAT, highlighting that only 2 layers are necessary for it to effectively learn dependency relations. By correlating GAT's dependency relation prediction scores with the translation quality of source sentences containing these relations, we discern a significant pattern: proficient dependency relation learning by GAT is mirrored in the improved translation quality of corresponding sentences. For example, GAT's adept prediction of the 'conj' dependency in Chinese translations leads to notable enhancements in the translation quality of sentences containing this relation. This observation is consistent across Russian and German as well. However, some dependency relations, such as 'iobj' and 'nusbj:pass', present challenges for GAT's prediction capabilities, and relations like 'obl:tmod' in Chinese and German show lower prediction scores yet still contribute to improved translation quality. This indicates that while robust dependency relation learning by GAT is a key factor in enhancing translation quality, it is not an absolute determinant. Factors such as the encoder or decoder requiring more explicit structural information from GAT, irrespective of the correctness of syntactic annotation, also play a crucial role.

Table 5. Efficacy of GAT in learning syntactic dependencies as reflected by F1-scores across three languages.

	Zh			**Ru**			**De**	
	Samples	Score		Samples	Score		Samples	Score
mark	291	0.986	det	476	0.990	case	2053	0.992
cc	283	0.984	root	1000	0.987	cc	724	0.987
conj	383	0.970	amod	1791	0.982	det	2771	0.987
nummod	809	0.965	case	2121	0.978	mark	459	0.981
root	1000	0.955	aux:pass	128	0.974	advmod	1103	0.932
cop	251	0.945	cop	87	0.971	root	1000	0.931
det	338	0.935	advmod	914	0.934	aux:pass	230	0.927
case	1319	0.934	cc	599	0.930	amod	1089	0.913
nmod	702	0.933	flat:foreign	97	0.921	flat:name	164	0.876
amod	420	0.927	obl	1465	0.900	aux	365	0.868

Table 6. Correlation between GAT's syntactic prediction and BERT's layer similarity as indicated by RSA scores for each language. *: Representations from Baseline and LSGIB-D for comparison.

			Zh		
	GAT	RSA	Layer	RSA*	Layer
mark	0.986	0.178	4	0.208	4
cc	0.984	0.274	4	0.354	5
conj	0.970	0.380	5	0.152	5
nummod	0.965	0.274	4	0.237	3
root	0.955	0.216	4	0.390	4
			Ru		
	GAT	RSA	Layer	RSA*	Layer
det	0.990	0.426	4	0.408	3
root	0.987	0.466	3	0.504	3
amod	0.982	0.444	3	0.391	4
case	0.978	0.462	4	0.413	4
aux:pass	0.974	0.357	3	0.327	3
			De		
	GAT	RSA	Layer	RSA*	Layer
case	0.992	0.686	5	0.759	2
cc	0.987	0.591	6	0.741	6
det	0.987	0.584	8	0.817	6
mark	0.981	0.676	6	0.769	6
advmod	0.932	0.733	6	0.774	8

6. Conclusions

In this study, we introduced the Linguistic Structure through Graphical Interpretation with BERT (LSGIB) model, which represents a novel approach to integrating syntactic knowledge into machine translation (MT) tasks. The LSGIB model leverages the capabilities of the Graph Attention Network (GAT) in conjunction with the advanced linguistic comprehension of BERT to enhance translation quality. Our experiments provided insights into the mechanisms through which syntactic knowledge contributes to the improvement of translation outcomes. Notably, the LSGIB model demonstrates how explicit syntactic structures, when effectively captured and integrated, can lead to translations that are not only more accurate but also more coherent and semantically rich. This paper, through its exploration of two distinct approaches under the LSGIB umbrella, lays the groundwork for a deeper understanding of the role of syntactic knowledge in MT. We have shown that by incorporating syntactic information via GAT, along with the deep learning capabilities of BERT, it is possible to achieve significant enhancements in translation quality. Our findings underscore the potential of combining graph-based representations with pre-trained language models in the realm of natural language processing. Looking ahead, our future work will delve further into exploring advanced methods for modeling critical linguistic knowledge through graphical representations in MT tasks. We aim to refine and extend the LSGIB model, seeking to uncover more nuanced ways in which syntactic and semantic information can be harnessed to push the boundaries of translation accuracy and fluency. The ultimate goal is to develop MT systems that not only translate languages but do so with an understanding of linguistic intricacies akin to that of human translators.

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
2. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
4. Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.
5. Rong Yan, Jiang Li, Xiangdong Su, Xiaoming Wang, and Guanglai Gao. Boosting the transformer with the bert supervision in low-resource machine translation. *Applied Sciences*, 12(14):7195, 2022.
6. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
7. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
8. Santiago Egea Gómez, Euan McGill, and Horacio Saggion. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, Online (Virtual Mode), September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.bucc-1.4.
9. Ru Peng, Tianyong Hao, and Yi Fang. Syntax-aware neural machine translation directed by syntactic dependency degree. *Neural Computing and Applications*, 33(23):16609–16625, 2021.
10. Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
11. Lianzhe Huang, Xin Sun, Sujian Li, Linhao Zhang, and Houfeng Wang. Syntax-aware graph attention network for aspect-level sentiment classification. In *Proceedings of the 28th international conference on computational linguistics*, pages 799–810, 2020.
12. Gang Li, Chengpeng Zheng, Min Li, and Haosen Wang. Automatic requirements classification based on graph attention network. *IEEE Access*, 10:30080–30090, 2022.
13. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
14. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
15. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
16. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.

17. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
18. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
19. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802, 2021.
20. Kenji Imamura and Eiichiro Sumita. Recycling a pre-trained bert encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, 2019.
21. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
22. Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385, 2020.
23. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
24. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
25. Anna Currey and Kenneth Heafield. Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5203. URL https://aclanthology.org/W19-5203.
26. Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. Sg-net: Syntax guided transformer for language representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
27. Colin McDonald and David Chiang. Syntax-based attention masking for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 47–52, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-srw.7. URL https://aclanthology.org/2021.naacl-srw.7.
28. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559, 2021.
29. Mingfei Chen, Wencong Wu, Yungang Zhang, and Ziyun Zhou. Combining adversarial training and relational graph attention network for aspect-based sentiment analysis with bert. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2021.
30. Xiaotang Zhou, Tao Zhang, Chao Cheng, and Shinan Song. Dynamic multichannel fusion mechanism based on a graph attention network and bert for aspect-based sentiment classification. *Applied Intelligence*, pages 1–14, 2022.
31. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
32. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
33. Jian Li, Pinjia He, Jieming Zhu, and Michael R Lyu. Software defect prediction via convolutional neural network. In *Proceedings of the 2017 International Conference on Software Quality, Reliability and Security*, pages 318–328. IEEE, 2017.
34. Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.57.

35. Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.
36. Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.