EventBERT: Incorporating Event-Based Semantics for Natural Language Understanding

Anni Zou[1,2,3], Zhuosheng Zhang[1,2,3], and Hai Zhao[1,2,3](✉)

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
zhaohai@cs.sjtu.edu.cn

[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China

[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract. Natural language understanding tasks require a comprehensive understanding of natural language and further reasoning about it, on the basis of holistic information at different levels to gain comprehensive knowledge. In recent years, pre-trained language models (PrLMs) have shown impressive performance in natural language understanding. However, they rely mainly on extracting context-sensitive statistical patterns without explicitly modeling linguistic information, such as semantic relationships entailed in natural language. In this work, we propose EventBERT, an event-based semantic representation model that takes BERT as the backbone and refines with event-based structural semantics in terms of graph convolution networks. EventBERT benefits simultaneously from rich event-based structures embodied in the graph and contextual semantics learned in pre-trained model BERT. Experimental results on the GLUE benchmark show that the proposed model consistently outperforms the baseline model.

Keywords: Event-based semantics · Graph convolution networks · Natural language understanding

1 Introduction

Recent years have witnessed deep pre-trained language models (PrLM) such as ELMo [28], BERT [8], XLNet [45] and ERNIE [38] significantly prospering the performance of a wide range of natural language understanding (NLU) tasks. The remarkable advancements brought by PrLM have shown the effectiveness of leveraging contextualized representation. However, they mainly rest on extracting context-sensitive statistical patterns without explicitly modeling linguistic information such as semantic relationships in natural language.

It is clear that natural language itself abounds with ample, multi-level linguistic information. Although PrLMs like BERT implicitly represent linguistic knowledge more or less [33], studies disclose that linguistic knowledge is far from fully

This work was supported in part by the Key Projects of National Natural Science Foundation of China under Grants U1836222 and 61733011.

M. Sun et al. (Eds.): CCL 2022, LNAI 13603, pp. 66–80, 2022.
https://doi.org/10.1007/978-3-031-18315-7_5

absorbed [10,33]. Therefore, there emerges a series of derivatives of PrLM intending to fuse explicit linguistic knowledge so as to acquire better language representation, including syntactic [1,44,47] and semantic information [14,17,46].

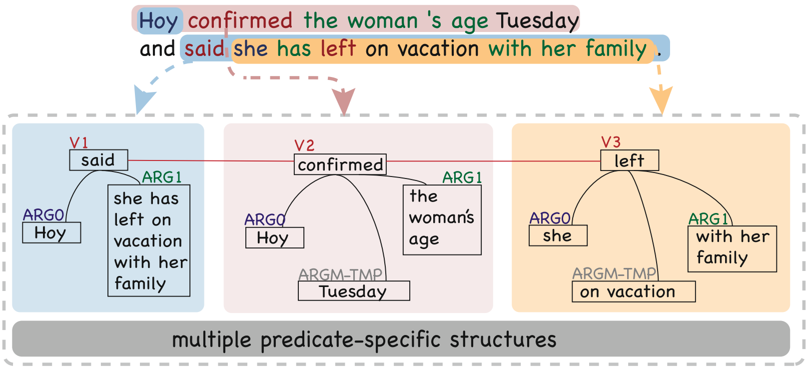

Fig. 1. An example showing how SRL parses sentences and the intuition of constructing event-based graph.

In cognition practice, human needs to distill semantics of different levels to gain a comprehensive understanding, whereas neural language models learn semantic representation to deal with downstream tasks [13]. Thus, effective learning of semantic knowledge plays a crucial role in NLU tasks and has gained growing attention recently. For instance, SemBERT [46] directly connects multiple predicate-argument structures acquired by semantic role labeler (SRL) to get the joint representation.

The essence of SRL [36] lies in that every sentence possesses multiple predicate-specific structures which can represent different frames of events, while semantic roles express the abstract role that arguments of a predicate can take in the event. Besides, the events inside a sentence have interactions with each other that serve together to present the overall semantic knowledge. As shown in Fig. 1, SRL parses every sentence with multiple predicate-specific structures which can serve as events inferring *who did what to whom*, *when and why*. Each event has an inner structure centered on the predicate to which several arguments are associated such as *Hoy[ARG0]*, *the woman's age[ARG1]* and *Tuesday[ARGM-TMP]* connected to *confirmed[V]*. Meanwhile, the multiple events work together to give a comprehensive meaning of a sentence, like the events centered on *said*, *confirmed* and *left*. With regard to delving into the inner interactions between the events and effectively capturing multiple objects, we are motivated to build a graph to reveal the intrinsic structures between and inside the events.

Inspired by the above ideas, we propose EventBERT: an event-based semantic representation model which takes BERT as the backbone and refines with event-based structural semantics. Our EventBERT benefits simultaneously from rich event-based structures embodied in the graph and contextual semantics learned in the pre-trained BERT.

Our proposed model works in three steps: it first applies an off-the-shelf SRL toolkit to parse every sentence with semantic role labels; then it constructs event-based graphs and employs Graph Convolutional Networks (GCNs) [35] to propagate and aggregate information from neighboring nodes on the graph; at last, it combines the contextualized representation acquired by BERT encoder together with the graph-level representation to obtain an event-based contextualized representation.

The key contributions of our work are summarized as follows:

1) We extract event-based semantic knowledge from SRL to enrich language representation.
2) We employ GCNs to construct sentence-level graphs which better reveal interactions inside and between the events in a sentence.

2 Related Work

2.1 Semantics in Language Representation

Recent studies show that current prominent pre-trained language models have already incorporated semantic information to some extent [6], yet such implicit semantic information is far from enough for comprehensive natural language understanding [10]. Thus there emerges a research line that focuses on fusing semantic information into contextualized language representation. ERNIE2.0 [38] adopts three-stage masking in which entity-level masking helps to obtain a word representation containing richer semantic information. SemBERT [46] makes use of PropBank [27] to fuse semantic role tags into language representation. FMSR [16] utilizes FrameNet [2] to extract multi-level semantic information within sentences. SS-MRC [15] takes advantage of syntax and frame semantics in an attempt to carve out information from two complementary perspectives to obtain richer language representation.

Besides simply employing semantic knowledge, other recent works shift the focus to exploring deeper structural semantics. For instance, frame semantics and graph neural networks are leveraged to model sentences from both intra-sentence level and inter-sentence level [14]. SIFT is introduced to inject predicate-argument semantic dependencies into pre-trained language models via R-GCNs [42]. Structured knowledge is introduced through multi-tasking to get a unified model, which inspires the potential of leveraging structural information [43]. Unlike previous works that attempt to capture shallow semantic structures by semantic tags, our model digs deeper into semantics itself and aims to find the structured event-based information behind semantics, thus unveiling richer structural-semantic information inside the sentence.

2.2 Graph Modeling for Language Understanding

As natural language itself abounds with dependencies and intricate relations between different levels of language units, graph neural networks (GNNs), which model the units as nodes in the graph and learn the weight via the message passing between nodes of the graph [18,34,39], stand out by explicitly and intuitively capturing the relations. Besides, a number of extensions to the original graph neural networks have been developed, the most notable of which include graph convolutional networks (GCNs) [18], graph attention networks (GANs) [39] and the models from [22] and [29] utilizing gating mechanisms to facilitate optimization.

In response to the outstanding performance of GCNs, several efforts have been made in recent years to improve performance on natural language understanding using GCNs, including GraphRel [12] which considers the interaction between named entities and relations via relation-weighted GCNs to better extract relations, NumNet [32] which utilizes a numerically-aware graph to perform numerical reasoning, DFGN [30] which dynamically builds the entity graph by adding the edges with co-occurrence relations, HGN [11] which creates a hierarchical graph by constructing nodes on different levels of granularity and social information reasoning [21] which uses GCNs to capture the documents' social context.

Moreover, R-GCNs [35] have shown effectiveness in relational graph modeling. For example, Entity-GCN [7] employs R-GCNs to link mentions of candidate answers for multi-document question answering. DFGN [30] dynamically builds the entity graph by adding the edges with co-occurrence relations and softly masking out irrelevant entities. DGM [26] constructs two discourse graphs and uses R-GCNs to fully capture interactions among the elements. R-GCNs are employed to enhance reference dependencies for dialogue disentanglement [23]. In contrast with previous works, our work proposes a sentence-level graph that is finely designed to mine the relationships between multiple elements in a sentence, extract rich structural semantics and facilitate information flow over the graph as well.

3 Model

Figure 2 gives an overview of our proposed EventBERT, which consists of two major components:

1. Context Encoder which acquires deep and contextualized representations for raw input sequences by following BERT architecture;
2. Event-based Encoder which obtains richer structural-semantic representation by modeling event-based intra-sentence graphs.

We omit the details of BERT which is widely used and ubiquitous and leave readers to resort to [8] for more information.

3.1 Context Encoder

The raw input sentence $X = \{x_1, \ldots, x_n\}$ is a sequence of words in length n. It is first tokenized to a sequence of sub-words with `[SEP]` inserted at the end as the end marker and `[CLS]` inserted at the beginning to get a sentence-level representation: $X' = \{token_1, \ldots, token_m\}$. Then we pass it through the embedding block and encoder block of BERT to produce a context-informed representation $C = \{c_1, \ldots, c_m\} \in \mathbb{R}^{m \times d_{hs}}$ using the equation below:

$$C = BERT(X'), \tag{1}$$

where m denotes the length of sentence on sub-word level and d_{hs} stands for the dimension of hidden states.

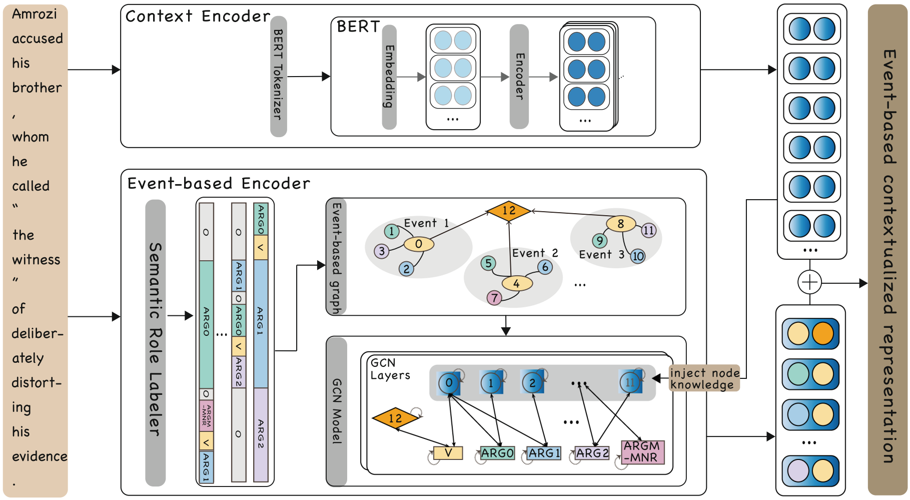

Fig. 2. The overall structure of EventBERT.

3.2 Event-Based Encoder

Semantic Role Labeler. The raw input sentence is simultaneously fed into Semantic Role Labeler [36] to fetch multiple predicate-specific structures tagged by PropBank semantic roles:

$$T = \{t_1, \ldots, t_d\}, \tag{2}$$

where d is the number of semantic structures for one sentence. Notably, t_i can be represented under the format $\{tag_1^i, tag_2^i, ..., tag_n^i\}$ and every tag span in t_i is recorded with its corresponding index in the context for further alignment.

Graph Construction. Figure 3 shows the process of graph construction: the predicates in the original input text are firstly extracted and an event subgraph is constructed with each predicate as the center; then a super event node (SEN) is applied to link all the predicates to collect the integral event information within the aggregated sentence; the Levi graph is finally constructed with reference to the method of [20], which is used to prepare the next stage of further computational operations on the graph.

For each sentence with the argument-predicate roles, we construct an event-based graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ with span-level nodes $v_i \in \mathcal{V}$ and labeled edges $(v_i, r, v_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ a relation type. Since every sentence has several semantic structures, here we take one structure as example and show the modeling method. Given $Seq_{tag} = \{tag_1, tag_2, ..., tag_n\}$ a word-level tag sequence,

1. We first transform it to a span-level sequence $Seq'_{tag} = \{tag'_1, tag'_2, ..., tag'_l\}$ by aggregating the same neighboring tags with $l \leq n$ representing the length of tags on span-level;
2. Then, we add a Super Event Node ($v = SEN$) to seize global graph information;

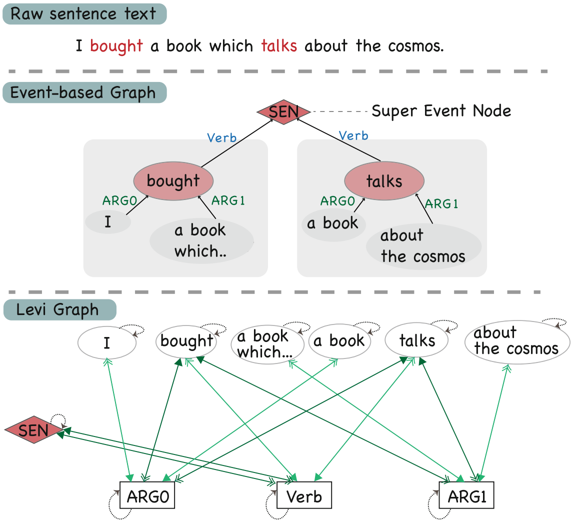

Fig. 3. The process of graph construction: from raw sentence text to event-based graph and corresponding Levi graph.

3. After that, we add other nodes and edges to G based on the following process:
   (a) we first find tag'_p which corresponds to predicate (*Verb* in e'),
   (b) we add a node $v = n_p$ and a directed edge $e = (n_p, Verb, SEN)$ with $r = Verb$,
   (c) for the rest tags referring to arguments of the predicate, tag'_q for example, we add a node $v = n_q$ and a directed edge linking to the predicate $e = (n_q, tag'_q, n_p)$ with relation $r = tag_q$;
4. Finally, the corresponding Levi graph [20] is extended from G to $G_L = (\mathcal{V}_L, \mathcal{E}_L, \mathcal{R}_L)$. For nodes $\mathcal{V}_L$, we add the nodes representing relations to the original: $\mathcal{V}_L = \mathcal{V} \cup \mathcal{R}$. For edges $\mathcal{E}_L$, we transform each edge $e = (n_q, tag'_q, n_p)$ in G into two corresponding edges: $e_1 = (n_q, tag'_q)$ and $e_2 = (tag'_q, n_p)$ in G_L. For $\mathcal{R}_L$, we follow the setting of [26] and refine it to five types: *default-in*, *default-out*, *reverse-in*, *reverse-out*, *self* according to the direction of edges towards the relation vertices, as is shown in Table 1.

Table 1. Relation types in our extended Levi graph

$\mathcal{R}_L$ in Levi graph	Illustration
default-in	The propagation path pointing to the node as the end point
default-out	The propagation path pointing to the node as the starting point
reverse-in	The propagation path in the opposite direction of *default-in*
reverse-out	The propagation path in the opposite direction of *default-out*
self	The propagation paths pointing to the node itself

Event-Based Contextualized Representation. We adopt Relational Graph Convolutional Networks (R-GCNs) [35] to implement explicit event graphs since traditional Graph Convolutional Networks (GCNs) cannot handle graphs containing edge features with multiple relations. For predicate and argument nodes, we inject the corresponding span-level encoding results obtained from Context Encoder in Sect. 3.1. For relation nodes, we regard the relations as embeddings and use a lookup table to get the initial representation. Given that the initial representation of each node v_i is h_i^0, the propagation process can be written as:

$$h_i^{(l+1)} = \mathrm{ReLU}\left(\sum_{r \in \mathcal{R}_L} \sum_{v_j \in \mathcal{N}_r(v_i)} \frac{1}{c_{i,r}} w_r^{(l)} h_j^{(l)}\right), \tag{3}$$

where $h_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ is the hidden state of node v_i in layer l with $d^{(l)}$ being the dimensionality of this layer's representations, $\mathcal{N}_r(v_i)$ denotes the set of neighbor indices of node v_i under the relation r, $c_{i,r}$ is a problem-specific normalization constant equal to $|\mathcal{N}_i^r|$, $w_r^{(l)}$ is the learnable parameters of layer l.

Since the importance of these relations cannot be treated the same, for example, the relation *Verb* is much more important than the relation *ARG2*, we introduce the gating mechanism [24]. The basic idea is to compute a value between 0 and 1 for message passing control as is shown in Eq. 4. Finally, the propagation process of R-GCNs under the gating mechanism is as follows:

$$g_j^{(l)} = \mathrm{Sigmoid}\left(h_j^{(l)} W_{r,g}^{(l)}\right) \tag{4}$$

$$h_i^{(l+1)} = \mathrm{ReLU}\left(\sum_{r \in \mathcal{R}_L} \sum_{v_j \in \mathcal{N}_r(v_i)} g_j^{(l)} \frac{1}{c_{i,r}} w_r^{(l)} h_j^{(l)}\right), \tag{5}$$

where $W_{r,g}^{(l)}$ is the learnable parameter under the l-th level relation type r.

With R-GCNs model, we obtain a graph-level semantic representation:

$$R = \{r_1, \ldots, r_f\} \in \mathbb{R}^{f \times d_{hs}} \tag{6}$$

where f is the number of nodes in the graph and d_{hs} is the same dimension as the representation C in Eq. 1 obtained from the context encoder.

At last, we concatenate R with the contextual sub-word-level representation C provided by Context Encoder and generate an event-based contextualized representation taking the mean value of both sub-word-level and graph-level information, which is then used as the new sequence representation for downstream tasks following the same way of [8].

4 Experiments

4.1 Setup

Datasets. We build EventBERT on the BERT backbone and fine-tune the model on GLUE (General Language Understanding Evaluation) benchmark [40] to evaluate the

Table 2. Comparisons between our models and baseline models on GLUE dev set.

Model	CoLA (mc)	SST-2 (acc)	MNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (acc)	QQP (acc)	STS-B (pc)	Avg -
Base-size									
$BERT_{BASE}$	58.4	92.8	83.2	88.6	68.5	86.0	86.5	87.8	81.5
$EventBERT_{BASE}$	59.6	93.3	83.9	91.8	69.7	89.7	89.8	88.9	83.3(↑1.8)
Large-size									
$BERT_{LARGE}$	60.3	93.1	85.2	91.5	70.3	88.5	90.2	89.3	83.6
$EventBERT_{LARGE}$	63.1	94.0	85.3	92.6	71.4	89.5	90.6	89.5	84.5(↑0.9)

performance, which includes two single-sentence tasks CoLA [41], SST-2 [37]), three similarity and paraphrase tasks MRPC [9], STS-B [4], QQP [5], three inference tasks MNLI [25], QNLI [31], RTE [3]. We exclude the controversial and problematic dataset WNLI [19].

Evaluation Metrics. According to [40], different datasets in GLUE correspond to different evaluation metrics, which include accuracy (acc), Matthew's correlation (mc) and Pearson correlation (pc). Among the eight datasets, STS-B is reported by Pearson correlation, CoLA is reported by Matthew's correlation, and other tasks are reported by accuracy.

Implementation Details. For the experiments, we use an initial learning rate in {1e−5, 2e−5, 3e−5} with warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size is selected in {16, 32}. The maximum number of epochs is set in $[2, 5]$ depending on tasks. Texts are tokenized with maximum length of 256 for the tasks. We use 2 layers of R-GCNs in our model.

4.2 Results

Table 2 presents the results on the GLUE benchmark, which show that EventBERT achieves consistent gains over all the subtasks under both base and large models.

The results indicate that our model performs better on longer sentences as shown in Sect. 5.3. Furthermore, our analysis shows that EventBERT can effectively benefit from the fine-grained graph-like event-based structures, as illustrated in case studies in Sect. 5.4. The results also disclose that modeling intrinsic structures between and inside events is crucial for language understanding.

In addition, the experimental results show that EventBERT has a significant performance gain on small datasets such as CoLA and MRPC, which indicates that semantic information involving event modeling is more advantageous and competitive in smaller datasets. In practice or industry, large-scale annotated data is rare and scarce due to the high cost and required expensive human resources, so language models that dominate in small-scale datasets are more valuable and important for most NLP tasks.

5 Analysis

5.1 Ablation Study

We conduct the ablation study to investigate the effects of the gating mechanism and the addition of global nodes in the event-based encoder module. Results in Table 3 show that both the gating mechanism and global nodes are non-trivial.

5.2 Methods of Aggregation

During the period of concatenating and aggregating the graph level semantic representation R and the contextual representation C, we further analyze the influence of different methods of aggregation such as max-pooling and mean-pooling by comparing the models with the same hyper-parameters on three datasets CoLA, MRPC and RTE respectively. Results in Table 3 demonstrate that employing mean-pooling presents better performance.

5.3 Effectiveness of Semantic Structures

In order to dig deeper into the rationale behind the effectiveness of the model, we select two datasets QNLI and MRPC, representing large-scale and small-scale datasets respectively. We statistically calculate the accuracy of the corresponding models on different word-level sequence length intervals for EventBERT and baseline. Figure 4 shows that our model outperforms the baseline especially when the sequence is relatively long and our model performs better on longer sentences compared with shorter ones, which implies that modeling intrinsic semantic structures is potential to guide the model to learn richer structural semantics more than contextualized information. Thus, the analysis of word sequence lengths shows that EventBERT performs better on data with longer sequence lengths, which indicates that event-level modeling is promising and competitive for understanding long texts. Under many practical situations where available data are long texts, the idea of extracting event-level structural-semantic information is promising in many NLP tasks.

Table 3. Ablation study and comparison of aggregation methods on three datasets.

Model	CoLA (mc)	MRPC (acc)	RTE (acc)
Ablation study			
EventBERT_{base}	59.6	89.7	69.7
w/o gating	58.6	86.8	69.0
w/o global node	58.4	87.0	67.9
Aggregation methods			
BERT_{base}	58.4	86.0	68.5
w/ max-pooling	59.1	86.8	68.2
w/ mean-pooling	59.6	89.7	69.7

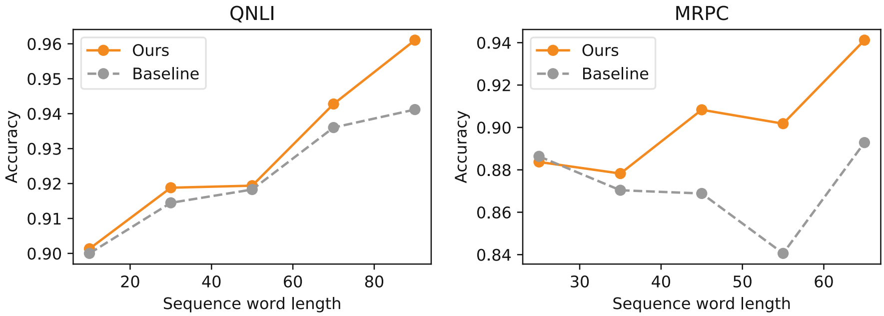

Fig. 4. Accuracy of different sequence word lengths on QNLI and MRPC.

5.4 Interpretability: Case Study

We select three cases in Classification, Sentence Similarity and Language Inference from SST-2, MRPC and QNLI respectively which are shown in Fig. 5, aiming to further explore the mechanism. It can be seen that our model can perceive explicit structural meaning to better understand the language. We will analyze each of the three cases in detail so as to analyze the advantages of EventBERT more intuitively.

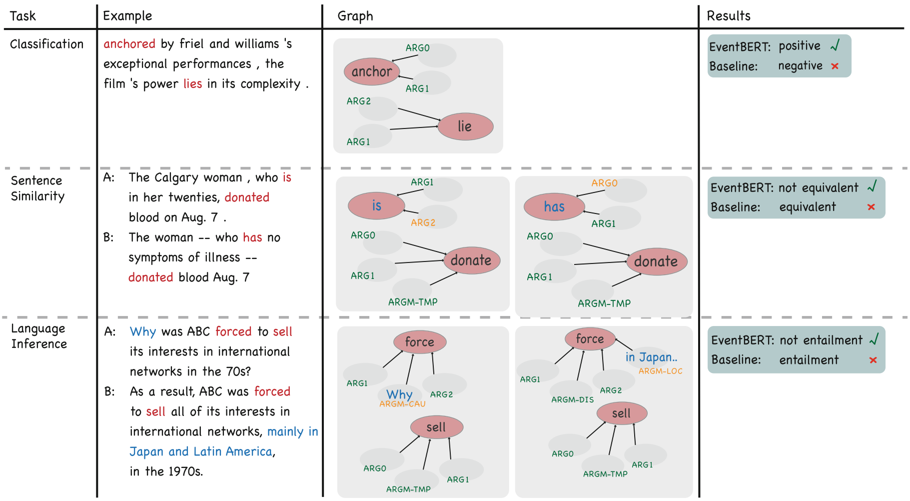

Fig. 5. Examples selected from the dev set of SST-2, MRPC and QNLI where baseline fails but our model succeeds.

Classification. In the case from SST-2, our model succeeds in capturing and understanding the event *Friel and william's exceptional performances[ARG0] anchored[V]* *the film's power[ARG1]*, whereas the baseline does not manage to capture this meaning, thus leading to the failure.

Sentence Similarity. The case from MRPC demonstrates that our model grabs the distinct semantic structures centered on *is* and *has* and thus gives the right answer *not equivalent*. The event centered on the predicate *donate* belongs to the same structure, which contains the arguments *ARG0*, *ARG1* and *ARGM-TMP* having the same contents (i.e., *the woman donated blood*). Nevertheless, the remaining events which center on the predicate *is* and the predicate *has* in the sentence pair are semantically different as one structure includes the arguments *ARG1* and *ARG2* while the other contains only *ARG0* and *ARG1*.

In Sentence Similarity tasks, two sentences in a sentence pair are likely to have one or several events in common, such as the event centered *donate* in this case. However, a subtle difference in a key element in the semantic structure of the sentence may also lead to a very different semantics of the whole sentence, such as the events centered on *is* and *has*. Our proposed model EventBERT precisely appreciates the value of abstracting structural semantics, benefiting from capturing event-based semantic knowledge to perceive the differences between sentences and thus make more accurate judgments.

Language Inference. Referring to the case from QNLI, as can be seen from Fig. 5, the question and paragraph texts are broadly similar in terms of *sell*-centered structure, both containing the arguments labeled *ARG0*, *ARG1*, and *ARGM-TMP*. However, by means of graph modeling, it can be clearly and explicitly observed that the structures centered on *force* are distinct, with the structure in the interrogative sentence containing the argument *ARGM-CAU* and the corresponding structure in the paragraph texts containing the argument *ARGM-LOC* instead. It is worth noting that one of the most crucial steps in determining whether a paragraph entails the correct answer to a question is whether the corresponding semantic structure in paragraph texts has the span labeled with the semantic role referring to the interrogative in the question. For example, in this case, the interrogative *Why* is exactly the *ARGM-CAU* of the predicate *force*; whereas the structure centered on *force* in the paragraph lacks the corresponding argument content and is replaced by *ARGM-LOC* instead. Therefore, it can be easily inferred that the paragraph focuses on the location (i.e., *in Japan and Latin America*) while the question concentrates on the cause (i.e., *Why*), which exactly reflects that there is no answer span for the interrogative of the question.

It is known that interrogative in the question and corresponding answer span should belong to the same semantic role. EventBERT takes full advantage of extracting abstracted semantics based on predicates, thus conducting language inference tasks more efficiently.

5.5 Error Analysis

We select bad cases of the baseline model and further investigate the ones of which our EventBERT also fails to predict the correct answers. We study two cases respectively from MRPC and QNLI as is shown in Table 4. The first error is caused by EventBERT's identification of the argument *in a written statement* of the predicate *said* in the first sentence, which is not entailed in the second sentence. However, the lack of this argument does not affect the main semantic information. The second error is due to argument reference confusion for the special predicate *is*. For instance, the interrogative *What* is labeled as *ARG2* whereas the correct answer *Hypersensitivity* is labeled as *ARG1*. From the above error cases, it may suggest that our model needs to have a more accurate perception of semantic relationships, which is left for future studies.

Table 4. Errors in predictions for cases in MRPC and QNLI dev set. The words in magenta indicate the key predicate. The words in blue indicate the key arguments referred to the predicate.

Example	EventBERT	Golden answer
This decision is clearly incorrect, FTC Chairman Timothy Muris *said in a written statement*. The decision is clearly incorrect, FTC Chairman Tim Muris *said*	*Not equivalent*	Equivalent
What is the name for a response of the immune system that damages the body's native tissues? *Hypersensitivity is* an immune response that damages the body's own tissues	*Not entailment*	Entailment

6 Conclusion

In this work, we propose EventBERT, an event-based semantic representation model that builds on BERT architecture and incorporates event-based structural semantics in terms of graph network modeling for fine-grained language representation. Experiments on a wide range of NLU tasks show the effectiveness of our model by consistently surpassing the baseline. While most existing works focus on fusing accurate semantic signals to enhance semantic information, we open up a novel perspective to model intrinsic structural semantics for deeper comprehension and inference in an intuitive and explicit way.

References

1. Bai, J., et al.: Syntax-BERT: improving pre-trained transformers with syntax trees. arXiv preprint arXiv:2103.04350 (2021)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, vol. 1, pp. 86–90. Association for Computational Linguistics (1998). https://doi.org/10.3115/980845.980860. https://aclanthology.org/P98-1013
3. Bentivogli, L., Clark, P., Dagan, I., Giampiccolo, D.: The fifth pascal recognizing textual entailment challenge. In: ACL-PASCAL (2009)

4. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055 (2017)
5. Chen, Z., Zhang, H., Zhang, X., Zhao, L.: Quora question pairs (2018)
6. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: 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, Florence, Italy, pp. 276–286. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-4828. https://aclanthology.org/W19-4828
7. De Cao, N., Aziz, W., Titov, I.: Question answering by reasoning across documents with graph convolutional networks. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long and Short Papers), Minneapolis, Minnesota, vol. 1, pp. 2306–2317. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1240. https://aclanthology.org/N19-1240
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, (Long and Short Papers), Minneapolis, Minnesota, vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
9. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: IWP 2005 (2005)
10. Ettinger, A.: What BERT is not: lessons from a new suite of psycholinguistic diagnostics for language models. Trans. Assoc. Comput. Linguist. **8**, 34–48 (2020). https://doi.org/10.1162/tacl_a_00298. https://aclanthology.org/2020.tacl-1.3
11. Fang, Y., Sun, S., Gan, Z., Pillai, R., Wang, S., Liu, J.: Hierarchical graph network for multi-hop question answering. arXiv preprint arXiv:1911.03631 (2019)
12. Fu, T.J., Li, P.H., Ma, W.Y.: GraphRel: modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1409–1418 (2019)
13. Geeraerts, D., Cuyckens, H.: Introducing cognitive linguistics. In: The Oxford Handbook of Cognitive Linguistics (2007)
14. Guan, Y., Guo, S., Li, R., Li, X., Tan, H.: Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, pp. 4045–4052. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.emnlp-main.331. https://aclanthology.org/2021.emnlp-main.331
15. Guo, S., Guan, Y., Li, R., Li, X., Tan, H.: Incorporating syntax and frame semantics in neural network for machine reading comprehension. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 2635–2641. International Committee on Computational Linguistics (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.237. https://aclanthology.org/2020.coling-main.237
16. Guo, S., Guan, Y., Li, R., Li, X., Tan, H.: Frame-based multi-level semantics representation for text matching. Knowl.-Based Syst. **232**, 107454 (2021). https://doi.org/10.1016/j.knosys.2021.107454. https://www.sciencedirect.com/science/article/pii/S0950705121007164
17. Guo, S., et al.: A frame-based sentence representation for machine reading comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 891–896. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.83. https://aclanthology.org/2020.acl-main.83

18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
19. Levesque, H., Davis, E., Morgenstern, L.: The Winograd schema challenge. In: Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (2012)
20. Levi, F.W.: Finite geometrical systems: six public lectures delivered in February, 1940, at the University of Calcutta. University of Calcutta (1942)
21. Li, C., Goldwasser, D.: Encoding social information with graph convolutional networks forpolitical perspective detection in news media. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2594–2604 (2019)
22. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.: Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493 (2015)
23. Ma, X., Zhang, Z., Zhao, H.: Structural characterization for dialogue disentanglement. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 285–297. Association for Computational Linguistics (2022). https://aclanthology.org/2022.acl-long.23
24. Marcheggiani, D., Titov, I.: Encoding sentences with graph convolutional networks for semantic role labeling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp. 1506–1515. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/D17-1159. https://aclanthology.org/D17-1159
25. Nangia, N., Williams, A., Lazaridou, A., Bowman, S.R.: The RepEval 2017 shared task: multi-genre natural language inference with sentence representations. In: RepEval (2017)
26. Ouyang, S., Zhang, Z., Zhao, H.: Dialogue graph modeling for conversational machine reading. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 3158–3169. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.findings-acl.279. https://aclanthology.org/2021.findings-acl.279
27. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Computat. Linguist. **31**(1), 71–106 (2005). https://doi.org/10.1162/0891201053630264. https://aclanthology.org/J05-1004
28. Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers), New Orleans, Louisiana, vol. 1, pp. 2227–2237. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-1202. https://aclanthology.org/N18-1202
29. Pham, T., Tran, T., Phung, D., Venkatesh, S.: Column networks for collective classification. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
30. Qiu, L., et al.: Dynamically fused graph network for multi-hop reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 6140–6150. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1617. https://aclanthology.org/P19-1617
31. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: EMNLP (2016)
32. Ran, Q., Lin, Y., Li, P., Zhou, J., Liu, Z.: NumNet: machine reading comprehension with numerical reasoning. arXiv preprint arXiv:1910.06701 (2019)
33. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in BERTology: what we know about how BERT works. Trans. Assoc. Comput. Linguist. **8**, 842–866 (2020). https://doi.org/10.1162/tacl_a_00349. https://aclanthology.org/2020.tacl-1.54
34. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2008)

35. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 593–607. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_38
36. Shi, P., Lin, J.J.: Simple BERT models for relation extraction and semantic role labeling. arXiv abs/1904.05255 (2019)
37. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP (2013)
38. Sun, Y., et al.: Ernie 2.0: a continual pre-training framework for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8968–8975 (2020)
39. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. Stat **1050**, 20 (2017)
40. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: 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, Brussels, Belgium, pp. 353–355. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-5446. https://aclanthology.org/W18-5446
41. Warstadt, A., Singh, A., Bowman, S.R.: Neural network acceptability judgments. arXiv preprint arXiv:1805.12471 (2018)
42. Wu, Z., Peng, H., Smith, N.A.: Infusing finetuning with semantic dependencies. Trans. Assoc. Computat. Linguist. **9**, 226–242 (2021). https://doi.org/10.1162/tacl_a_00363
43. Xie, T., et al.: UnifiedSKG: unifying and multi-tasking structured knowledge grounding with text-to-text language models. arXiv preprint arXiv:2201.05966 (2022)
44. Xu, Z., et al.: Syntax-enhanced pre-trained model. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5412–5422. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.420. https://aclanthology.org/2021.acl-long.420
45. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
46. Zhang, Z., et al.: Semantics-aware BERT for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9628–9635 (2020)
47. Zhang, Z., Wu, Y., Zhou, J., Duan, S., Zhao, H., Wang, R.: SG-Net: syntax-guided machine reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9636–9643 (2020)