EventBERT

Anonymous ACL submission

Abstract

Pre-trained language models (PrLMs) have shown impressive performance in natural language understanding. However, they mainly rest on extracting context-sensitive statistical patterns without explicit modeling of 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 network. 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 the effectiveness.

1 Introduction

Recent years have witnessed deep pre-trained language models (PrLM) such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLnet (Yang et al., 2019) and ERNIE (Sun et al., 2020) 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 learning 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 (Rogers et al., 2020), we have to confess that linguistic knowledge is far from fully absorbed (Ettinger, 2020; Rogers et al., 2020). Therefore, there emerges a series of derivatives of PrLM intending to fuse explicit linguistic knowledge so as to acquire better language representation, including syntactic and semantic information (Zhang et al., 2020b,a; Xu et al., 2021).

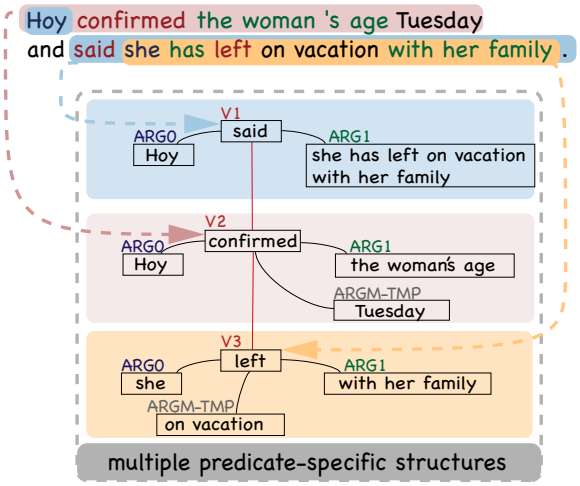

Figure 1: An example showing how SRL parses sentences and the intuition of constructing event-based graph.

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 (Geeraerts and Cuyckens, 2007). Thus, effective learning of semantic knowledge plays a crucial role in NLU tasks and has gained growing attention recently. For instance, Zhang et al. (2020a) proposed SemBERT, which directly connects multiple predicate-argument structures acquired by semantic role labeler (SRL) to get the joint representation.

The essence of SRL (Shi and Lin, 2019) 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 which serve together to present the overall semantic knowledge.

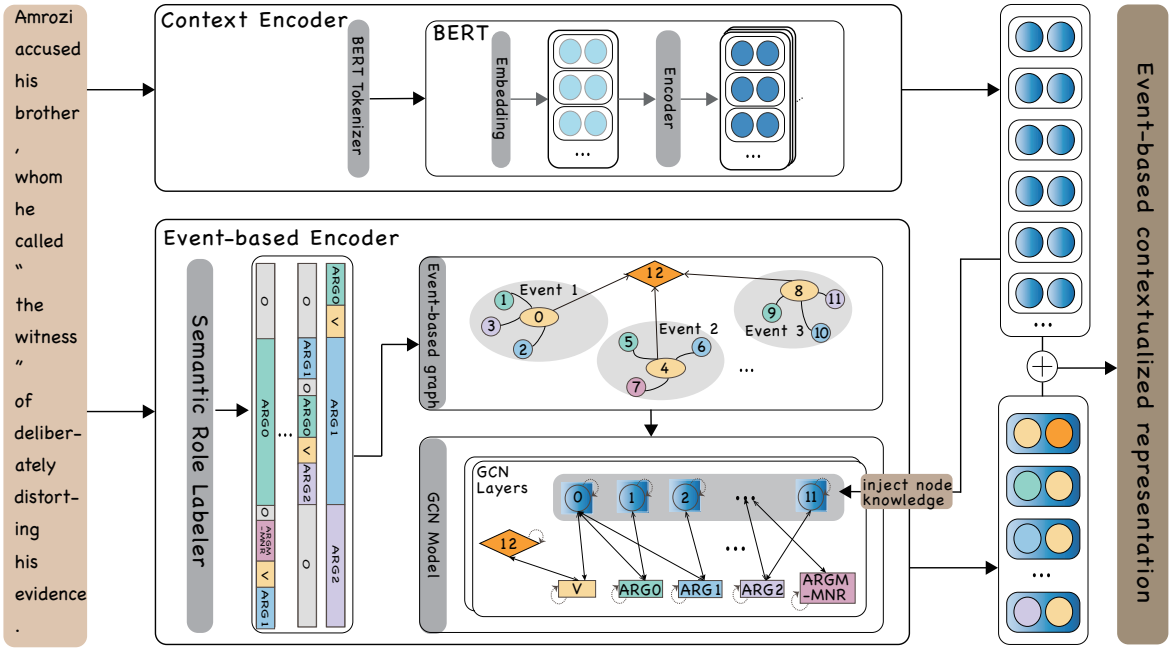

Figure 2: The overall structure of EventBERT.

As shown in Figure 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 *Hog[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*, *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 developed by AllenNLP to parse every sentence with semantic role labels; then it constructs event-based graphs and employs Graph Convolutional Networks (GCNs) (Schlichtkrull et al., 2018) 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 finally.

The key contributions of our work are summarized as follows:

1. To our best knowledge, our work is the first attempt to extract event-based semantic knowledge employing 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

Recent studies show that current prominent pre-trained language models have already incorporated semantic information to some extent (Clark et al., 2019), yet such implicit semantic information is far from enough for comprehensive natural language understanding (Ettinger, 2020). Thus there emerges a research line that focuses on fusing semantic information into contextualized language representation. ERNIE2.0 (Sun et al., 2020) adopts three-stage masking in which entity-level masking helps to obtain a word representation containing richer semantic information. Other works like SemBERT (Zhang et al., 2020a) and FMSR (Guo et al.,

2021) make use of two major semantic frames PropBank (Palmer et al., 2005) and FrameNet (Baker et al., 1998) respectively to capture explicit semantic information. Unlike previous works that attempt to capture semantic structures by semantic tags, our model is the first to construct event-based graphs on sentence level, unveiling richer structural semantic information inside the sentence.

3 Model

Figure 2 gives an overview of our proposed model, EventBERT with 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 Devlin et al. (2019) for more information.

3.1 Context Encoder

The raw input sentence $X = \{x_1, \ldots, x_n\}$ is a sequence of words of length n. It is first tokenized to a sequence of sub-words with `[CLS]` inserted at the beginning to get a sentence-level representation. Then we pass it through the embedding block and encoder block of BERT to produce a context-informed representation: $C = \{c_1, \ldots, c_m\} \in \mathbb{R}^{m \times d_{hs}}$, where m denotes the length of sentence on sub-word level and d_{hs} stands for the dimension of hidden states.

3.2 Event-based Encoder

Semantic Role Labeler The raw input sentence is simultaneously fed into SRL to fetch multiple predicate-specific structures tagged by PropBank (Shi and Lin, 2019) semantic roles: $T = \{t_1, \ldots, t_d\}$, where d is the number of semantic structures for one sentence. Notably, t_i can be represented under the format $\{tag_1^i, tag_2^i, ..., tag_n^i\}$.

Graph Construction Figure 3 shows the process of graph construction. 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 $e = \{tag_1, tag_2, ..., tag_n\}$ a word-level tag

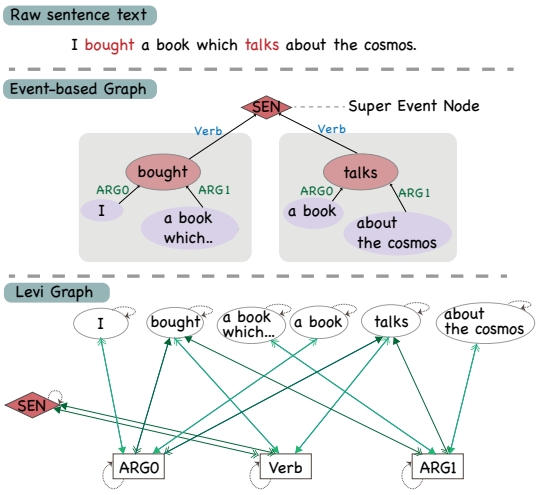

Figure 3: The process of graph construction: from raw sentence text to event-based graph and corresponding Levi graph.

sequence, we first transform it to a span-level sequence $e' = \{tag'_1, tag'_2, ..., tag'_l\}$ by aggregating the same neighboring tags with $l \leq n$ representing the length of tags on span-level. Then we add a Super Event Node ($v = SEN$) to seize global graph information. After that, we add other nodes and edges to G based on the following process: We first find tag'_p which corresponds to predicate (*Verb* in e'). We add a node $v = n_p$ and a directed edge $e = (n_p, Verb, SEN)$ with $r = Verb$. 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 $e = (n_p, tag'_q, n_p)$ with relation $r = tag_q$.

Finally, the corresponding Levi graph (Levi, 1942) is extended from G to $G_L = (\mathcal{V}_L, \mathcal{E}_L, \mathcal{R}_L)$ with $\mathcal{V}_L = \mathcal{V} \cup \mathcal{R}$ and $\mathcal{E}_L$ under the form of (n_q, tag'_q), (tag'_q, n_p). For $\mathcal{R}_L$, we follow the setting of Ouyang et al. (2021) 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.

Event-based Contextualized Representation A natural way to model relational graph is to adopt Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018). For predicate and argument nodes, we inject the corresponding span-level encoding results from Context Encoder in Section 3.1. For relation nodes, we regard the relations as embeddings and use a lookup table to get initial representation. Given initial representation h_i^0 for every node v_i, we implement the propagation

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

Table 1: Comparisons between our models and baseline models on GLUE dev set. STS-B is reported by Pearson correlation, CoLA is reported by Matthew's correlation, and other tasks are reported by accuracy.

process as follows:

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

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. $g_j^{(l)}$ is a gated value (Marcheggiani and Titov, 2017) between 0 and 1 for information passing control. $c_{i,r}$ is a problem-specific normalization constant equal to $|\mathcal{N}_i^r|$. $w_r^{(l)}$ is the learnable parameters of layer l. Through the R-GCN model, we obtain a graph-level semantic representation: $R = \{r_1, \ldots, r_f\} \in \mathbb{R}^{f\times d_{hs}}$, where f is the number of nodes in the graph and d_{hs} is the same dimension as C.

At last, we concatenate R with the pooled contextual sub-word-level representation c_1 of special token `[CLS]` 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 Devlin et al. (2019).

4 Experiments

4.1 Setup

We build EventBERT on the BERT backbone and fine-tune the model on GLUE (General Language Understanding Evaluation) benchmark (Wang et al., 2018) to evaluate the performance, which includes two single-sentence tasks (CoLA (Warstadt et al., 2018), SST-2 (Socher et al., 2013)), three similarity and paraphrase tasks (MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017), QQP (Chen et al., 2018)), three inference tasks (MNLI (Nangia et al., 2017), QNLI (Rajpurkar et al., 2016), RTE (Bentivogli et al., 2009).

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.

4.2 Results

Table 1 presents the results on the GLUE benchmark, which show that EventBERT achieves consistent gains over all the subtasks under both of the base and large models. The results indicate that EventBERT can effectively benefit from the fine-grained graph-like event-based structures, as illustrated in case studies in Appendix B. Our analysis shows that our model performs better on longer sentences as shown in Appendix A. The results also disclose that modeling intrinsic structures between and inside events is critical for language understanding.

5 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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *ACL-PASCAL*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP2005*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Dirk Geeraerts and Hubert Cuyckens. 2007. Introducing cognitive linguistics. In *The Oxford handbook of cognitive linguistics*.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2021. Frame-based multi-level semantics representation for text matching. *Knowledge-Based Systems*, 232:107454.

Friedrich Wilhelm Levi. 1942. *Finite geometrical systems: six public lectues delivered in February, 1940, at the University of Calcutta*. University of Calcutta.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *RepEval*.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Dialogue graph modeling for conversational machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3158–3169, Online. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Peng Shi and Jimmy J. Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020a. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643.

A Effectiveness of semantic structures

Figure 4 shows that our model surpasses 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.

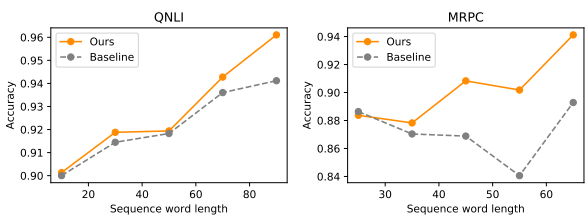

Figure 4: Accuracy of different sequence word lengths on QNLI and MRPC.

B Interpretability: Case Study

We select three cases in Classification, Sentence Similarity and Language Inference from SST-2, MRPC and QNLI, which are shown in Figure 5. It can be seen that our model can perceive explicit structural meaning to better understand the language. For example, in the first case, our model succeeds in understanding the event *Friel and william's exceptional performances[ARG0] anchored[V] the film's power[ARG1]*, whereas baseline does not manage to capture this meaning, thus leading to the failure. The second case demonstrates that our model grabs the distinct semantic structures centered on *is* and *has* and thus gives the right answer *not equivalent*. Referring to the third case, it can be easily observed that the argument structures centered on *force* are different which exactly reflects that there is no answer span for the interrogative *Why*.

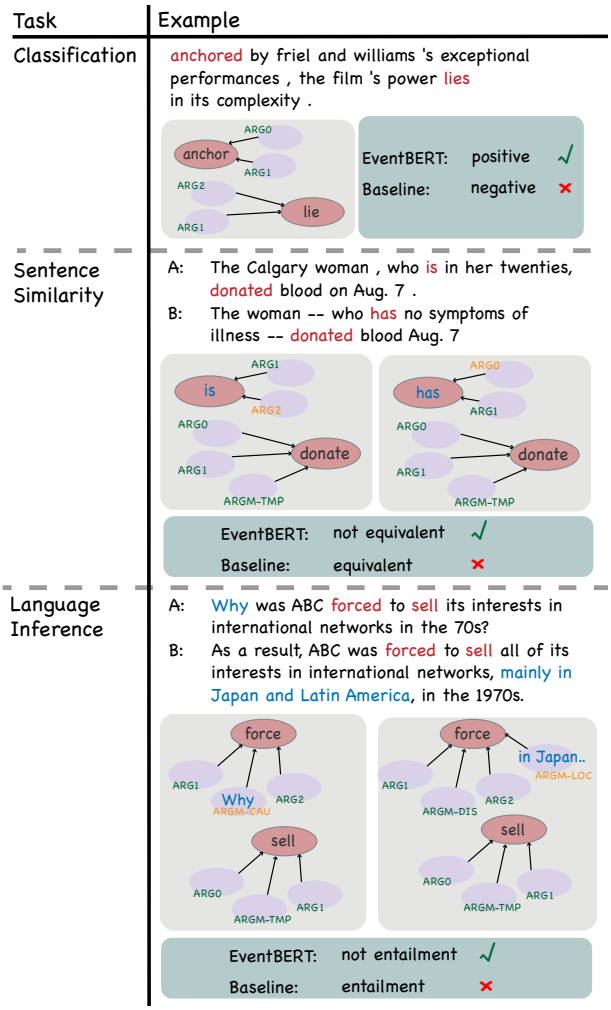

Figure 5: Examples selected from the dev set of SST-2, MRPC and QNLI where baseline fails but our model succeeds.