# Event-Event Relation Extraction using Probabilistic Box Embedding

Anonymous ACL submission

## Abstract

To understand a story with multiple events, it is important to capture the proper relations across these events. However, existing event relation extraction (ERE) framework regards it as a multi-class classification task and do not guarantee any coherence between different relation types, such as anti-symmetry. If a phone line *died* after *storm*, then it is obvious that the *storm* happened before the *died*. Current framework of event relation extraction do not guarantee this coherence and thus enforces it via constraint loss function (Wang et al., 2020). In this work, we propose to modify the underlying ERE model to guarantee coherence by representing each event as a box representation (BERE) without applying explicit constraints. From our experiments, BERE also shows stronger conjunctive constraint satisfaction while performing on par or better in $F_1$ compared to previous models with constraint injection.

## 1 Introduction

A piece of text can contain several events. In order to truly understand this text, it is vital to understand the subevent and temporal relationships between these events.(Mani et al., 2006a; Chambers and Jurafsky, 2008; Yang and Mitchell, 2016; Araki et al., 2014). Both temporal as well as subevent relationships between events satisfy transitivity constraints. For instance, "There was a *storm* in Atlanta in the night. All the phone lines were *dead* the next morning. I was not able to *call* for help.", the event marked by *dead* occurs after *storm* and the event *call* occurs after *dead*. Hence, by transitivity, a sensible model should predict that *storm* occurs before *call*. In general, predicting the relationships between different events in the same document, such that these predictions hold coherent structure, is a challenging task (Xiang and Wang, 2019).

While previous work utilizing neural methods provide competitive performances, these works employ multi-class classification per event-pair independently and are not capable of preserving logical constraints among relations, such as asymmetry and transitivity, during training time (Ning et al., 2019; Han et al., 2019a). To address this problem Wang et al. (2020) introduced a constrained learning framework, wherein they enforce logical coherence amongst the predicted event types through extra loss terms. However, since the coherence is enforced in a soft manner using extra loss terms, there is still room for incoherent predictions. In this work, we show that it is possible to induce coherence in a much stronger manner by representing each event using a box (Dasgupta et al., 2020).

We propose a Box Event Relation Extraction (`BERE`) model that represents each event as a probabilistic box. Box embeddings (Vilnis et al., 2018) were first introduced to embed nodes of hierarchical graphs in to into euclidean space using hyperrectangles, which were later extended to jointly embed multi-relational graphs and perform logical queries (Patel et al., 2020; Abboud et al., 2020). In this paper, we represent an event complex using boxes–one box for each event. Such a model enforces logical constraints by design (see Section 3.2). Consider the example in Figure 1. Event *dead* ($e_2$) follows event *storm* ($e_1$), indicating $e_2$ is child of $e_1$. Boxes can represent these two events as separate representations and by making $e_1$ to contain the box $e_2$, which not only preserve their semantics, but also can infer its antisymmetric relation that event $e_3$ is a parent of event $e_1$. However, the previous models based on pairwise-event vector representations have no real relation between representations $(e_1, e_2)$ and $(e_2, e_1)$ that can guarantee the logical coherence.

Experimental results over three datasets, HiEve, MATRES, and Event StoryLine (ESL), show that our method improves the baseline (Wang et al., 2020) by 6.8 and 4.2 $F_1$ points on single task and by 0.95 and 3.29 $F_1$ points on joint task over symmetrical dataset. Furthermore, our approach without using constrained learning clearly decreases conjunctive constraints by 4.36% and 3.29% on single task and by 0.4% and 1.14% on joint task over asymmetrical and symmetrical datasets, respectively. We show that handling antisymmetric constraints, that exist among different relations, can satisfy the interwined conjunctive constraints and encourage the model towards a coherent output across temporal and subevent tasks.

## 2 Background

**Task description** Given a document consisting of multiple events $e_1, e_2, \dots, e_n$, we wish to predict the relationship between each event pair $(e_i, e_j)$. We denote by $\mathbf{R}_{e_i,e_j}$ the relation between event pair ($e_i$, $e_j$). It value in the label space { PARENT-CHILD, CHILD-PARENT, COREF, NOREL} for subevent relationship (HiEve) and {BEFORE, AFTER, EQUAL, VAGUE} for temporal relationship (MATRES).[1] Both subevent and temporal relationships have four similar-category relationship labels where the first two labels, (PARENT-CHILD,CHILD-PARENT) and (BEFORE, AFTER) hold reciprocal relationship, the third label (COREF and EQUAL) occurs when it is hard to tell which of the first two labels that event pair should be classified to. Lastly, the last label NOREL and VAGUE represents a case when an event pair is not related at all.

**Logical constraints** *Symmetry constraint* indicate the event pair $(e_1, e_2)$ with relation $\mathbf{R}_{e_1,e_2}$ (BEFORE) flipping orders will have the reversed relation $\bar{\mathbf{R}}_{e_2,e_1}$ (AFTER), i.e. $\mathbf{R}_{e_i,e_j} \leftrightarrow \bar{\mathbf{R}}_{e_i,e_j}$. *Conjunctive constraints* refer to the constraints that exist in the relations among any event triplet. Given three event pairs, $(e_i, e_j), (e_j, e_k)$, and $(e_i, e_k)$, then the relation of $R_{e_i,e_k}$ has to fall into the conjunction set $\mathcal{D}(R_{e_i,e_j}, R_{e_j,e_k})$ specified based on relations of $(e_i, e_j)$ and $(e_j, e_k)$ (see Appendix G for more details).

**Box embeddings** A box $b = \prod_{i=1}^{d} [b_{m,i}, b_{M,i}]$ such that $b \subseteq R^d$ is characterized by its min and max endpoints $b_m, b_M \in \mathbb{R}^d$, with $b_{m,i} < b_{M,i}\ \forall i$. In the probabilistic gumbel box, these min and max points are taken to be independent gumbel-max and gumbel-min random variables, respectively. As shown in Dasgupta et al. (2020), if $b$ and $c$ are two such gumbel boxes then their volume and intersection is given as:

$$\mathrm{Vol}(b) = \prod_{i=1}^{d} \log\Big(1 + \exp\big(\frac{b_{M,i} - b_{m,i}}{\beta} - 2\gamma\big)\Big)$$
$$b \cap c = \prod_{i=1}^{d} \Big[l(b_{m,i}, c_{m,i}; \beta), l(b_{M,i}, c_{M,i}; -\beta)\Big],$$

where $l(x, y; \beta) = \beta \log(e^{\frac{x}{\beta}} + e^{\frac{y}{\beta}})$, $\beta$ is the temperature, which is a hyperparameter, and $\gamma$ is the Euler-Mascheroni constant.[2]

## 3 BERE model

In this section, we present the proposed box model BERE for event-event relation extraction. As depicted in Figure 1, the proposed model encodes each event $e_i$ as a box $b_i$ in $\mathbb{R}^d$ based on $e_i$'s contextualized vector representation $h_i$. As described in §3.1, the relation between $(e_i, e_j)$ is then predicted using conditional probability scores $P(b_i|b_j) = \mathrm{Vol}(b_i \cap b_j)/\mathrm{Vol}(b_j)$, $P(b_j|b_i) = \mathrm{Vol}(b_i \cap b_j)/\mathrm{Vol}(b_i)$ defined on box space. Lastly, §3.2 describes loss function used to learn the parameters of the model.

### 3.1 Inference rule on conditional probability

Notice that given two boxes $b_i$ and $b_j$, a higher value of $P(b_i|b_j)$ (resp. $P(b_j|b_i)$) implies that box $b_j$ is contained in $b_i$ (resp. $b_i$ contained in $b_j$). Moreover, other than complete containment in either direction, there are other two prominent configurations possible, i.e. one where $b_i$, $b_j$ overlap but none contains the other, and the one where $b_i$, $b_j$ do not overlap. It is possible to capture all four configurations by comparing the values of $P(b_i|b_j)$ and $P(b_j|b_i)$ with a threshold $\delta$. Figure 1(B) states our classification rule formulated based on this observation. With this formulation we have the desired symmetry constraint, i.e., $\mathbf{R}_{e_i,e_j}$ = PARENT-CHILD $\iff$ $\mathbf{R}_{e_j,e_i}$ = CHILD-PARENT, satisfied by design.

### 3.2 Loss functions for training

**BCE loss** As we require two dimensions of scalar $P(b_i|b_j)$ and $P(b_j|b_i)$ to classify $\mathbf{R}_{e_i,e_j}$, and for ease of notation, we define our label space with 2-dimensional binary variable $y^{(i,j)}$ as shown in Figure1(b). Where $y_0^{(i,j)} = I(P(b_i|b_j) \geq \delta)$ and $y_1^{(i,j)} = I(P(b_j|b_i) \geq \delta)$ where $I(\cdot)$ stands for

[1] See Appendix C for the detailed information of HiEve and Matres.

[2] https://en.wikipedia.org/wiki/Euler%27s_constant

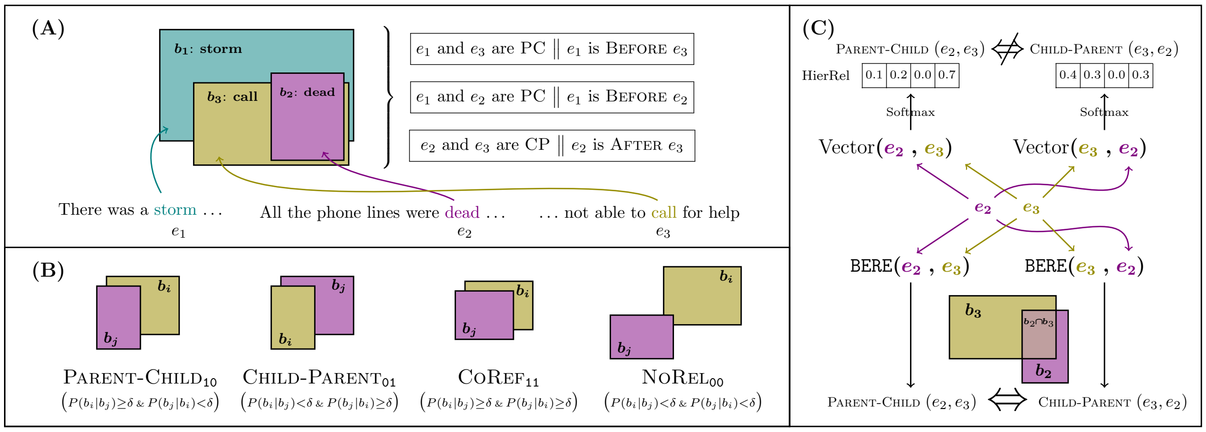


Figure 1: (A) BOX model architecture. (B) Mapping from box positions to event relations with classification rule below. (C) An example shows the fundamental difference between VECTOR and BOX model: BOX model will map events into consistent box representations regardless of the order; VECTOR model treats both cases separately and may not persist logical consistency.

indicator function. Now given batch $B$, BCE loss ($\mathbf{L}_1$) is defined as:

$$-\sum_{(i,j)\in B}\Big(y_0^{(i,j)}\log P(b_i|b_j)+(1-y_0^{(i,j)})\log\left(1-P(b_i|b_j)\right) + y_1^{(i,j)}\log P(b_j|b_i)+(1-y_1^{(i,j)})\log\left(1-P(b_j|b_i)\right)\Big).$$

**Pairwise loss** Motivated from previous papers using pairwise features to characterize relations, we also incorporate a pairwise box into our learning objective, and only in learning time, to encourage relevant boxes to be concentrated together. For the event-pair representation, two contextualized event embeddings $(h_i, h_j)$ are combined as $[h_i, h_j, h_i \odot h_j]$ where $\odot$ represents element-wise multiplication. Then, a multi-layer perceptron (MLP) is used to transform pairwise vectors to box representations $b_{ij}$. The pairwise features we use here are similar to (Zhou et al., 2020) except that we do not use subtraction in order to preserve symmetry between pairwise features of $(e_i, e_j)$ and $(e_j, e_i)$, i.e. $b_{ij} = b_{ji}$. For two related events, we enforce the intersection of corresponding boxes $b_i \cap b_j$ to be inside the pairwise box. For irrelevant event pairs such as having NOREL or VAGUE, their intersection and pairwise boxes are forced to be disjoint. The pairwise loss $\mathbf{L}_2$ is defined as:

$$-\sum_{i,j\in R^+}\log P(b_i\cap b_j|b_{ij}) - \sum_{i,j\in R^-}\log\Big(1-P(b_i\cap b_j|b_{ij})\Big)$$

where $R^-$ for irrelevant relations, such as NOREL and VAGUE, and $R^+$ stands for complement set of $R^-$, i.e. all the set of relations that indicates two events have some relation.

In the remainder of the paper, `BERE` refers to a model trained with loss $\mathbf{L}_1$ and `BERE-p` refers to a model trained with two losses $\mathbf{L}_1, \mathbf{L}_2$ combined.

Table 1: $F_1$ scores of `BERE` and `BERE-p`

| Model | $F_1$ Score | |
|---|---|---|
| | HiEve | MATRES |
| `BERE` | 0.4483 | 0.7069 |
| `BERE-p` | 0.4771 | 0.7105 |

## 4 Experiments

In this section, we describe datasets, baseline methods, and evaluation metrics. Lastly, we provide experimental results and a detailed analysis of logical consistency.

### 4.1 Experimental Setup

**Datasets** Experiments are conducted over three asymmetrical event relation extraction corpus, HiEve (Glavaš and Šnajder, 2014), MATRES (Ning et al., 2018), and Event StoryLine (ESL) (Caselli and Vossen, 2017). Since knowing $R_{e_1,e_2}$ (PARENT-CHILD or BEFORE) implies $R_{e_2,e_1}$ (CHILD-PARENT or AFTER), we expand our test set to be symmetrical for these reciprocal relations PARENT-CHILD, CHILD-PARENT, BEFORE and AFTER. See Appendix C for the dataset details.

**Baseline** We compare our `BERE`, `BERE-p` against the state-of-the-art event-event relation extraction model proposed by (Wang et al., 2020). This model utilizes RoBERTa with frozen parameters and further trains BiLSTM to represent text inputs into vector $h_i$ (for $e_i$) and then further utilizes MLP to represent pairwise representation $v_{ij}$ for $(e_i, e_j)$. Given $v_{ij}$, vector model (`Vector`) simply computes softmax over projected logits to produce probability for every possible relations. On top of

Table 2: $F_1$ scores with symmetric and conjunctive constraint violation results over original and symmetrical datasets. symm const. and conj const. denote symmetric and conjunctive constraint violations (%), respectively; H, M, and ESL are HiEve, MATRES, Event StoryLine datasets, respectively; single task(top) and joint task(bottom)

| Model | $F_1$ Score | | | | | | symmetry const. | | | conjunctive const. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Original data | | | Symmetric evaluation | | | | | | | | |
| | H | M | ESL | H | M | ESL | H | M | ESL | H | M | ESL |
| `Vector` | 0.4437 | **0.7274** | 0.2660 | 0.5385 | 0.7288 | 0.4444 | 22.49 | 35.81 | 60.9 | 4.91 | 2.53 | 6.1 |
| `BERE-p` | **0.4771** | 0.7105 | **0.3214** | **0.6064** | **0.7714** | **0.5379** | **0** | **0** | **0** | **0.71** | **0.30** | **0** |
| | Joint | | | | | | H+M | | | H+M | | |
| `Vector` | 0.4727 | **0.7291** | | 0.5517 | 0.7405 | | 86.77 | | | 6.17 | | |
| `Vector-c` | **0.5262** | 0.7068 | n/a | 0.6166 | 0.7106 | n/a | 46.03 | | n/a | 2.98 | | n/a |
| `BERE-p` | 0.5053 | 0.7125 | | **0.6261** | **0.7734** | | **0** | | | **1.84** | | |

this, as (Wang et al., 2020) showed that constraint injection improves performance, we also compare with the constraint-injected model (`Vector-c`).

For a fair comparison, we utilize the same RoBERTa + BiLSTM + MLP architecture for projecting event to box representation.

**Metrics** Following the same evaluation setting in previous works, we report the micro-$F_1$ score of all pairs, except VAGUE pairs, on MATRES (Han et al., 2019b; Wang et al., 2020). On HiEve and ESL, the micro-$F_1$ score of PARENT-CHILD and CHILD-PARENT pairs is reported (Glavaš and Šnajder, 2014; Wang et al., 2020).

### 4.2 Results and Discussion

**Impact of pairwise box, Table 1** We first show the results of the `BERE` and `BERE-p` with and without pairwise loss. The model with pairwise loss shows about 2.8 $F_1$ point improvement on HiEve and 1 $F_1$ point improvement on MATRES. It indicates that promoting the relevant event pairs to mingle together in the geometrical space is helpful and it is particularly useful when most of the relation extraction model encodes individual sentences independently.

**Vector-based vs. Box-based, Table 2** Table 2 shows a comparison of our box approach to the baseline with the ratio of symmetric and conjunctive constraint violations. Our approach clearly outperforms the baseline methods on symmetric evaluation with a gain of 6.79, 4.26, and 9.34 $F_1$ points on the single task over HiEve, MATRES, and ESL datasets, respectively and with a gain of 0.95 and 3.29 $F_1$ points on the joint task over HiEve and MATRES. The performance gains from asymmetrical to symmetrical datasets with `BERE-p` are much larger compared to the increase of `Vectors`. This demonstrates the `BERE-p` successfully capture symmetrical relations, while previous vector models do not. In addition, it is noteworthy that our method without constrained learning excels `Vector-c`, which is trained with constrained learning. This suggests that the inherent ability to model symmetrical relations helps satisfy the intertwined conjunctive constraints, thus producing more coherent results from a model. See Appendix F for constraint violation statistics for asymmetric dataset.

**Constraint Violation Analysis, Table 7 (Appendix)** We analyze constraint violations for each label from both HiEve and MATRES. For label pairs from the same dataset, our approach excels in almost every cases. For label pairs across datasets, our approach also shows fewer or similar levels of violation. This further indicates, without explicitly injecting constraints into objectives, our model can persist logical consistency among different relations.

## 5 Conclusion

We propose a novel event relation extraction method that utilizes box representation. The proposed method projects each event to a box representation which can model asymmetric relationships between entities. Utilizing this box representation, we design our relation extraction model to handle antisymmetry between events of $(e_i, e_j)$ and $(e_j, e_i)$ which previous vector models were not capable of. Thorough experiment on three datasets, we show that the proposed method not only free of antisymmetric constraint violations but also have drastically lower conjunctive constraint violations while maintaining similar or better performance in $F_1$. Our model shows that box representation can provide coherent classification across multiple event relations and opens up future research for box representations in event-to-event relation classification.

## References


Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. In *Proceedings of the Thirty-Fourth Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Anonymous. 2022. Modeling label space interactions in multi-label classification using box embeddings. In *Submitted to The Tenth International Conference on Learning Representations*. Under review.

Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers and Daniel Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 698–706, Honolulu, Hawaii. Association for Computational Linguistics.

Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. 2021. Box embeddings: An open-source library for representation learning using geometric structures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 203–211, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Lorraine Li, and Andrew McCallum. 2020. Improving local identifiability in probabilistic box embeddings. In *NeurIPS*.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2014. Constructing coherent event hierarchies from news stories. In *Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*, pages 34–38, Doha, Qatar. Association for Computational Linguistics.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019a. Joint event and temporal relation extraction with shared representations and structured prediction. In *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with shared representations and structured prediction. *CoRR*, abs/1909.05360.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006a. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006b. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 753–760, USA. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *EMNLP*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. 2021. Modeling fine-grained entity types with box embeddings. 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 2051–2064, Online. Association for Computational Linguistics.

Dhruvesh Patel, Shib Sankar Dasgupta, Michael Boratko, Xiang Li, Luke Vilnis, and Andrew McCallum. 2020. Representing joint hierarchies with box embeddings. In *Automated Knowledge Base Construction*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, page 75–80, USA. Association for Computational Linguistics.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the tarsqi toolkit. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, COLING '08, page 189–192, USA. Association for Computational Linguistics.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *ACL*. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 696–706. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.

Guangyu Zhou, Muhao Chen, Chelsea J T Ju, Zheng Wang, Jyun-Yu Jiang, and Wei Wang. 2020. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genomics and Bioinformatics*, 2(2). Lqaa015.

Table 3: An overview of dataset statistics.

| | HiEve | MATRES | ESL |
|---|---|---|---|
| # of Documents | | | |
| Train | 80 | 183 | 155 |
| Dev | - | 72 | 51 |
| Test | 20 | 20 | 52 |
| # of Pairs | | | |
| Train | 35001 | 6332 | 2238 |
| Test | 7093 | 827 | 619 |

Table 4: Mapped relation labels from ESL to HiEve

| Original labels in ESL | Mapped Labels |
|---|---|
| RISING_ACTION | PARENT-CHILD |
| CONTAINS | |
| BEFORE | |
| PRECONDITION | |
| ENDED_ON | |
| FALLING_ACTION | CHILD-PARENT |
| AFTER | |
| BEGUN_ON | |
| CAUSE | |
| OVERLAP | NOREL |

## A Hyperparameters

We utilize 768 dimensional pretrained RoBERTa model to compute word embeddings for events. models are trained for 100 epochs with AMSGrad optimizer and the learning rate is set to be 0.001. On HiEve and ESL, we sample NOREL in trainset using downsample ratio, which is fixed to 0.015, and the downsample ratio for valid and testset is fixed to 0.4. This is to encourage the models to learn and evaluate all types of relations that exist in the datasets when NOREL overwhelmingly represents the dataset. We use three weights, $\lambda_1, \lambda_2$, and $\lambda_3$, to balance our three learning objectives $L_1$, $L_2$, and $L_3$ (see Section 3.2 and Appendix B), in which the weights are selected between 0.1 and 1. A threshold $\delta$ for HiEve is selected between -0.4 and -0.3 and a threshold for MATRES is chosen between -0.7 and -0.6. We use wandb (Biewald, 2020) tool for efficient hyperparameter tuning.

## B Conjunctive Consistency Loss

With consistency requirements on conjunctive relations over temporal and subevent datasets (as shown in Table 5), we incorporate the loss function introduced by (Wang et al., 2020) into our box model to handle conjunctive constraints. Three events are grouped into three pairs, $(e1, e2), (e2, e3)$ and $(e1, e3)$, and the relation score for each class is calculated based on conditional probabilities and its binary logits. With the relation labels defined for each class (see Section 3.2), the relation score, $r(e_1, e_2)$, is calculated as:

$$r_i = y_0^{(i,j)} \log P(b_i|b_j) + y_1^{(i,j)} \log P(b_j|b_i) \quad (1)$$

where $y_0^{(i,j)} = I(P(b_i|b_j) \geq \delta)$ and $y_1^{(i,j)} = I(P(b_j|b_i) \geq \delta)$ and $y_0^{(i,j)}$ and $y_1^{(i,j)}$ are the first and second binary logits in relation label, respectively. Using this relation score, we now define the loss function for modeling conjunction constraints:

$$L_3 = \sum |L_{t1}| + \sum |L_{t2}|, \quad (2)$$

where the two transitivity losses are defined as

$$L_{t1} = \log r_{(e1,e2)} + \log r_{(e2,e3)} + \log r_{(e1,e3)}$$
$$L_{t2} = \log r_{(e1,e2)} + \log r_{(e2,e3)} + \log(1 - r_{(e1,e3)})$$

Table 6 presents the results of `BERE-p` combined with the above learning objective, denoted as `BERE-c`. Compared to the results from `BERE-p`, `BERE-c` shows a significantly smaller ratio of constraint violations than `BERE-p`, while sacrificing $F_1$ by $\sim$2 point from the performance with `BERE-p`.

## C Additional Details on the Data

Table 3 shows a brief summary of dataset statistics. HiEve consists of 100 articles and the narratives in news stories are represented as event hierarchies (Glavaš and Šnajder, 2014). The annotations include subevent and coreference relations. MATRES is a four-class temporal relation dataset, which contains 275 news articles drawn from a number of different sources (Ning et al., 2018). Event StoryLine (ESL) corpus is a dataset that contains 258 news documents and includes event temporal and subevent relations (Caselli and Vossen, 2017). ESL labels are mapped to the relation types that exist in the HiEve dataset as shown in Table 4.

For creating symmetrical dataset, we augment PARENT-CHILD and CHILD-PARENT (BEFORE and AFTER) pairs by their reversed relations CHILD-PARENT and PARENT-CHILD (AFTER and BEFORE), respectively.

## D Vector model architecture

Refer to Figure 2 for architecture of previous vector models.

Table 5: The induction table for conjunctive constraints on temporal and subevent relations (Wang et al., 2020). Given three events, $e1$, $e2$, and $e3$, the left-most column is $r_1(e_1, e_2)$ and the top row is $r_2(e_2, e_3)$.

| | PC | CP | CR | NR | BF | AF | EQ | VG |
|---|---|---|---|---|---|---|---|---|
| PC | PC, $\cancel{AF}$ | – | PC, -AF | -CP, -CR | BF, -CP, -CR | – | BF, -CP, -CR | – |
| CP | – | CP, -BF | CP, -BF | -PC, -CR | – | AF, -PC, -CR | AF, -PC, -CR | – |
| CR | PC, -AF | CP, -BF | CR, EQ | NR | BF, -CP, -CR | AF, -PC, -CR | EQ | VG |
| NR | -CP, -CR | -PC, -CR | NR | – | – | – | – | – |
| BF | BF, -CP, -CR | – | BF, -CP, -CR | – | BF, -CP, -CR | – | BF, –CP, –CR | -AF, -EQ |
| AF | – | AF, -PC, -CR | AF, -PC, -CR | – | – | AF, -PC, -CR | AF, -PC, -CR | -BF, -EQ |
| EQ | -AF | -BF | EQ | – | BF, -CP, -CR | AF, -PC, -CR | EQ | VG, -CR |
| VG | – | – | VG, -CR | – | -AF, -EQ | -BF, -EQ | VG | - |

Table 6: $F_1$ scores and the ratio of symmetric and conjunctive constraint violations of box model with constrained learning over `Eval-A` and `Eval-S`; `Eval-A` and `Eval-S` denote asymmetrical and symmetrical evaluation datasets, respectively. const. means constraint violations; results are on joint task.

| Model | $F_1$ Score | | | | symmetry const. (%) | | conjunctive const. (%) | |
|---|---|---|---|---|---|---|---|---|
| | `Eval-A` | | `Eval-S` | | `Eval-A` | `Eval-S` | `Eval-A` | `Eval-S` |
| | HiEve | MATRES | HiEve | MATRES | | | | |
| `BERE-p` | 0.5053 | 0.7125 | 0.6261 | 0.7734 | 0 | 0 | 3.12 | 1.84 |
| `BERE-c` | 0.5083 | 0.7021 | 0.6183 | 0.7562 | 0 | 0 | 0.39 | 0.19 |

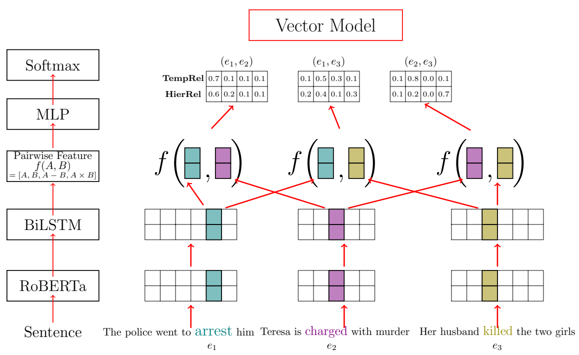


Figure 2: VECTOR model architecture.

Table 7: Constraint violation analysis over HiEve and MATRES. See Appendix B for conjunctive consistency requirements; PARENT-CHILD (PC), CHILD-PARENT (CP), COREF (CR), NOREL (NR), BEFORE (BF), AFTER (AF), EQUAL (EQ), VAGUE (VG); "-" means no existing constraint violations; constraint injected vector model (top), box model with using pairwise loss (bottom).

| **`Vector-c`** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **PC** | **CP** | **CR** | **NR** | **BF** | **AF** | **EQ** | **VG** |
| **PC** | 0.05 | - | 0.13 | 0.02 | 0.20 | - | 0.5 | - |
| **CP** | - | 0.33 | 0.46 | 0.01 | - | 0.25 | n/a | - |
| **CR** | 0.12 | 0.42 | 0.68 | 0.08 | 0.19 | 0.43 | n/a | 0.27 |
| **NR** | 0.01 | 0.03 | 0.13 | - | - | - | - | - |
| **BF** | 0.23 | - | 0.41 | - | 0.12 | - | 0.42 | 0.02 |
| **AF** | - | 0.33 | 0.30 | - | - | 0.01 | 0.13 | 0.05 |
| **EQ** | 0.00 | 0.50 | n/a | - | 0.25 | 0.00 | n/a | 0.50 |
| **VG** | - | - | 0.34 | - | 0.03 | 0.02 | n/a | - |
| **`BERE-p`** | | | | | | | | |
| | **PC** | **CP** | **CR** | **NR** | **BF** | **AF** | **EQ** | **VG** |
| **PC** | 0.13 | - | n/a | 0.00 | 0.16 | - | 0.30 | - |
| **CP** | - | 0.23 | n/a | 0 | - | 0.28 | 0.34 | - |
| **CR** | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| **NR** | 0.00 | 0.00 | n/a | - | - | - | - | - |
| **BF** | 0.24 | - | n/a | - | 0.08 | - | 0.32 | 0.00 |
| **AF** | - | 0.17 | n/a | - | - | 0.05 | 0.12 | 0.00 |
| **EQ** | 0.23 | 0.29 | n/a | - | 0.15 | 0.18 | n/a | 0.00 |
| **VG** | - | - | n/a | - | 0.00 | 0.00 | 0.13 | - |

## E Detailed analysis on conjunctive constraint violation

**Constraint Violation Analysis, Table 7** We further break down constraint violations for each label on HiEve and MATRES. The comparison of constraint violations between the vector model with constrained learning (`Vector-c`) and the box model without constrained learning (`BERE-p`) is shown in Table 7. "n/a" refers to no predictions and this frequently appears on COREF and EQUAL due to their sparsity in the corpus. Our approach shows a smaller ratio of constraint violations in most of the categories, with only a few exceptions. 2nd and 3rd quadrants (HiEve→MATRES and MATRES→HiEve) stand for cross-category, while 1st and 4th quadrants (HiEve→HiEve and MATRES→MATRES) stand for the same-category. Interestingly, our approach without any injected constraints shows a smaller or similar ratio to `Vector-c` in the cross-category as well as in the same-category. We calculated $r_c = \frac{\text{total \# of cross-category constraint violations}}{\text{total \# of cross-category event triplets}}$ and $r_s = \frac{\text{total \# of same-category constraint violations}}{\text{total \# of same-category event triplets}}$. $r_c$ for `Vector-c` is 6.26% and for `BERE-p` is 4.55% and $r_s$ for `Vector-c` is 0.05% and for `BERE-p` is 0.017%. This confirms the effectiveness of having boxes in handling logical consistency among different relations.

## F Symmetric and conjunctive constraint violations over origianl data

Table 8 shows the $F_1$ and symmetry and conjunctive constraint violation results over original dataset. The results of symmetry and conjunctive constraint violations confirm our expectation and exhibit a similar observation from Table 2.

# G Symmetry and Conjunction Consistency

We define symmetry and conjunction constraints of relations. Symmetry constraints indicate the event pair with flipping orders will have the reversed relation. For example, if $\mathbf{R}_{e_i,e_j}$ = PARENT-CHILD (BEFORE), then $\tilde{\mathbf{R}}_{e_j,e_i}$ = CHILD-PARENT (AFTER). Given any two events, $e_i$ and $e_j$, the symmetry consistency is defined as follows:

$$\bigwedge_{e_i,e_j \in \mathcal{E}, r \in \mathcal{R}_{\mathcal{S}}} \mathbf{R}_{(e_i,e_j)} \leftrightarrow \bar{\mathbf{R}}_{(e_j,e_i)} \tag{3}$$

where $r$ is the relation between events, the $\mathcal{E}$ is the set of all possible events and the $\mathcal{R}_{\mathcal{S}}$ is the set of relations, in which symmetry constraints hold.

Conjunctive constraints refer to the constraints that exist in the relations among any event triplet. The conjunctive constraints rules indicate that given any three event pairs, $(e_i, e_j)$, $(e_j, e_k)$, and $(e_i, e_k)$, then the relation of $(e_i, e_k)$ has to fall into the conjunction set specified based on $(e_i, e_j)$ and $(e_j, e_k)$ pairs (see Table 5). The conjunctive consistency can be defined as:

$$\bigwedge_{\substack{e_i,e_j,e_k \in \mathcal{E} \\ \mathbf{R}_1,\mathbf{R}_2 \in \mathcal{R}, \mathbf{R}_3 \in \mathcal{D}(\mathbf{R}_1,\mathbf{R}_2)}} \mathbf{R}_1(e_i,e_j) \wedge \mathbf{R}_2(e_j,e_k) \rightarrow \mathbf{R}_3(e_i,e_k)$$

$$\bigwedge_{\substack{e_i,e_j,e_k \in \mathcal{E} \\ \mathbf{R}_1,\mathbf{R}_2 \in \mathcal{R}, \mathbf{R}'_3 \notin \mathcal{D}(\mathbf{R}_1,\mathbf{R}_2)}} \mathbf{R}_1(e_i,e_j) \wedge \mathbf{R}_2(e_j,e_k) \rightarrow \neg\mathbf{R}'_3(e_i,e_k)$$

where the $\mathcal{E}$ is the set of all possible events, $r_1$ and $r_2$ are any possible relations exist in the set of all relations $\mathcal{R}$, $r_3$ is the relation, which is specified by $r_1$ and $r_2$ based on conjunctive induction table, and $\mathcal{D}$ is the set of all possible relations, in which $r_1$ and $r_2$ have no conflicts in between. The full explanation on symmetry and conjunction consistency can be found in Wang et al. (2020).

# H Related Work

## H.1 Event-Event Relation Extraction

This task has been traditionally modeled as a pairwise classification task with hand-engineered features and early attempts applied conventional machine learning methods, such as logistic regressions and SVM (Mani et al., 2006b; Verhagen et al., 2007; Verhagen and Pustejovsky, 2008). Later works utilized a structured learning (Ning et al., 2017) and neural methods to characterize relations. The neural methods have been shown effective and ensure logical consistency on relations through inference step (Dligach et al., 2017; Ning et al., 2018, 2019; Han et al., 2019a). More recent works proposed a constrained learning framework, which facilitates constraints during training time (Han et al., 2019b; Wang et al., 2020). Motivated by these works, we propose a box model to automatically handle inherent constraints without heavily relying on constrained learning across two different tasks.

## H.2 Box Embeddings

Box embeddings (Vilnis et al., 2018) were introduced as a shallow model to embed nodes of hierarchical graphs into euclidean space using hyper-rectangles, which were later extended to jointly embed multi-relational graphs and perform logical queries (Patel et al., 2020; Abboud et al., 2020). Recent works have successfully used box representations in conjunction with neural networks to represent input text for tasks like entity typing (Onoe et al., 2021), multi-label classification (Anonymous, 2022), natural language entailment (Chheda et al., 2021), etc. In all these works, the input is represented using a single box by transforming the output of the neural network into a hyper-rectangle. In this paper, we take this a step forward by representing the input event complex using multiple boxes. Our single box model represents each even in an input paragraph using a box and the pairwise box model adds on top of these, one box each for every pair of events (see section 3.2).

Table 8: $F_1$ scores with symmetric and conjunctive constraint violation results over original datasets. symm const. and conj const. denote symmetric and conjunctive constraint violations, respectively; H, M, and ESL are HiEve, MATRES, Event StoryLine datasets, respectively; single task(top) and joint task(bottom)

| Model | F1 Score | | | symmetry const. (%) | | | conjunctive const.(%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Original data | | | | | | | | |
| | H | M | ESL | H | M | ESL | H | M | ESL |
| `Vector` | 0.4437 | **0.7274** | 0.2660 | 22.73 | 38.63 | 56.7 | 5.66 | 0.69 | 9.4 |
| `BERE-p` | **0.4771** | 0.7105 | **0.3214** | **0** | **0** | **0** | **0.75** | **0.46** | **0** |
| Joint | | | | H+M | | | H+M | | |
| `Vector` | 0.4727 | **0.7291** | n/a | 23.04 | | n/a | 10.85 | | n/a |
| `Vector-c` | **0.5262** | 0.7068 | | 23.83 | | | 3.52 | | |
| `BERE-p` | 0.5053 | 0.7125 | | **0** | | | **3.12** | | |