Joint Constrained Learning for Event-Event Relation Extraction

Haoyu Wang[1], **Muhao Chen**[1], **Hongming Zhang**[2*] **& Dan Roth**[1]
[1]Department of Computer and Information Science, UPenn
[2]Department of Computer Science and Engineering, HKUST
{why16gzl, muhao, danroth}@seas.upenn.edu; hzhangal@cse.ust.hk

Abstract

Understanding natural language involves recognizing how multiple event mentions structurally and temporally interact with each other. In this process, one can induce event complexes that organize multi-granular events with temporal order and membership relations interweaving among them. Due to the lack of jointly labeled data for these relational phenomena and the restriction on the structures they articulate, we propose a joint constrained learning framework for modeling event-event relations. Specifically, the framework enforces logical constraints within and across multiple temporal and subevent relations by converting these constraints into differentiable learning objectives. We show that our joint constrained learning approach effectively compensates for the lack of jointly labeled data, and outperforms SOTA methods on benchmarks for both temporal relation extraction and event hierarchy construction, replacing a commonly used but more expensive global inference process. We also present a promising case study showing the effectiveness of our approach in inducing event complexes on an external corpus.[1]

1 Introduction

Human languages evolve to communicate about real-world events. Therefore, understanding events plays a critical role in natural language understanding (NLU). A key challenge to this mission lies in the fact that events are not just simple, standalone predicates. Rather, they are often described at different granularities and may form complex structures. Consider the example in Figure 1, where the description of a storm ($\boldsymbol{e}_1$) involves more fine-grained event mentions about people killed ($\boldsymbol{e}_2$), flights canceled ($\boldsymbol{e}_3$) and passengers affected ($\boldsymbol{e}_4$). Some of those mentions also follow strict temporal order ($\boldsymbol{e}_3$, $\boldsymbol{e}_4$ and $\boldsymbol{e}_5$). Our goal is to induce such an *event complex* that recognizes the membership of multi-granular events described in the text, as well as their temporal order. This is not only at the core of text understanding, but is also beneficial to various applications such as question answering (Khashabi et al., 2018), narrative prediction (Chaturvedi et al., 2017), timeline construction (Do et al., 2012a) and summarization (Daumé III and Marcu, 2006).

On Tuesday, there was a typhoon-strength (***e*$_1$*:storm***) in Japan. One man got (***e*$_2$*:killed***) and thousands of people were left stranded. Police said an 81-year-old man (***e*$_3$*:died***) in central Toyama when the wind blew over a shed, trapping him underneath. Later this afternoon, with the agency warning of possible tornadoes, Japan Airlines (***e*$_4$*:canceled***) 230 domestic flights, (***e*$_5$*:affecting***) 31,600 passengers.

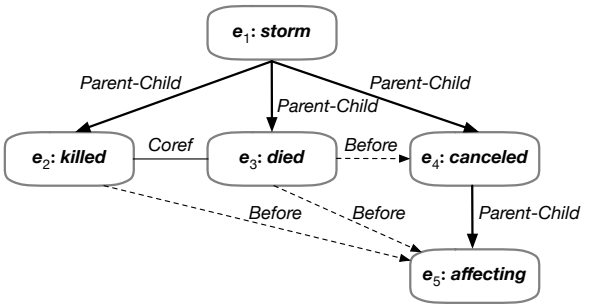

Figure 1: An example of an event complex described in the document. Bold arrows denote PARENT-CHILD relation; dotted arrows represent BEFORE relation; solid line represents two events are COREF to each other. For clarity, not all event mentions are shown in the figure.

Recently, significant research effort has been devoted to several event-event relation extraction tasks, such as event temporal relation (TempRel) extraction (Ning et al., 2018a, 2019) and subevent

* This work was done when the author was visiting the University of Pennsylvania.

[1]Our code is publicly available at https://cogcomp.seas.upenn.edu/page/publication_view/914.

relation extraction (Liu et al., 2018; Aldawsari and Finlayson, 2019). Addressing such challenging tasks requires a model to recognize the inherent connection between event mentions as well as their contexts in the documents. Accordingly, a few previous methods apply statistical learning methods to characterize the grounded events in the documents (Glavaš et al., 2014; Ning et al., 2017b, 2018c). Such methods often require designing various features to characterize the structural, discourse and narrative aspects of the events, which are costly to produce and are often specific to a certain task or dataset. More recent works attempted to use data-driven methods based on neural relation extraction models (Dligach et al., 2017; Ning et al., 2019; Han et al., 2019a,b) which refrain from feature engineering and offer competent performances.

While data-driven methods provide a general and tractable way for event-event relation extraction, their performance is restricted by the limited annotated resources available (Glavaš et al., 2014; Ning et al., 2018b). For example, the largest temporal relation extraction dataset MATRES (Ning et al., 2018b) only has 275 articles, which is far from enough for training a well-performing supervised model. The observation that relations and, in particular, event-event relations should be constrained by their logical properties (Roth and Yih, 2004; Chambers and Jurafsky, 2008), led to employing global inference to comply with transitivity and symmetry consistency, specifically on TempRel (Do et al., 2012b; Ning et al., 2017b; Han et al., 2019a). However, in an event complex, the logical constraints may globally apply to different task-specific relations, and form more complex conjunctive constraints. Consider the example in Figure 1: given that ***e2:died*** is BEFORE ***e3:canceled*** and ***e3:canceled*** is a PARENT event of ***e4:affecting***, the learning process should enforce ***e2:died*** BEFORE ***e4:affecting*** by considering the conjunctive constraints on both TempRel and subevent relations. While previous works focus on preserving logical consistency through (post-learning) inference or structured learning (Ning et al., 2017a), there was no effective way to endow neural models with the sense of global logical consistency during training. This is key to bridging the learning processes of TempRel and subevent relations, which is a research focus of this paper.

The *first* contribution of this work is proposing a joint constrained learning model for multi-faceted event-event relation extraction. The joint constrained learning framework seeks to regularize the model towards consistency with the logical constraints across both temporal and subevent relations, for which three types of consistency requirements are considered: *annotation consistency*, *symmetry consistency* and *conjunction consistency*. Such consistency requirements comprehensively define the interdependencies among those relations, essentially unifying the ordered nature of time and the topological nature of multi-granular subevents based on a set of declarative logic rules. Motivated by the logic-driven framework proposed by Li et al. (2019), the declarative logical constraints are converted into differentiable functions that can be incorporated into the learning objective for relation extraction tasks. Enforcing logical constraints across temporal and subevent relations is also a natural way to combine the supervision signals coming from two different datasets, one for each of the relation extraction tasks with a shared learning objective. Despite the scarce annotation for both tasks, the proposed method surpasses the SOTA TempRel extraction method on MATRES by relatively 3.27% in F_1; it also offers promising performance on the HiEve dataset for subevent relation extraction, relatively surpassing previous methods by at least 3.12% in F_1.

From the NLU perspective, the *second* contribution of this work lies in providing a general method for inducing an event complex that comprehensively represents the relational structure of several related event mentions. This is supported by the memberships vertically identified between multi-granular events, as well as the horizontal temporal reasoning within the event complex. As far as we know, this is different from all previous works that only formulated relations along a single axis. Our model further demonstrates the potent capability of inducing event complexes when evaluated on the RED dataset (O'Gorman et al., 2016).

2 Related Work

Various approaches have been proposed to extract event TempRels. Early effort focused on characterizing event pairs based on various types of semantic and linguistic features, and utilizing statistical learning methods, such as logistic regression (Mani et al., 2006; Verhagen and Pustejovsky, 2008) and SVM (Mirza and Tonelli, 2014), to capture the relations. Those methods typically require

extensive feature engineering, and do not comprehensively consider the contextual information and global constraints among event-event relations. Recently, data-driven methods have been developed for TempRel extraction, and have offered promising performance. Ning et al. (2019) addressed this problem using a system combining an LSTM document encoder and a Siamese multi-layer perceptron (MLP) encoder for temporal commonsense knowledge from TEMPROB (Ning et al., 2018a). Han et al. (2019a) proposed a bidirectional LSTM (BiLSTM) with structured prediction to extract TempRels. Both of these works incorporated global inference to facilitate constraints on TempRels.

Besides TempRels, a couple of efforts have focused on event hierarchy construction, a.k.a. subevent relation extraction. This task seeks to extract the hierarchy where each parent event contains child events that are described in the same document. To cope with this task, both Araki et al. (2014) and Glavaš and Šnajder (2014) introduced a variety of features and employed logistic regression models for classifying event pairs into subevent relations (PARENT-CHILD and CHILD-PARENT, coreference (COREF), and no relation (NOREL). Aldawsari and Finlayson (2019) further extended the characterization with more features on the discourse and narrative aspects. Zhou et al. (2020a) presented a data-driven method by fine-tuning a time duration-aware BERT (Devlin et al., 2019) on corpora of time mentions, and used the estimation of time duration to predict subevent relations.

Though previous efforts have been devoted to preserving logical consistency through inference or structured learning (Roth and Yih, 2004; Roth and tau Yih, 2007; Chang et al., 2008), this is difficult to do in the context of neural networks. Moreover, while it is a common strategy to combine multiple training data in multi-task learning (Lin et al., 2020), our work is distinguished by enhancing the learning process by pushing the model towards a coherent output that satisfies logical constraints across separate tasks.

3 Methods

In this section, we present the joint learning framework for event-event relation extraction. We start with the problem formulation (§3.1), followed by the techniques for event pair characterization (§3.2), constrained learning (§3.3) and inference (§3.4).

3.1 Preliminaries

A document D is represented as a sequence of tokens $D = [t_1, \cdots, e_1, \cdots, e_2, \cdots, t_n]$. Some of the tokens belong to the set of annotated event triggers, i.e., $\mathcal{E}_D = \{e_1, e_2, \cdots, e_k\}$, whereas the rest are other lexemes. The goal is to induce event complexes from the document, which is through extracting the multi-faceted event-event relations. Particularly, we are interested in two subtasks of relation extraction, corresponding to the label set $\mathcal{R} = \mathcal{R}_T \cup \mathcal{R}_H$. $\mathcal{R}_T$ thereof denotes the set of temporal relations defined in the literature (Ning et al., 2017b, 2018b, 2019; Han et al., 2019b), which contains BEFORE, AFTER, EQUAL, and VAGUE. To be consistent with previous studies (Ning et al., 2018b, 2019), the temporal ordering relations between two events are decided by the order of their starting time, without constraining on their ending time. $\mathcal{R}_H$ thereof denotes the set of relation labels defined in the subevent relation extraction task (Hovy et al., 2013; Glavaš et al., 2014), i.e., PARENT-CHILD, CHILD-PARENT, COREF and NOREL. Following the definitions by Hovy et al. (2013), an event e_1 is said to have a child event e_2 if e_1 is a collector event that contains a sequence of activities, where e_2 is one of these activities, and e_2 is spatially and temporally contained within e_1. Note that each pair of events can be annotated with one relation from each of $\mathcal{R}_H$ and $\mathcal{R}_T$ respectively, as the labels within each task-specific relation set are mutually exclusive.

Our learning framework first obtains the event pair representation that combines contextualized and syntactic features along with commonsense knowledge, and then use an MLP to get confidence scores for each relation in $\mathcal{R}$. The joint learning objective seeks to enforce the logical consistency of outputs for both TempRel and subevent relations. The overall architecture is shown in Figure 2.

3.2 Event Pair Representation

To characterize the event pairs in the document, we employ a neural encoder architecture which provides event representations from two groups of features. Specifically, the representation here incorporates the contextualized representations of the event triggers along with statistical commonsense knowledge from several knowledge bases. On top of the features that characterize an event pair (e_1, e_2), we use an MLP with $|\mathcal{R}|$ outputs to estimate the confidence score for each relation r,

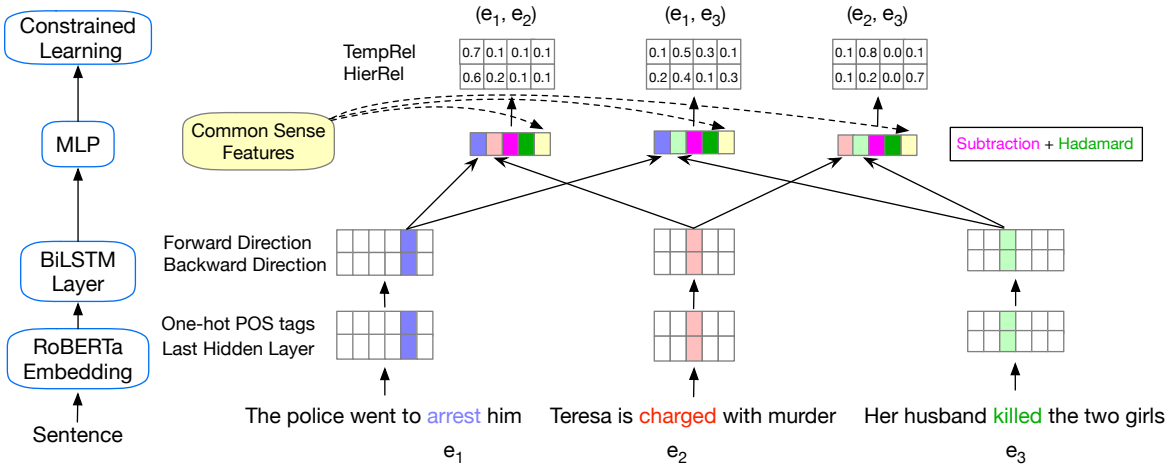

Figure 2: Model architecture. The model incorporates contextual features and commonsense knowledge to represent event pairs (§3.2). The joint learning enforces logical consistency on TempRel and subevent relations (§3.3).

denoted as $r_{(e_1,e_2)}$. Two separate softmax functions are then added to normalize the outputs for two task-specific label sets $\mathcal{R}_T$ and $\mathcal{R}_H$.

3.2.1 Contextualized Event Trigger Encoding

Given a document, we first use a pre-trained language model, RoBERTa (Liu et al., 2019), to produce the contextualized embeddings for all tokens of the entire document. The token embeddings are further concatenated with the one-hot vectors of POS (part-of-speech) tags, and fed into a BiLSTM. The hidden state of the last BiLSTM layer that is stacked on top of each event trigger e is therefore treated as the embedding representation of the event, denoted as h_e. For each event pair (e_1, e_2), the contextualized features are obtained as the concatenation of h_{e_1} and h_{e_2}, along with their element-wise Hadamard product and subtraction. This is shown to be a comprehensive way to model embedding interactions (Zhou et al., 2020b).

3.2.2 Commonsense Knowledge

We also incorporate the following sources of commonsense knowledge to characterize event pairs. Specifically, we first extract relevant knowledge from ConceptNet (Speer et al., 2017), which is a large-scale commonsense knowledge graph for commonsense concepts, entities, events and relations. A portion of the relations in ConceptNet that are relevant to our tasks include "HasSubevent", "HasFirstSubevent" and "HasLastSubevent" relations. From ConceptNet we extract around 30k pairs of event concepts labeled with the aforementioned relations, along with 30k randomly corrupted negative samples. We also incorporate commonsense knowledge from TEMPROB (Ning et al., 2018a). This provides prior knowledge of the temporal order that some events usually follow.

We use the event pairs from those knowledge bases to train two MLP encoders. Each takes the concatenated token embeddings of two event triggers as inputs, and is trained with *contrastive loss* to estimate the likelihood that if a relation holds. For subevent and temporal related commonsense knowledge, two MLPs are separately trained. After the encoders are well-trained, we fix their parameters and combine them as a black box that corresponds to "Common Sense Features" in Figure 2.

3.3 Joint Constrained Learning

Given the characterization of grounded event pairs from the document, we now define the learning objectives for relation prediction. The goal of learning is to let the model capture the data annotation, meanwhile regularizing the model towards consistency on logic constraints. Inspired by the logic-driven framework for consistency of neural models (Li et al., 2019), we specify three types of consistency requirements, i.e. *annotation consistency*, *symmetry consistency* and *conjunction consistency*. We hereby define the requirements with declarative logic rules, and show how we transform them into differentiable loss functions.

Annotation Consistency For labeled cases, we expect the model to predict what annotations specify. That is to say, if an event pair is annotated with

α \ β	PC	CP	CR	NR	BF	AF	EQ	VG
PC	PC, ¬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 1: The induction table for conjunctive constraints on temporal and subevent relations. Given the relations $\alpha(e_1, e_2)$ in the left-most column and $\beta(e_2, e_3)$ in the top row, each entry in the table includes all the relations and negations that can be deduced from their conjunction for e_1 and e_3, i.e. $\text{De}(\alpha, \beta)$. The abbreviations PC, CP, CR, NR, BF, AF, EQ and VG denote PARENT-CHILD, CHILD-PARENT, COREF, NOREL, BEFORE, AFTER, EQUAL and VAGUE, respectively. Vertical relations are in black, and TempRel are in blue. "–" denotes no constraints.

relation r, then the model should predict so:

$$\bigwedge_{e_1,e_2 \in \mathcal{E}_D} \top \rightarrow r(e_1, e_2).$$

To obtain the learning objective that preserves the annotation consistency, we use the product t-norm to get the learning objective of maximizing the probability of the true labels, by transforming to the negative log space to capture the inconsistency with the product t-norm. Accordingly, the annotation loss is equivalently defined as the cross entropy

$$L_A = \sum_{e_1,e_2 \in \mathcal{E}_D} -w_r \log r_{(e_1,e_2)},$$

in which w_r is the label weight that seeks to balance the loss for training cases of each relation r.

Symmetry Consistency Given any event pair (e_1, e_2), the grounds for a model to predict a relation $\alpha(e_1, e_2)$ to hold between them should also implies the hold of the converse relation $\bar{\alpha}(e_2, e_1)$. The logical formula is accordingly written as

$$\bigwedge_{e_1,e_2 \in \mathcal{E}_D,\, \alpha \in \mathcal{R}_S} \alpha(e_1, e_2) \leftrightarrow \bar{\alpha}(e_2, e_1),$$

where the $\mathcal{R}_S$ is the set of relations enforcing the symmetry constraint. Particularly for the TempRel extraction task, $\mathcal{R}_S$ contains a pair of reciprocal relations BEFORE and AFTER, as well as two reflexive ones EQUAL and VAGUE. Similarly, the subevent relation extraction task adds reciprocal relations PARENT-CHILD and CHILD-PARENT as well as reflexive ones COREF and NOREL.

Using the product t-norm and transformation to the negative log space as before, we have the symmetry loss:

$$L_S = \sum_{e_1,e_2 \in \mathcal{E}, \alpha \in \mathcal{R}_S} |\log \alpha_{(e_1,e_2)} - \log \bar{\alpha}_{(e_2,e_1)}|.$$

Conjunction Consistency This set of constraints are applicable to any three related events e_1, e_2 and e_3. If we group the events into three pairs, namely $(e_1, e_2), (e_2, e_3)$ and (e_1, e_3), the relation definitions mandate that not all of the possible assignments to these three pairs are allowed. More specifically, if two relations $\alpha(e_1, e_2)$ and $\beta(e_2, e_3)$ apply to the first two pairs of events, then the conjunction consistency may enforce the following two conjunctive rules.

In the first rule, the conjunction of the first two relations infers the hold of another relation γ between the third event pair (e_1, e_3), namely

$$\bigwedge_{\substack{e_1,e_2,e_3 \in \mathcal{E}_D \\ \alpha,\beta \in \mathcal{R},\, \gamma \in \text{De}(\alpha,\beta)}} \alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \gamma(e_1, e_3).$$

$\text{De}(\alpha, \beta)$ thereof is a set composed of all relations from $\mathcal{R}$ that do not conflict with α and β, which is a subset of the deductive closure (Stine, 1976) of the conjunctive clause for these two relations. A special case that the above formula expresses is a (task-specific) transitivity constraint, where $\alpha = \beta = \gamma$ present the same transitive relation.

Another condition could also hold, where the former two relations always infer the negation of a certain relation δ on (e_1, e_3), for which we have

$$\bigwedge_{\substack{e_1,e_2,e_3 \in \mathcal{E}_D \\ \alpha,\beta \in \mathcal{R},\, \delta \notin \text{De}(\alpha,\beta)}} \alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \neg\delta(e_1, e_3).$$

Table 1 is an induction table that describes all the conjunctive rules for relations in $\mathcal{R}$. To illustrate the conjunction consistency requirement (see the orange cell in Table 1), assume that (e_1, e_2) and (e_2, e_3) are respectively annotated with BEFORE and PARENT-CHILD. Then the two conjunctive formulae defined above infer that we have the relation

BEFORE hold on (e_1, e_3), whereas we should not have CHILD-PARENT hold.

Similar to the other consistency requirements, the loss function dedicated to the conjunction consistency is derived as follows:

$$L_C = \sum_{\substack{e_1,e_2,e_3 \in \mathcal{E}_D, \\ \alpha,\beta \in \mathcal{R}, \gamma \in \text{De}(\alpha,\beta)}} |L_{t_1}| + \sum_{\substack{e_1,e_2,e_3 \in \mathcal{E}_D, \\ \alpha,\beta \in \mathcal{R}, \delta \notin \text{De}(\alpha,\beta)}} |L_{t_2}|,$$

where the two terms of triple losses are defined as

$$L_{t_1} = \log \alpha_{(e_1,e_2)} + \log \beta_{(e_2,e_3)} - \log \gamma_{(e_1,e_3)}$$
$$L_{t_2} = \log \alpha_{(e_1,e_2)} + \log \beta_{(e_2,e_3)} - \log(1 - \delta_{(e_1,e_3)})$$

It is noteworthy that modeling the conjunctive consistency is key to the combination of two different event-event relation extraction tasks, as this general consistency requirement can be enforced between both TempRels and subevent relations.

Joint Learning Objective After expressing the logical consistency requirements with different terms of cross-entropy loss, we combine all of those into the following joint learning objective loss

$$L = L_A + \lambda_S L_S + \lambda_C L_C.$$

The λ's are non-negative coefficients to control the influence of each loss term. Note that since the consistency requirements are defined on both temporal and subevent relations, the model therefore seamlessly incorporates both event-event relation extraction tasks with a shared learning objective. In this case, the learning process seeks to unify the ordered nature of time and the topological nature of subevents, therefore supporting the model to comprehensively understand the event complex.

3.4 Inference

To support task-specific relation extraction, i.e. extracting either a TempRel or a subevent relation, our framework selects the relation r with highest confident score $r_{(e_1,e_2)}$ from either of $\mathcal{R}_T$ and $\mathcal{R}_H$. When it comes to extracting event complexes with both types of relations, the prediction of subevent relations has higher priority. The reason lies in the fact that a relation in $\mathcal{R}_H$, except for NOREL, always implies a TempRel, yet there is not a single TempRel that necessitates a subevent relation.

We also incorporate ILP in the inference phase to further ensure the logical consistency in predicted results. Nevertheless, we show in experiments that a well-trained constrained learning model may not additionally require global inference (§4.5).

4 Experiments

In this section, we present the experiments on event-event relation extraction. Specifically, we conduct evaluation for TempRel and subevent relation extraction based on two benchmark datasets (§4.1-§4.4). To help understand the significance of each model component in the framework, we also give a detailed ablation study (§4.5). Finally, a case study on the RED dataset is described to demonstrate the capability of inducing event complexes (§4.6).

4.1 Datasets

Since there is not a large-scale dataset that amply annotates for both TempRel and subevent relations, we evaluate the joint training and prediction of both categories of relations on two separate datasets. Specifically, we use MATRES (Ning et al., 2018b) for TempRel extraction and HiEve (Glavaš et al., 2014) for subevent relation extraction.

MATRES is a new benchmark dataset for TempRel extraction, which is developed from TempEval3 (UzZaman et al., 2013). It annotates on top of 275 documents with TempRels BEFORE, AFTER, EQUAL, and VAGUE. Particularly, the annotation process of MATRES has defined four axes for the actions of events, i.e. *main*, *intention*, *opinion*, and *hypothetical* axes. The TempRels are considered for all event pairs on the same axis and within a context of two adjacent sentences. The labels are decided by comparing the starting points of the events. The multi-axis annotation helped MATRES to achieve a high IAA of 0.84 in Cohen's Kappa.

The HiEve corpus is a news corpus that contains 100 articles. Within each article, annotations are given for both subevent and coreference relations. The HiEve adopted the IAA measurement proposed for TempRels by (UzZaman and Allen, 2011), resulting in 0.69 F_1.

In addition to these two datasets, we also present a case study on an updated version of the RED dataset (O'Gorman et al., 2016). This dataset contains 35 news articles with annotations for event complexes that contain both membership relations and TempRels. Since small dataset is not sufficient for training, we use it only to demonstrate our method's capability of inducing event complexes on data that are external to training.

We briefly summarize the data statistics for HiEve, MATRES, and RED dataset in Table 3.

Model	P	R	F_1
CogCompTime (Ning et al., 2018c)	0.616	0.725	0.666
Perceptron (Ning et al., 2018b)	0.660	0.723	0.690
BiLSTM+MAP (Han et al., 2019b)	-	-	0.755
LSTM+CSE+ILP (Ning et al., 2019)	0.713	0.821	0.763
Joint Constrained Learning (ours)	**0.734**	**0.850**	**0.788**

Table 2: TempRel extraction results on MATRES. Precision and recall are not reported by (Han et al., 2019b).

	HiEve	MATRES	RED
# of Documents			
Train	80	183	-
Dev	-	72	-
Test	20	20	35
# of Pairs			
Train	35001	6332	-
Test	7093	827	1718

Table 3: Data statistics of HiEve, MATRES, and RED.

4.2 Baselines and Evaluation Protocols

On MATRES, we compare with four baseline methods. Ning et al. (2018b) present a baseline method based on a set of linguistic features and an averaged perceptron classifier (Perceptron). Han et al. (2019b) introduce a BiLSTM model that incorporates MAP inference (BiLSTM+MAP). Ning et al. (2019) present the SOTA data-driven method incorporating ILP and commonsense knowledge from TEMPROB with LSTM (LSTM+CSE+ILP). We also compare with the CogCompTime system (Ning et al., 2018c). On HiEve[2], we compare with a structured logistic regression model (StructLR, Glavaš and Šnajder 2014) and a recent data-driven method based on fined-tuning a time duration-aware BERT on large time-related web corpora (TACOLM, Zhou et al. 2020a).

MATRES comes with splits of 183, 72 and 20 documents respectively used for training, development and testing. Following the settings in previous work (Ning et al., 2019; Han et al., 2019b), we report the micro-average of precision, recall and F1 scores on test cases. On HiEve, we use the same evaluation setting as Glavaš and Šnajder (2014) and Zhou et al. (2020a), leaving 20% of the documents out for testing. The results in terms of F_1 of PARENT-CHILD and CHILD-PARENT and the micro-average of them are reported. Note that in the previous setting by Glavaš and Šnajder (2014), the relations are only considered for event pairs (e_1, e_2) where e_1 appears before e_2 in the document. We also follow Glavaš and Šnajder (2014) to populate the annotations by computing the transitive closure of COREF and subevent relations.

[2]Despite carefully following the details described in (Aldawsari and Finlayson, 2019) and communicating with the authors, we were not able to reproduce their results. Therefore, we choose to compare with other methods.

Model	F_1 score PC	CP	Avg.
StructLR (Glavaš et al., 2014)	0.522	**0.634**	0.577
TACOLM (Zhou et al., 2020a)	0.485	0.494	0.489
Joint Constrained Learning (ours)	**0.625**	0.564	**0.595**

Table 4: Subevent relation extraction results on HiEve. PC, CP and Avg. respectively denote PARENT-CHILD, CHILD-PARENT and their micro-average.

4.3 Experimental Setup

To encode the tokens of each document, we employ the officially released 768 dimensional RoBERTa (Liu et al., 2019), which is concatenated with 18 dimensional one-hot vectors representing the tokens' POS tags. On top of those embeddings, the hidden states of the trainable BiLSTM are 768 dimensional, and we only apply one layer of BiLSTM. Since the TempRel extraction and subevent relation extraction tasks are considered with two separate sets of labels, we use two separate softmax functions for normalizing the outputs for each label set from the single MLP. For all the MLPs we employ one hidden layer each, whose dimensionality is set to the average of the input and output space following convention (Chen et al., 2018).

We use AMSGrad (Reddi et al., 2018) to optimize the parameters, with the learning rate set to 0.001. Label weights in the annotation loss L_A is set to balance among training cases for different relations. The coefficients λ_S and λ_D in the learning objective function are both fixed to 0.2. Training is limited to 80 epochs, which is sufficient to converge.

4.4 Results

In Table 2 we report the TempRel extraction results on MATRES. Among the baseline methods, Ning et al. (2019) offer the best performance in terms of F_1 by incorporating an LSTM with global inference and commonsense knowledge. In contrast, the proposed joint constrained learning framework surpasses the best baseline method by a relative gain of 3.27% in F_1, and excels in terms of both precision and recall. While both methods ensure logical constraints in learning or inference phases,

Model	SUBEVENT			TEMPREL		
	P	R	F_1	P	R	F_1
Single-task Training	32.5	**73.1**	45.0	67.7	80.3	73.5
Joint Training	50.4	43.1	46.5	68.4	82.0	74.6
+ Task-specific constrained learning	51.6	59.7	55.4	71.3	82.7	76.6
+ Cross-task constrained learning	51.1	67.0	58.0	72.2	83.8	77.6
+ Commonsense knowledge	56.9	61.6	59.2	73.3	84.2	78.4
+ Global inference (ILP)	**57.4**	61.7	**59.5**	**73.4**	**85.0**	**78.8**
All but constrained learning	54.2	41.8	47.2	72.1	80.8	76.2

Table 5: Ablation study results (§4.5). The results on HiEve are the micro-average of PARENT-CHILD and CHILD-PARENT. Results in the middle group are achieved by incrementally adding the corresponding model components. The gray-scaled row shows the results of the complete model.

the improvement by the proposed method is largely due to the joint constraints combining both TempRel and subevent relations. Learning to capture subevent relations from an extrinsic resource simultanously offer auxiliary supervision signals to improve the comprehension on TempRel, even though the resources dedicated to the later is limited.

The results in Table 4 for subevent relation extraction exhibit similar observation. Due to scarcer annotated data, the pure data-driven baseline method (TACOLM) falls behind the statistical learning one (i.e. StructLR) with comprehensively designed features. However, our model successfully complements the insufficient supervision signals, partly by incorporating linguistic and commonsense knowledge. More importantly, while our model is able to infer TempRel decently, the global consistency ensured by cross-task constraints naturally makes up for the originally weak supervision signals for subevent relations. This fact leads to promising results, drastically surpassing TACOLM with a relative gain of 21.4% in micro-average F_1, and outperforming StructLR by $\sim$3% relatively.

In general, the experiments here show that the proposed joint constrained learning approach effectively combines the scarce supervision signals for both tasks. Understanding the event complex by unifying the ordered nature of time and the topological nature of multi-granular subevents, assists the comprehension on both TempRel and memberships among multi-granular events.

4.5 Ablation Study

To help understand the model components, we conduct an ablation study and report the results in Table 5. Starting from the vanilla single-task BiLSTM model with only RoBERTa features, changing to joint training both tasks with only annotation brings along 1.1-1.5% of absolute gain in F_1. Incorporating task-specific constraints to learning for relations only in $\mathcal{R}_T$ or $\mathcal{R}_H$ notably brings up the F_1 2.0-8.9%, whereas the cross-task constraints bring along an improvement of 1.0-2.6% in F_1. This indicates that the global consistency ensured within and across TempRel and subevent relations is important for enhancing the comprehension for both categories of relations. The commonsense knowledge leads to another 0.8-1.2% of improvement. Lastly, global inference does not contribute much to the performance in our setting, which indicates that the rest model components are already sufficient to preserve global consistency through joint constrained learning.

To compare both ways of ensuring logical consistency, we also report a set of results in the last row of Table 5, where constrained learning is removed and only global inference is used to cope with consistency requirements in prediction. As expected, this leads to significant performance drop of 2.6-12.3% in F_1. This fact implies that ensuring the logical consistency in the learning phase is essential, in terms of both complementing task-specific training and enhancing the comprehension of event complex components.

4.6 Case Study on the RED Dataset

We use the RED dataset (2019 updated version) to further evaluate our model trained on MATRES and HiEve for inducing complete event complexes, as well as to show the model's generalizability to an external validation set. Since the labels of RED are defined differently from those in the datasets we train the model on, Table 6 shows the details about how some RED lables are mapped to MATRES and HiEve labels. Other event-event relations in RED are mapped to VAGUE or NOREL according to their relation types, and the relations annotated between entities are discarded. To obtain the event

Original labels in RED	Mapped labels
BEFORE, BEFORE/CAUSES, BEFORE/PRECONDITION, ENDS-ON, OVERLAP/PRECONDITION	BEFORE
SIMULTANEOUS	EQUAL
OVERLAP, REINITIATES	VAGUE
CONTAINS, CONTAINS-SUBEVENT	PARENT-CHILD & BEFORE
BEGINS-ON	AFTER

Table 6: Mapping from relations annotated in the RED dataset to the relations studied in this work.

A (***e1:convoy***) of 280 Russian trucks (***e2:headed***) for Ukraine, which Moscow says is (***e3:carrying***) relief goods for war-weary civilians, has suddenly (***e4:changed***) course, according to a Ukrainian state news agency.

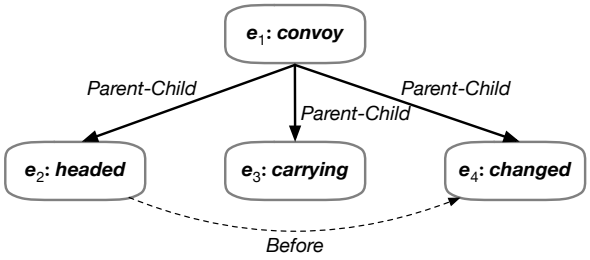

Figure 3: An example of an event complex extracted from a document in RED. Bold arrows denote the PARENT-CHILD relation, and dotted arrows represent the BEFORE relation.

complexes, as stated in §3.4, prediction of subevent relations is given higher priority than that of TempRels. In this way, our model achieves 0.72 F_1 on TempRel extraction and 0.54 F_1 on subevent relation extraction.

Here we give an example of an event complex extracted from the RED dataset in Figure 3, using our joint constrained learning method.

5 Conclusion

We propose a joint constrained learning framework for extracting event complexes from documents. The proposed framework bridges TempRel and subevent relation extraction tasks with a comprehensive set of logical constraints, which are enforced during learning by converting them into differentiable objective functions. On two benchmark datasets, the proposed method outperforms SOTA statistical learning methods and data-driven methods for each task, without using data that is jointly annotated with the two classes of relations. It also presents promising event complex extraction results on RED that is external to training. Thus, our work shows that the global consistency of the event complex significantly helps understanding both temporal order and event membership. For future work, we plan to extend the framework towards an end-to-end system with event extraction. We also seek to extend the conjunctive constraints along with event argument relations.

Acknowledgement

We appreciate the anonymous reviewers for their insightful comments. Also, we would like thank Jennifer Sheffield and other members from the UPenn Cognitive Computation Group for giving suggestions to improve the manuscript.

This research is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the BETTER Program, and by contract FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy. Association for Computational Linguistics.

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-2014)*, Reykjavik, Iceland. European Languages Resources Association (ELRA).

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.

Ming-Wei Chang, Lev Ratinov, Nicholas Rizzolo, and Dan Roth. 2008. Learning and Inference with Constraints. In *Proc. of the Conference on Artificial Intelligence (AAAI)*.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614.

Muhao Chen, Changping Meng, Gang Huang, and Carlo Zaniolo. 2018. Neural article pair modeling for wikipedia sub-article matching. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 3–19. Springer.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. 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.

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.

Quang Do, Wei Lu, and Dan Roth. 2012a. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687, Jeju Island, Korea. Association for Computational Linguistics.

Quang Do, Wei Lu, and Dan Roth. 2012b. Joint Inference for Event Timeline Construction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3678–3683, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhengzhong Liu, Teruko Mitamura, and Eduard Hovy. 2018. Graph based decoding for event sequencing and coreference resolution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3645–3657, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. 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.

Paramita Mirza and Sara Tonelli. 2014. Classifying temporal relations with simple features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317, Gothenburg, Sweden. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017a. A Structured Learning Approach to Temporal Relation Extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1038–1048, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017b. 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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. Improving temporal relation extraction with a globally acquired statistical resource. 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 841–851, New Orleans, Louisiana. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. 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.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018c. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*.

Dan Roth and Scott Yih. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*, pages 1–8. Association for Computational Linguistics.

Dan Roth and Wen tau Yih. 2007. Global Inference for Entity and Relation Identification via a Linear Programming Formulation.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Gail C Stine. 1976. Skepticism, relevant alternatives, and deductive closure. *Philosophical Studies*, 29(4):249–261.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356, Portland, Oregon, USA. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the TARSQI toolkit. In *Coling 2008: Companion volume: Demonstrations*, pages 189–192, Manchester, UK. Coling 2008 Organizing Committee.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020a. Temporal Common Sense Acquisition with Minimal Supervision. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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