ACL'18

Joint Reasoning for Temporal and Causal Relations

Qiang Ning,[1] **Zhili Feng,**[2] **Hao Wu,**[3] **Dan Roth**[1,3]
Department of Computer Science
[1]University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]University of Wisconsin-Madison, Madison, WI 53706, USA
[3]University of Pennsylvania, Philadelphia, PA 19104, USA
qning2@illinois.edu, zfeng49@cs.wisc.edu, {haowu4,danroth}@seas.upenn.edu

Abstract

Understanding temporal and causal relations between events is a fundamental natural language understanding task. Because a cause must occur earlier than its effect, temporal and causal relations are closely related and one relation often dictates the value of the other. However, limited attention has been paid to studying these two relations jointly. This paper presents a joint inference framework for them using constrained conditional models (CCMs). Specifically, we formulate the joint problem as an integer linear programming (ILP) problem, enforcing constraints that are inherent in the nature of time and causality. We show that the joint inference framework results in statistically significant improvement in the extraction of both temporal and causal relations from text.[1]

1 Introduction

Understanding events is an important component of natural language understanding. An essential step in this process is identifying relations between events, which are needed in order to support applications such as story completion, summarization, and timeline construction.

Among the many relation types that could exist between events, this paper focuses on the joint extraction of temporal and causal relations. It is well known that temporal and causal relations interact with each other and in many cases, the decision of one relation is made primarily based on evidence from the other. In Example 1, identifying the temporal relation between ***e1:died*** and ***e2:exploded*** is in fact a very hard case: There are no explicit temporal markers (e.g., "before", "after", or "since"); the events are in separate sentences so their syntactic connection is weak; although the occurrence time of ***e2:exploded*** is given (i.e., Friday) in text, it is not given for ***e1:died***. However, given the causal relation, ***e2:exploded*** caused ***e1:died***,it is clear that ***e2:exploded*** happened before ***e1:died***. The temporal relation is dictated by the causal relation.

Ex 1: Temporal relation dictated by causal relation.
More than 10 people ***(e1:died)*** on their way to the nearest hospital, police said. A suicide car bomb ***(e2:exploded)*** on Friday in the middle of a group of men playing volleyball in northwest Pakistan.
Since ***e2:exploded*** *is the reason of* ***e1:died****, the temporal relation is thus e2 being before e1.*

Ex 2: Causal relation dictated by temporal relation.
Mir-Hossein Moussavi ***(e3:raged)*** after government's efforts to ***(e4:stifle)*** protesters.
Since ***e3:raged*** *is temporally after* ***e4:stifle****, e4 should be the cause of e3.*

On the other hand, causal relation extraction can also benefit from knowing temporal relations. In Example 2, it is unclear whether the government stifled people because people raged, or people raged because the government stifled people: both situations are logically reasonable. However, if we account for the temporal relation (that is, ***e4:stifle*** happened before ***e3:raged***), it is clear that ***e4:stifle*** is the cause and ***e3:raged*** is the effect. In this case, the causal relation is dictated by the temporal relation.

The first contribution of this work is proposing a joint framework for **T**emporal and **C**ausal **R**easoning (TCR), inspired by these examples. Assuming the availability of a temporal extraction system and a causal extraction system, the proposed joint framework combines these two using a constrained conditional model (CCM) (Chang et al., 2012) framework, with an integer linear pro-

[1] The dataset and code used in this paper are available at http://cogcomp.org/page/publication_view/835

gramming (ILP) objective (Roth and Yih, 2004) that enforces declarative constraints during the inference phase. Specifically, these constraints include: (1) A cause must temporally precede its effect. (2) Symmetry constraints, i.e., when a pair of events, (A, B), has a temporal relation r (e.g., *before*), then (B, A) must have the reverse relation of r (e.g., *after*). (3) Transitivity constraints, i.e., the relation between (A, C) must be temporally consistent with the relation derived from (A, B) and (B, C). These constraints originate from the one-dimensional nature of time and the physical nature of causality and build connections between temporal and causal relations, making CCM a natural choice for this problem. As far as we know, very limited work has been done in joint extraction of both relations. Formulating the joint problem in the CCM framework is novel and thus the first contribution of this work.

A key obstacle in jointly studying temporal and causal relations lies in the absence of jointly annotated data. The second contribution of this work is the development of such a jointly annotated dataset which we did by augmenting the Event-Causality dataset (Do et al., 2011) with dense temporal annotations. This dataset allows us to show statistically significant improvements on both relations via the proposed joint framework.

This paper also presents an empirical result of improving the temporal extraction component. Specifically, we incorporate explicit time expressions present in the text and high-precision knowledge-based rules into the ILP objective. These sources of information have been successfully adopted by existing methods (Chambers et al., 2014; Mirza and Tonelli, 2016), but were never used within a global ILP-based inference method. Results on TimeBank-Dense (Cassidy et al., 2014), a benchmark dataset with temporal relations only, show that these modifications can also be helpful within ILP-based methods.

2 Related Work

Temporal and causal relations can both be represented by directed acyclic graphs, where the nodes are events and the edges are labeled with either *before, after*, etc. (in temporal graphs), or *causes* and *caused by* (in causal graphs). Existing work on *temporal* relation extraction was initiated by (Mani et al., 2006; Chambers et al., 2007; Bethard et al., 2007; Verhagen and Pustejovsky, 2008), which formulated the problem as that of learning a classification model for determining the label of each edge locally (i.e., *local* methods). A disadvantage of these early methods is that the resulting graph may break the symmetric and transitive constraints. There are conceptually two ways to enforce such graph constraints (i.e., *global* reasoning). CAEVO (Chambers et al., 2014) grows the temporal graph in a multi-sieve manner, where predictions are added sieve-by-sieve. A graph closure operation had to be performed after each sieve to enforce constraints. This is solving the global inference problem greedily. A second way is to perform exact inference via ILP and the symmetry and transitivity requirements can be enforced as ILP constraints (Bramsen et al., 2006; Chambers and Jurafsky, 2008; Denis and Muller, 2011; Do et al., 2012; Ning et al., 2017).

Ex 3: Global considerations are needed when making local decisions.

The FAA on Friday ***(e5:announced)*** it will close 149 regional airport control towers because of forced spending cuts. Before Friday's ***(e6:announcement)***, it ***(e7:said)*** it would consider keeping a tower open if the airport convinces the agency it is in the ”national interest” to do so.

We adopt the ILP approach in the temporal component of this work for two reasons. First, as we show later, it is straightforward to build a joint framework with both temporal and causal relations as an extension of it. Second, the relation between a pair of events is often determined by the relations among other events. In Ex 3, if a system is unaware of $(e5, e6)$=*simultaneously* when trying to make a decision for $(e5, e7)$, it is likely to predict that $e5$ is *before* $e7$[2]; but, in fact, $(e5, e7)$=*after* given the existence of $e6$. Using global considerations is thus beneficial in this context not only for generating globally consistent temporal graphs, but also for making more reliable pairwise decisions.

Prior work on *causal* relations in *natural language text* was relatively sparse. Many causal extraction work in other domains assumes the existence of ground truth timestamps (e.g., (Sun et al., 2007; Güler et al., 2016)), but gold timestamps rarely exist in natural language text. In NLP, people have focused on causal relation identification using lexical features or discourse relations. For

[2] Consider the case that “The FAA ***e5:announced***...it ***e7:said*** it would...”. Even humans may predict that $e5$ is *before* $e7$.

example, based on a set of explicit causal discourse markers (e.g., “because”, “due to”, and “as a result”), Hidey and McKeown (2016) built parallel Wikipedia articles and constructed an open set of implicit markers called AltLex. A classifier was then applied to identify causality. Dunietz et al. (2017) used the concept of construction grammar to tag causally related clauses or phrases. Do et al. (2011) considered global statistics over a large corpora, the cause-effect association (CEA) scores, and combined it with discourse relations using ILP to identify causal relations. These work only focused on the causality task and did not address the temporal aspect.

However, as illustrated by Examples 1-2, temporal and causal relations are closely related, as assumed by many existing works (Bethard and Martin, 2008; Rink et al., 2010). Here we argue that being able to capture both aspects in a joint framework provides a more complete understanding of events in natural language documents. Researchers have started paying attention to this direction recently. For example, Mostafazadeh et al. (2016b) proposed an annotation framework, CaTeRs, which captured both temporal and causal aspects of event relations in common sense stories. CATENA (Mirza and Tonelli, 2016) extended the multi-sieve framework of CAEVO to extracting both temporal and causal relations and exploited their interaction through post-editing temporal relations based on causal predictions. In this paper, we push this idea forward and tackle the problem in a joint and more principled way, as shown next.

3 Temporal and Causal Reasoning

In this section, we explain the proposed joint inference framework, **T**emporal and **C**ausal **R**easoning (TCR). To start with, we focus on introducing the temporal component, and clarify how to design the transitivity constraints and how to enforce other readily available prior knowledge to improve its performance. With this temporal component already explained, we further incorporate causal relations and complete the TCR joint inference framework. Finally, we transform the joint problem into an ILP so that it can be solved using off-the-shelf packages.

3.1 Temporal Component

Let $\mathcal{R}_T$ be the label set of temporal relations and $\mathcal{E}$ and $\mathcal{T}$ be the set of all events and the set of all time expressions (a.k.a. timex) in a document. For notation convenience, we use $\mathcal{EE}$ to represent the set of all event-event pairs; then $\mathcal{ET}$ and $\mathcal{TT}$ have obvious definitions. Given a pair in $\mathcal{EE}$ or $\mathcal{ET}$, assume for now that we have corresponding classifiers producing confidence scores for every temporal relation in $\mathcal{R}_T$. Let them be $s^{ee}(\cdot)$ and $s^{et}(\cdot)$, respectively. Then the inference formulation for all the temporal relations within this document is:

$$\hat{Y} = \arg\max_{Y\in\mathcal{Y}} \sum_{i\in\mathcal{EE}} s^{ee}\{i \mapsto Y_i\} + \sum_{j\in\mathcal{ET}} s^{et}\{j \mapsto Y_j\} \quad (1)$$

where $Y_k \in \mathcal{R}_T$ is the temporal label of pair $k \in \mathcal{MM}$ (Let $\mathcal{M} = \mathcal{E} \cup \mathcal{T}$ be the set of all temporal nodes), “$k \mapsto Y_k$” represents the case where the label of pair k is predicted to be Y_k, Y is a vectorization of all these Y_k’s in one document, and $\mathcal{Y}$ is the constrained space that Y lies in.

We do not include the scores for $\mathcal{TT}$ because the temporal relationship between timexes can be trivially determined using the normalized dates of these timexes, as was done in (Do et al., 2012; Chambers et al., 2014; Mirza and Tonelli, 2016). We impose these relations via equality constraints denoted as $\mathcal{Y}_0$. In addition, we add symmetry and transitivity constraints dictated by the nature of time (denoted by $\mathcal{Y}_1$), and other prior knowledge derived from linguistic rules (denoted by $\mathcal{Y}_2$), which will be explained subsequently. Finally, we set $\mathcal{Y} = \cap_{i=0}^{2}\mathcal{Y}_i$ in Eq. (1).

Transitivity Constraints. Let the dimension of Y be n. Then a standard way to construct the symmetry and transitivity constraints is shown in (Bramsen et al., 2006; Chambers and Jurafsky, 2008; Denis and Muller, 2011; Do et al., 2012; Ning et al., 2017)

$$\mathcal{Y}_1 = \big\{Y \in \mathcal{R}_T^n | \forall m_{1,2,3} \in \mathcal{M}, Y_{(m_1,m_2)} = \bar{Y}_{(m_2,m_1)}, \\ Y_{(m_1,m_3)} \in \text{Trans}(Y_{(m_1,m_2)}, Y_{(m_2,m_3)})\big\}$$

where the bar sign is used to represent the reverse relation hereafter, and Trans(r_1, r_2) is a set comprised of all the temporal relations from $\mathcal{R}_T$ that do not conflict with r_1 and r_2.

The construction of Trans(r_1, r_2) necessitates a clearer definition of $\mathcal{R}_T$, the importance of which is often overlooked by existing methods. Existing approaches all followed the interval representation of events (Allen, 1984), which yields 13 temporal relations (denoted by $\tilde{\mathcal{R}}_T$ here) as shown in Fig. 1. Most systems used a reduced set, for ex-

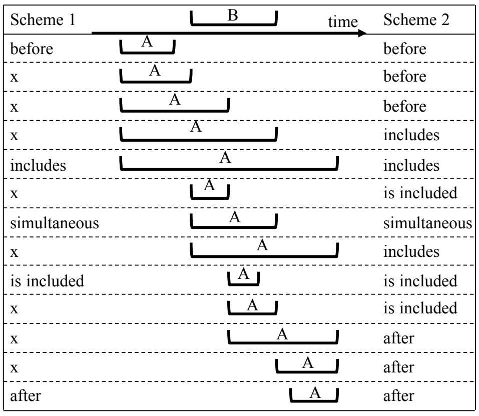

Figure 1: **Two possible interpretations to the label set of $\mathcal{R}_T = \{b, a, i, ii, s, v\}$** for the temporal relations between (A, B). "x" means that the label is ignored. Brackets represent time intervals along the time axis. Scheme 2 is adopted consistently in this work.

ample, {*before*, *after*, *includes*, *is included*, *simultaneously*, *vague*}. For notation convenience, we denote them $\mathcal{R}_T = \{b, a, i, ii, s, v\}$. Using a reduced set is more convenient in data annotation and leads to better performance in practice.

However, there has been limited discussion in the literature on how to interpret the reduced relation types. For example, is the "*before*" in $\mathcal{R}_T$ exactly the same as the "*before*" in the original set ($\tilde{\mathcal{R}}_T$) (as shown on the left-hand-side of Fig. 1), or is it a combination of multiple relations in $\tilde{\mathcal{R}}_T$ (the right-hand-side of Fig. 1)? We compare two reduction schemes in Fig. 1, where scheme 1 ignores low frequency labels directly and scheme 2 absorbs low frequency ones into their temporally closest labels. The two schemes barely have differences when a system only looks at a single pair of mentions at a time (this might explain the lack of discussion over this issue in the literature), but they lead to different Trans(r_1, r_2) sets and this difference can be magnified when the problem is solved jointly and when the label distribution changes across domains. To completely cover the 13 relations, we adopt scheme 2 in this work.

The resulting transitivity relations are shown in Table 1. The top part of Table 1 is a compact representation of three generic rules; for instance, Line 1 means that the labels themselves are transitive. Note that during human annotation, if an annotator looks at a pair of events and decides that multiple well-defined relations can exist, he/she labels it *vague*; also, when aggregating the labels from multiple annotators, a label will be changed to *vague* if the annotators disagree with each other. In either case, *vague* is chosen to be the label when a single well-defined relation cannot be uniquely determined by the contextual information. This explains why a *vague* relation (v) is always added in Table 1 if more than one label in Trans(r_1, r_2) is possible. As for Lines 6, 9-11 in Table 1 (where *vague* appears in Column r_2), Column Trans(r_1,r_2) was designed in such a way that r_2 cannot be uniquely determined through r_1 and Trans(r_1,r_2). For instance, r_1 is *after* on Line 9, if we further put *before* into Trans(r_1,r_2), then r_2 would be uniquely determined to be *before*, conflicting with r_2 being *vague*, so *before* should not be in Trans(r_1,r_2).

No.	r_1	r_2	Trans(r_1, r_2)
1	r	r	r
2	r	**s**	r
3	r_1	r_2	$\overline{\text{Trans}(\bar{r}_2, \bar{r}_1)}$
4	**b**	**i**	**b**, **i**, **v**
5	**b**	**ii**	**b**, **ii**, **v**
6	**b**	**v**	**b**, **i**, **ii**, **v**
7	**a**	**i**	**a**, **i**, **v**
8	**a**	**ii**	**a**, **ii**, **v**
9	**a**	**v**	**a**, **i**, **ii** ,**v**
10	**i**	**v**	**b**, **a**, **i**, **v**
11	**ii**	**v**	**b**, **a**, **ii**, **v**

Table 1: **Transitivity relations** based on the label set reduction scheme 2 in Fig. 1. If $(m_1, m_2) \mapsto r_1$ and $(m_2, m_3) \mapsto r_2$, then the relation of (m_1, m_3) must be chosen from Trans(r_1, r_2), $\forall m_1, m_2, m_3 \in \mathcal{M}$. The top part of the table uses r to represent generic rules compactly. Notations: before (**b**), after (**a**), includes (**i**), is included (**ii**), simultaneously (**s**), vague (**v**), and $\bar{r}$ represents the reverse relation of r.

Enforcing Linguistic Rules. Besides the transitivity constraints represented by $\mathcal{Y}_1$ above, we also propose to enforce prior knowledge to further constrain the search space for Y. Specifically, linguistic insight has resulted in rules for predicting the temporal relations with special syntactic or semantic patterns, as was done in CAEVO (a state-of-the-art method). Since these rule predictions often have high-precision, it is worthwhile incorporating them in global reasoning methods as well.

In the CCM framework, these rules can be represented as hard constraints in the search space for Y. Specifically,

$$\mathcal{Y}_2 = \left\{ Y_j = rule(j), \forall j \in \mathcal{J}^{(rule)} \right\}, \qquad (2)$$

where $\mathcal{J}^{(rule)} \subseteq \mathcal{M}\mathcal{M}$ is the set of pairs that can be determined by linguistic rules, and $rule(j) \in$

$\mathcal{R}_T$ is the corresponding decision for pair j according to these rules. In this work, we used the same set of rules designed by CAEVO for fair comparison.

3.2 Full Model with Causal Relations

Now we have presented the joint inference framework for temporal relations in Eq. (1). It is easier to explain our complete TCR framework on top of it. Let W be the vectorization of all causal relations and add the scores from the scoring function for causality $s^c(\cdot)$ to Eq. (1). Specifically, the full inference formulation is now:

$$\begin{aligned}\hat{Y},\hat{W} = \arg\max_{Y\in\mathcal{Y}, W\in\mathcal{W}_Y} \sum_{i\in\mathcal{EE}} s^{ee}\{i\mapsto Y_i\} \quad (3)\\ + \sum_{j\in\mathcal{ET}} s^{et}\{j\mapsto Y_j\} + \sum_{k\in\mathcal{EE}} s^{c}\{k\mapsto W_k\}\end{aligned}$$

where $\mathcal{W}_Y$ is the search space for W. $\mathcal{W}_Y$ depends on the temporal labels Y in the sense that

$$\begin{aligned}\mathcal{W}_Y \;=\; & \{W\in\mathcal{R}_C^m | \forall i,j\in\mathcal{E}, \text{if } W_{(i,j)}=c, \quad (4)\\ & \text{then } W_{(j,i)}=\bar{c}, \text{ and } Y_{(i,j)}=b\}\end{aligned}$$

where m is the dimension of W (i.e., the total number of causal pairs), $\mathcal{R}_C = \{c, \bar{c}, null\}$ is the label set for causal relations (i.e., "causes", "caused by", and "no relation"), and $W_{(i,j)}$ is the causal label for pair (i,j). The constraint represented by $\mathcal{W}_Y$ means that if a pair of events i and j are labeled to be "causes", then the causal relation between j and i must be "caused by", and the temporal relation between i and j must be "before".

3.3 Scoring Functions

In the above, we have built the joint framework on top of scoring functions $s^{ee}(\cdot)$, $s^{et}(\cdot)$ and $s^c(\cdot)$. To get $s^{ee}(\cdot)$ and $s^{et}(\cdot)$, we trained classifiers using the averaged perceptron algorithm (Freund and Schapire, 1998) and the same set of features used in (Do et al., 2012; Ning et al., 2017), and then used the soft-max scores in those scoring functions. For example, that means

$$s^{ee}\{i\mapsto r\} = \frac{w_r^T\phi(i)}{\sum_{r'\in\mathcal{R}_T} w_{r'}^T\phi(i)},\ i\in\mathcal{EE},\ r\in\mathcal{R}_T,$$

where $\{w_r\}$ is the learned weight vector for relation $r\in\mathcal{R}_T$ and $\phi(i)$ is the feature vector for pair $i\in\mathcal{EE}$.

Given a pair of ordered events, we need $s^c(\cdot)$ to estimate the scores of them being "causes" or "caused by". Since this scoring function has the same nature as $s^{ee}(\cdot)$, we can reuse the features from $s^{ee}(\cdot)$ and learn an averaged perceptron for $s^c(\cdot)$. In addition to these existing features, we also use prior statistics retrieved using our temporal system from a large corpus[3], so as to know *probabilistically* which event happens before another event. For example, in Example 1, we have a pair of events, ***e1:died*** and ***e2:exploded***. The prior knowledge we retrieved from that large corpus is that ***die*** happens before ***explode*** with probability 15% and happens after ***explode*** with probability 85%. We think this prior distribution is correlated with causal directionality, so it was also added as features when training $s^c(\cdot)$.

Note that the scoring functions here are implementation choice. The TCR joint framework is fully extensible to other scoring functions.

3.4 Convert the Joint Inference into an ILP

Conveniently, the joint inference formulation in Eq. (3) can be rewritten into an ILP and solved using off-the-shelf optimization packages, e.g., (Gurobi Optimization, Inc., 2012). First, we define indicator variables $y_i^r = \mathbb{I}\{Y_i = r\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function, $\forall i\in\mathcal{MM}$, $\forall r\in\mathcal{R}_T$. Then let $p_i^r = s^{ee}\{i\mapsto r\}$ if $i\in\mathcal{EE}$, or $p_i^r = s^{et}\{i\mapsto r\}$ if $i\in\mathcal{ET}$; similarly, let $w_j^r = \mathbb{I}\{W_i = r\}$ be the indicator variables for W_j and q_j^r be the score for $W_j = r\in\mathcal{R}_C$. Therefore, without constraints $\mathcal{Y}$ and $\mathcal{W}_Y$ for now, Eq. (3) can be written as:

$$\begin{aligned}\hat{y},\hat{w} = \arg\max \sum_{i\in\mathcal{EE}\cup\mathcal{ET}}\sum_{r\in\mathcal{R}_T} p_i^r y_i^r + \sum_{j\in\mathcal{EE}}\sum_{r\in\mathcal{R}_C} q_j^r w_j^r\\ \text{s.t.}\quad y_i^r, w_j^r\in\{0,1\},\ \sum_{r\in\mathcal{R}_T} y_i^r = \sum_{r\in\mathcal{R}_C} w_j^r = 1\end{aligned}$$

The prior knowledge represented as $\mathcal{Y}$ and $\mathcal{W}_Y$ can be conveniently converted into constraints for this optimization problem. Specifically, $\mathcal{Y}_1$ has two components, symmetry and transitivity:

$$\begin{aligned}\mathcal{Y}_1:\quad & \forall i,j,k\in\mathcal{M},\ y_{i,j}^r = y_{j,i}^{\bar{r}},\ \text{(symmetry)}\\ & y_{i,j}^{r_1} + y_{j,k}^{r_2} - \sum_{r_3\in\text{Trans}(r_1,r_2)} y_{i,k}^{r_3} \le 1\ \text{(transitivity)}\end{aligned}$$

where $\bar{r}$ is the reverse relation of r (i.e., $\bar{b} = a$, $\bar{i} = ii$, $\bar{s} = s$, and $\bar{v} = v$), and Trans(r_1, r_2) is defined in Table 1. As for the transitivity constraints,

[3] https://catalog.ldc.upenn.edu/LDC2008T19, which is disjoint to the test set used here. Please refer to (Ning et al., 2018a) for more analysis on using this corpus to acquire prior knowledge that aids temporal relation classification.

if both $y^{r_1}_{i,j}$ and $y^{r_2}_{j,k}$ are 1, then the constraint requires at least one of $y^{r_3}_{i,k}, r_3 \in \text{Trans}(r_1, r_2)$ to be 1, which means the relation between i and k has to be chosen from $\text{Trans}(r_1, r_2)$, which is exactly what $\mathcal{Y}_1$ is intended to do.

The rules in $\mathcal{Y}_2$ is written as

$$\mathcal{Y}_2: y^r_j = \mathbb{I}_{\{rule(j)=r\}}, \forall j \in \mathcal{J}^{(rule)} \text{ (linguistic rules)}$$

where $rule(j)$ and $\mathcal{J}^{(rule)}$ have been defined in Eq. (2). Converting the $\mathcal{TT}$ constraints, i.e., $\mathcal{Y}_0$, into constraints is as straightforward as $\mathcal{Y}_2$, so we omit it due to limited space.

Last, converting the constraints $\mathcal{W}_Y$ defined in Eq. (4) can be done as following:

$$\mathcal{W}_Y: w^c_{i,j} = w^{\bar{c}}_{j,i} \leq y^b_{i,j}, \; \forall i, j \in \mathcal{E}.$$

The equality part, $w^c_{i,j} = w^{\bar{c}}_{j,i}$ represents the symmetry constraint of causal relations; the inequality part, $w^c_{i,j} \leq y^b_{i,j}$ represents that if event i causes event j, then i must be before j.

4 Experiments

In this section, we first show on TimeBank-Dense (TB-Dense) (Cassidy et al., 2014), that the proposed framework improves temporal relation identification. We then explain how our new dataset with both temporal and causal relations was collected, based on which the proposed method improves for both relations.

4.1 Temporal Performance on TB-Dense

Multiple datasets with temporal annotations are available thanks to the TempEval (TE) workshops (Verhagen et al., 2007, 2010; UzZaman et al., 2013). The dataset we used here to demonstrate our improved temporal component was TB-Dense, which was annotated on top of 36 documents out of the classic TimeBank dataset (Pustejovsky et al., 2003). The main purpose of TB-Dense was to alleviate the known issue of sparse annotations in the evaluation dataset provided with TE3 (UzZaman et al., 2013), as pointed out in many previous work (Chambers, 2013; Cassidy et al., 2014; Chambers et al., 2014; Ning et al., 2017). Annotators of TB-Dense were forced to look at each pair of events or timexes within the same sentence or contiguous sentences, so that much fewer links were missed. Since causal link annotation is not available on TB-Dense, we only show our improvement in terms of temporal performance on TB-Dense. The standard train/dev/test split of TB-Dense was used and parameters were tuned to optimize the F_1 performance on dev. Gold events and time expressions were also used as in existing systems.

#	System	P	R	F_1
	Ablation Study			
1	Baseline	39.1	56.8	46.3
2	+Transitivity†	42.9	54.9	48.2
3	+$\mathcal{ET}$	44.3	54.8	49.0
4	+Rules	45.4	58.7	51.2
5	+Causal	**45.8**	**60.5**	**52.1**
	Existing Systems‡			
6	ClearTK	53.0	26.4	35.2
7	CAEVO	**56.0**	41.6	47.8
8	Ning et al. (2017)	47.1	**53.3**	**50.0**

†This is technically the same with Do et al. (2012), or Ning et al. (2017) without its structured learning component.
‡We added gold $\mathcal{TT}$ to both gold and system prediction. Without this, Systems 6-8 had F_1=28.7, 45.7, and 48.5, respectively, same with the reported values in Ning et al. (2017).

Table 2: **Ablation study of the proposed system in terms of the standard temporal awareness metric**. The baseline system is to make inference locally for each event pair without looking at the decisions from others. The "+" signs on lines 2-5 refer to adding a new source of information on top of its preceding system, with which the inference has to be global and done via ILP. All systems are significantly different to its preceding one with p<0.05 (McNemar's test).

The contributions of each proposed information sources are analyzed in the ablation study shown in Table 2, where we can see the F_1 score was improved step-by-step as new sources of information were added. Recall that $\mathcal{Y}_1$ represents transitivity constraints, $\mathcal{ET}$ represents taking event-timex pairs into consideration, and $\mathcal{Y}_2$ represents rules from CAEVO (Chambers et al., 2014). System 1 is the baseline we are comparing to, which is a local method predicting temporal relations one at a time. System 2 only applied $\mathcal{Y}_1$ via ILP on top of all $\mathcal{EE}$ pairs by removing the 2nd term in Eq. (1); for fair comparison with System 1, we added the same $\mathcal{ET}$ predictions from System 1. System 3 incorporated $\mathcal{ET}$ into the ILP and mainly contributed to an increase in precision (from 42.9 to 44.3); we think that there could be more gain if more time expressions existed in the testset. With the help of additional high-precision rules ($\mathcal{Y}_2$), the temporal performance can further be improved, as shown by System 4. Finally, using the causal extraction obtained via (Do et al., 2011) in the joint framework, the proposed method achieved the best precision, recall, and F_1 scores in our ablation study (Systems 1-5). According to the McNemar's test (Everitt, 1992; Dietterich, 1998), all Systems 2-5 were significantly different to its preceding system with $p<0.05$.

The second part of Table 2 compares several state-of-the-art systems on the same test set. ClearTK (Bethard, 2013) was the top performing system in TE3 in temporal relation extraction. Since it was designed for TE3 (not TB-Dense), it expectedly achieved a moderate recall on the test set of TB-Dense. CAEVO (Chambers et al., 2014) and Ning et al. (2017) were more recent methods and achieved better scores on TB-Dense. Compared with these state-of-the-art methods, the proposed joint system (System 5) achieved the best F_1 score with a major improvement in recall. We think the low precision compared to System 8 is due to the lack of structured learning, and the low precision compared to System 7 is propagated from the baseline (System 1), which was tuned to maximize its F_1 score. However, the effectiveness of the proposed information sources is already justified in Systems 1-5.

4.2 Joint Performance on Our New Dataset

4.2.1 Data Preparation

TB-Dense only has temporal relation annotations, so in the evaluations above, we only evaluated our temporal performance. One existing dataset with both temporal and causal annotations available is the Causal-TimeBank dataset (Causal-TB) (Mirza and Tonelli, 2014). However, Causal-TB is sparse in temporal annotations and is even sparser in causal annotations: In Table 3, we can see that with four times more documents, Causal-TB still has fewer temporal relations (denoted as T-Links therein), compared to TB-Dense; as for causal relations (C-Links), it has less than two causal relations in each document on average. Note that the T-Link sparsity of Causal-TB originates from TimeBank, which is known to have missing links (Cassidy et al., 2014; Ning et al., 2017). The C-Link sparsity was a design choice of Causal-TB in which C-Links were annotated based on only explicit causal markers (e.g., "A happened *because* of B").

Another dataset with parallel annotations is CaTeRs (Mostafazadeh et al., 2016b), which was primarily designed for the Story Cloze Test (Mostafazadeh et al., 2016a) based on common sense stories. It is different to the newswire domain that we are working on. Therefore, we decided to augment the EventCausality dataset provided in Do et al. (2011) with a modified version of the dense temporal annotation scheme proposed in Cassidy et al. (2014) and use this new dataset to showcase the proposed joint approach.

	Doc	Event	T-Link	C-Link
TB-Dense	36	1.6k	5.7k	-
EventCausality	25	0.8k	-	580
Causal-TB	183	6.8k	5.1k	318
New Dataset	25	1.3k	3.4k	172

Table 3: **Statistics of our new dataset with both temporal and causal relations annotated**, compared with existing datasets. T-Link: Temporal relation. C-Link: Causal relation. The new dataset is much denser than Causal-TB in both T-Links and C-Links.

The EventCausality dataset provides relatively dense causal annotations on 25 newswire articles collected from CNN in 2010. As shown in Table 3, it has more than 20 C-Links annotated per document on average (10 times denser than Causal-TB). However, one issue is that its notion for events is slightly different to that in the temporal relation extraction regime. To construct parallel annotations of both temporal and causal relations, we preprocessed all the articles in EventCausality using ClearTK to extract events and then manually removed some obvious errors in them. To annotate temporal relations among these events, we adopted the annotation scheme from TB-Dense given its success in mitigating the issue of missing annotations with the following modifications. First, we used a crowdsourcing platform, CrowdFlower, to collect temporal relation annotations. For each decision of temporal relation, we asked 5 workers to annotate and chose the majority label as our final annotation. Second, we discovered that comparisons involving ending points of events tend to be ambiguous and suffer from low inter-annotator agreement (IAA), so we asked the annotators to label relations based on the starting points of each event. This simplification does not change the nature of temporal relation extraction but leads to better annotation quality. For more details about this data collection scheme, please refer to (Ning et al., 2018b) for more details.

4.2.2 Results

Result on our new dataset jointly annotated with both temporal and causal relations is shown in Ta-

	Temporal			Causal
	P	R	F_1	Accuracy
1. Temporal Only	67.2	72.3	69.7	-
2. Causal Only	-	-	-	70.5
3. Joint System	**68.6**	**73.8**	**71.1**	**77.3**
Enforcing Gold Relations in Joint System				
4. Gold Temporal	100	100	100	*91.9*
5. Gold Causal	*69.3*	*74.4*	*71.8*	100

Table 4: **Comparison between the proposed method and existing ones, in terms of both temporal and causal performances**. See Sec. 4.2.1 for description of our new dataset. Per the McNemar's test, the joint system is significantly better than both baselines with p<0.05. Lines 4-5 provide the best possible performance the joint system could achieve if gold temporal/causal relations were given.

ble 4. We split the new dataset into 20 documents for training and 5 documents for testing. In the training phase, the training parameters were tuned via 5-fold cross validation on the training set.

Table 4 demonstrates the improvement of the joint framework over individual components. The "temporal only" baseline is the improved temporal extraction system for which the joint inference with causal links has NOT been applied. The "causal only" baseline is to use $s^c(\cdot)$ alone for the prediction of each pair. That is, for a pair i, if $s^c\{i \mapsto \text{causes}\} > s^c\{i \mapsto \text{caused by}\}$, we then assign "causes" to pair i; otherwise, we assign "caused by" to pair i. Note that the "causal accuracy" column in Table 4 was evaluated only on gold causal pairs.

In the proposed joint system, the temporal and causal scores were added up for all event pairs. The temporal performance got strictly better in precision, recall, and F_1, and the causal performance also got improved by a large margin from 70.5% to 77.3%, indicating that temporal signals and causal signals are helpful to each other. According to the McNemar's test, both improvements are significant with p<0.05.

The second part of Table 4 shows that if gold relations were used, how well each component would possibly perform. Technically, these gold temporal/causal relations were enforced via adding extra constraints to ILP in Eq. (3) (imagine these gold relations as a special rule, and convert them into constraints like what we did in Eq. (2)). When using gold temporal relations, causal accuracy went up to 91.9%. That is, 91.9% of the C-Links satisfied the assumption that the cause is temporally before the effect. First, this number is much higher than the 77.3% on line 3, so there is still room for improvement. Second, it means in this dataset, there were 8.1% of the C-Links in which the cause is temporally *after* its effect. We will discuss this seemingly counter-intuitive phenomenon in the Discussion section. When gold causal relations were used (line 5), the temporal performance was slightly better than line 3 in terms of both precision and recall. The small difference means that the temporal performance on line 3 was already very close to its best. Compared with the first line, we can see that gold causal relations led to approximately 2% improvement in precision and recall in temporal performance, which is a reasonable margin given the fact that C-Links are often much sparser than T-Links in practice.

Note that the temporal performance in Table 4 is consistently better than those in Table 2 because of the higher IAA in the new dataset. However, the improvement brought by joint reasoning with causal relations is the same, which further confirms the capability of the proposed approach.

5 Discussion

We have consistently observed that on the TB-Dense dataset, if automatically tuned to optimize its F_1 score, a system is very likely to have low precisions and high recall (e.g., Table 2). We notice that our system often predicts non-vague relations when the TB-Dense gold is vague, resulting in lower precision. However, on our new dataset, the same algorithm can achieve a more balanced precision and recall. This is an interesting phenomenon, possibly due to the annotation scheme difference which needs further investigation.

The temporal improvements in both Table 2 and Table 4 are relatively small (although statistically significant). This is actually not surprising because C-Links are much fewer than T-Links in newswires which focus more on the temporal development of stories. As a result, many T-Links are not accompanied with C-Links and the improvements are diluted. But for those event pairs having both T-Links and C-Links, the proposed joint framework is an important scheme to synthesize both signals and improve both. The comparison between Line 5 and Line 3 in Table 4 is a showcase of the effectiveness. We think that a deeper reason for the improvement achieved via a joint framework is that causality often encodes

Ex 4: Cause happened after effect.
The shares fell to a record low of ¥60 and ***(e8:finished)*** at ¥67 before the market ***(e9:closed)*** for the New Year holidays.
As she ***(e10:prepares)*** to ***(e11:host)*** her first show, Crowley writes on what viewers should expect.

humans prior knowledge as global information (e.g., "death" is *caused by* "explosion" rather than *causes* "explosion", regardless of the local context), while temporality often focuses more on the local context. From this standpoint, temporal information and causal information are complementary and helpful to each other.

When doing error analysis for the fourth line of Table 4, we noticed some examples that break the commonly accepted temporal precedence assumption. It turns out that they are not annotation mistakes: In Example 4, ***e8:finished*** is obviously *before* ***e9:closed***, but ***e9*** is a cause of ***e8*** since if the market did not close, the shares would not finish. In the other sentence of Example 4, she prepares *before* hosting her show, but ***e11:host*** is the cause of ***e10:prepares*** since if not for hosting, no preparation would be needed. In both cases, the cause is temporally after the effect because people are inclined to make projections for the future and change their behaviors before the future comes. The proposed system is currently unable to handle these examples and we believe that a better definition of what can be considered as events is needed, as part of further investigating how causality is expressed in natural language.

Finally, the constraints connecting causal relations to temporal relations are designed in this paper as "if A is the cause of B, then A must be *before* B". People have suggested other possibilities that involve the *includes* and *simultaneously* relations. While these other relations are simply different interpretations of temporal precedence (and can be easily incorporated in our framework), we find that they rarely happen in our dataset.

6 Conclusion

We presented a novel joint framework, **T**emporal and **C**ausal **R**easoning (TCR), using CCMs and ILP to the extraction problem of temporal and causal relations between events. To show the benefit of TCR, we have developed a new dataset that jointly annotates temporal and causal annotations, and then exhibited that TCR can improve both temporal and causal components. We hope that this notable improvement can foster more interest in jointly studying multiple aspects of events (e.g., event sequencing, coreference, parent-child relations) towards the goal of understanding events in natural language.

Acknowledgements

We thank all the reviewers for providing insightful comments and critiques. This research is supported in part by a grant from the Allen Institute for Artificial Intelligence (allenai.org); the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network; by DARPA under agreement number FA8750-13-2-0008; and by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053 (the ARL Network Science CTA).

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, of the Army Research Laboratory or the U.S. Government. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the view of the ARL.

References

James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence* 23(2):123–154.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *SemEval*. volume 2, pages 10–14.

Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, pages 177–180.

Steven Bethard, James H Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *IEEE International Conference on Semantic Computing (ICSC)*. pages 11–18.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. pages 189–198.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 501–506.

Nathanael Chambers. 2013. NavyTime: Event and time ordering from raw text. In *SemEval*. volume 2, pages 73–77.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics* 2:273–284.

Nathanael Chambers and Dan Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. pages 173–176.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine Learning* 88(3):399–431.

Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. volume 22, page 1788.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7):1895–1923.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics* 5:117–133.

Brian S Everitt. 1992. *The analysis of contingency tables*. CRC Press.

Yoav Freund and Robert E. Schapire. 1998. Large margin classification using the Perceptron algorithm. In *Proceedings of the Annual ACM Workshop on Computational Learning Theory (COLT)*. pages 209–217.

Başak Güler, Aylin Yener, and Ananthram Swami. 2016. Learning causal information flow structures in multi-layer networks. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. pages 1340–1344.

Gurobi Optimization, Inc. 2012. Gurobi optimizer reference manual. `http://www.gurobi.com`.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 753–760.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings the International Conference on Computational Linguistics (COLING)*. pages 2097–2106.

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *The 26th International Conference on Computational Linguistics*. pages 64–75.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*. pages 839–849.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation*. pages 51–61.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1038–1048.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 841–851.

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

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The TIMEBANK corpus. In *Corpus linguistics*. volume 2003, page 40.

Bryan Rink, Cosmin Adrian Bejan, and Sanda M Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *FLAIRS Conference*.

Dan Roth and Wen-Tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In Hwee Tou Ng and Ellen Riloff, editors, *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*. pages 1–8.

Yizhou Sun, Kunqing Xie, Ning Liu, Shuicheng Yan, Benyu Zhang, and Zheng Chen. 2007. Causal relation of queries from temporal logs. In *The International World Wide Web Conference*. pages 1141–1142.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, 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*. volume 2, pages 1–9.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval temporal relation identification. In *SemEval*. pages 75–80.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the TARSQI toolkit. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*. pages 189–192.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *SemEval*. pages 57–62.

Appendix

For the crowdsourcing task described in Sec. 4.2.1, we have a gold set annotated by experts, based on which CrowdFlower has two types of quality controls: First, each crowdsourcer has to pass a qualifying test, which is randomly picked from the gold set, with accuracy higher than 70% before the crowdsourcer is allowed on a job; one is also tested on the gold set randomly throughout the process without notice and once the accuracy drops below 70%, the crowdsourcer is kicked out automatically and his or her annotations are all considered tainted and not used. Finally kept are those annotations from crowdsourcers who survive in the end (so called Trusted Contributors). **In our case, trusted contributors had 87% accuracy on our gold set**, as reported by CrowdFlower.

Accuracy on the gold set reflects the crowdsourcers' level of understanding of the job owner's intent. Another quality metric is the level of agreement among themselves. Since more than two annotators are involved in crowdsourcing, Cohen's Kappa cannot be applied to crowdsourced data. Instead, CrowdFlower adopts a default IAA metric called WAWA (Worker Agreement with Aggregate). WAWA indicates the average number of crowdsourcers' responses agreed with the aggregate answer for each question. For example, if N individual responses were obtained in total, and n of them were correct when compared to the aggregate answer, then WAWA is simply n/N. **The WAWA score of the proposed new dataset was 78%.**

In terms of the cost of the crowdsourcing job, we paid \$0.01 for every individual response and the total cost was approximately \$500 (including overhead fees). The actual annotation time was about 15 hours for the entire dataset to be finished.