Joint Inference for Event Timeline Construction

Quang Xuan Do **Wei Lu** **Dan Roth**
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{quangdo2,luwei,danr}@illinois.edu

Abstract

This paper addresses the task of constructing a timeline of events mentioned in a given text. To accomplish that, we present a novel representation of the temporal structure of a news article based on time intervals. We then present an algorithmic approach that jointly optimizes the temporal structure by coupling local classifiers that predict associations and temporal relations between pairs of temporal entities with global constraints. Moreover, we present ways to leverage knowledge provided by event coreference to further improve the system performance. Overall, our experiments show that the joint inference model significantly outperformed the local classifiers by 9.2% of relative improvement in F_1. The experiments also suggest that good event coreference could make remarkable contribution to a robust event timeline construction system.

1 Introduction

Inferring temporal relations amongst a collection of events in a text is a significant step towards various important tasks such as automatic information extraction and document comprehension. Over the past few years, with the development of the TimeBank corpus (Pustejovsky et al., 2003) , there have been several works on building automatic systems for such a task (Mani et al., 2006; Chambers and Jurafsky, 2008; Yoshikawa et al., 2009; Denis and Muller, 2011).

Most previous works devoted much efforts to the task of identifying relative temporal relations (such as *before*, or *overlap*) amongst events (Chambers and Jurafsky, 2008; Denis and Muller, 2011), without addressing the task of identifying correct associations between events and their absolute time of occurrence. Even if this issue is addressed, certain restrictions are often imposed for efficiency reasons (Yoshikawa et al., 2009; Verhagen et al., 2010). In practice, however, being able to automatically infer the correct time of occurrence associated with each event is crucial. Such information not only leads to better text comprehension, but also enables fusion of event structures extracted from multiple articles or domains.

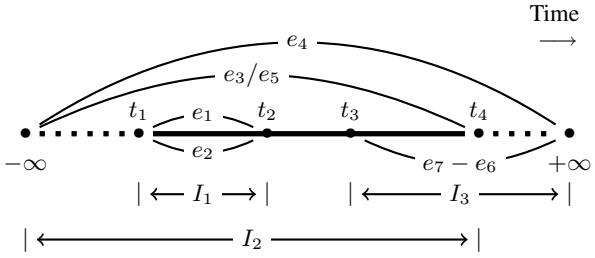

Figure 1: A graphical illustration of our timeline representation. The e's, t's and I's are events, time points and time intervals, respectively.

In this work, we are specifically interested in mapping events into an universal *timeline* representation. Besides inferring the relative temporal relations amongst the events, we would also like to automatically infer a specific absolute time of occurrence for each event mentioned in the text. Unlike previous work, we associate each event with a specific absolute time interval inferred from the text. An example timeline representation is illustrated in Fig.

1. Further details of our timeline representation are given in Sec. 2.3.

We perform global inference by combining a collection of local pairwise classifiers through the use of an Integer Linear Programming (ILP) formulation that promotes global coherence among local decisions. The formulation allows our model to predict both event-event relations and event-time interval associations simultaneously. We show that, with the use of time intervals instead of time points, our approach leads to a more concise ILP formulation with reduced number of variables and constraints.

Moreover, we observed that event coreference can reveal important information for such a task. We propose that different event mentions that refer to the same event can be grouped together before classification and performing global inference. This can reduce the amount of efforts in both classification and inference stages and can potentially eliminate mistakes that would be made otherwise without such coreference information. To the best of our knowledge, our proposal of leveraging event coreference to support event timeline construction is novel.

Our experiments on a collection of annotated news articles from the standard ACE dataset demonstrate that our approach produces robust timelines of events. We show that our algorithmic approach is able to combine various local evidences to produce a global coherent temporal structure, with improved overall performance. Furthermore, the experiments show that the overall performance can be further improved by exploiting knowledge from event coreference.

2 Background

We focus on the task of mapping event mentions in a news article to a timeline. We first briefly describe and define several basic concepts.

2.1 Events

Following the annotation guidelines of the ACE project, we define an event as an action or occurrence that happens with associated participants or arguments. We also distinguish between events and event mentions, where a unique event can be coreferred to by a set of explicit event mentions in an article. Formally, an event E^i is co-referred to by a set of event mentions $(e_1^i, e_2^i, \ldots, e_k^i)$. Each event mention e can be written as $p(a_1, a_2, \ldots, a_l)$, where the predicate p is the word that triggers the presence of e in text, and $a_1, a_2, \ldots a_l$ are the arguments associated with e. In this work we focus on four temporal relations between two event mentions including *before*, *after*, *overlap* and *no relation*.

2.2 Time Intervals

Similar to Denis and Muller (2011), we define time intervals as pairs of time endpoints. Each time interval I is denoted by $[t^-, t^+]$, where t^- and t^+ are two time endpoints representing the lower and upper bound of the interval I, respectively, with $t^- \leq t^+$. The general form of a time endpoint is written as "*YYYY-MM-DD hh:mm:ss*". An endpoint can be undefined, in which case it is set to an infinity value: $-\infty$, or $+\infty$. There are two types of time intervals:

Explicit intervals are time intervals that can be extracted directly from a given text. For example, consider the following snippet of an article in our data set: *The litigation covers buyers in auctions outside the United States between January 1, 1993 and February 7, 2000.* In this example, we can extract and normalize two time intervals which are explicitly written, including *January 1, 1993* $\rightarrow$ [1993-01-01 00:00:00, 1993-01-01 23:59:59] and *February 7, 2000* $\rightarrow$ [2000-02-07 00:00:00, 2000-02-07 23:59:59]. Moreover, an explicit interval can also be formed by one or more separate explicit temporal expressions. In the example above, the connective term *between* relates the two expressions to form a single time interval: *between January 1, 1993 and February 7, 2000* $\rightarrow$ [1993-01-01 00:00:00, 2000-02-07 23:59:59]. To extract explicit time intervals from text, we use the time interval extractor described in Zhao et al. (2012).

Implicit intervals are time intervals that are not explicitly mentioned in the text. We observed that there are events that cannot be assigned to any precise time interval but are roughly known to occur in the past or in the future relative to the Document Creation Time (DCT) of the article. We introduce two implicit time intervals to represent the past and the future events as $(-\infty, t^-_{DCT}]$ and $[t^+_{DCT}, +\infty)$, respectively. In addition, we also allow an event mention to be assigned into the entire timeline, which is denoted by $(-\infty, +\infty)$ if we can-

not identify its time of occurrence. We also consider DCT as an implicit interval.

We say that the time interval I_i precedes the time interval I_j on a timeline if and only if $t_i^+ \leq t_j^-$, which also implies that I_i succeeds I_j if and only if $t_i^- \geq t_j^+$. The two intervals overlap, otherwise.

2.3 Timeline

We define a timeline as a partially ordered set of time intervals. Fig. 1 gives a graphical illustration of an example timeline, where events are annotated and associated with time intervals. Relations amongst events can be properly reflected in the timeline representation. For example, in the figure, the events e_1 and e_2 are both associated with the interval I_1. The relation between them is *no relation*, since it is unclear which occurs first. On the other hand, e_5 and e_3 both happen in the interval I_2 but they form an *overlap* relation. The events e_6 and e_7 occur within the same interval I_3, but e_7 precedes (i.e. *before*) e_6 on the timeline. The event e_4 is associated with the interval $(-\infty, +\infty)$, indicating there is no knowledge about its time of occurrence.

We believe that such a timeline representation for temporally ordering events has several advantages over the temporal graph representations used in previous works (Chambers and Jurafsky, 2008; Yoshikawa et al., 2009; Denis and Muller, 2011). Unlike previous works, in our model the events are partially ordered in a single timeline, where each event is associated with a precise time interval. This improves human interpretability of the temporal relations amongst events and time. This property of our timeline representation, thus, facilitates merging multiple timelines induced from different articles. Furthermore, as we will show later, the use of time intervals within the timeline representation simplifies the global inference formulation and thus the inference process.

3 A Joint Timeline Model

Our task is to induce a globally coherent timeline for a given article. We thus adopt a global inference model for performing the task. The model consists of two components: (1) two local pairwise classifiers, one between event mentions and time intervals (the E–T classifier) and one between event mentions themselves (the E–E classifier), and (2) a joint inference module that enforces global coherency constraints on the final outputs of the two local classifiers. Fig. 2 shows a simplified temporal structure of event mentions and time intervals of an article in our model.

Our E–T classifier is different from previous work (Chambers and Jurafsky, 2008; Yoshikawa et al., 2009; Denis and Muller, 2011), where such classifiers were trained to identify temporal relations between event mentions and a temporal *expression*. In our work, in order to construct absolute timeline of event mentions, temporal expressions are captured and normalized as absolute time intervals. The E–T classifiers are then used to assign event mentions to their contextually corresponding time intervals.

We also lifted several restrictions imposed in previous work (Bethard et al., 2007; Yoshikawa et al., 2009; Verhagen et al., 2010). Specifically, we do not require that event mentions and time expressions have to appear in the same sentence, and we do not require two event mentions have to appear very close to each other (e.g., main event mentions in adjacent sentences) in order to be considered as candidate pairs for classification. Instead, we performed classifications over all pairs of event mentions and time intervals as well as over all pairs of event mentions. We show through experiments that lifting these restrictions is indeed important (see Sec. 5).

Another important improvement over previous work is our global inference model We would like to highlight that our work is also distinct from most previous works in the global inference component. Specifically, our global inference model jointly optimizes the *E-E* relations amongst event mentions and their associations, *E-T*, with temporal information (intervals in our case). Previous work (Chambers and Jurafsky, 2008; Denis and Muller, 2011), on the other hand, assumed that the *E-T* information is given and only tried to improve *E-E*.

3.1 The Pairwise Classifiers

We first describe our local classifiers that associate event mention with time interval and classify temporal relations between event mentions, respectively.

C_{E-T}: is the E–T classifier that associates an event mention with a time interval. Given an event mention and a time interval, the classifier predicts

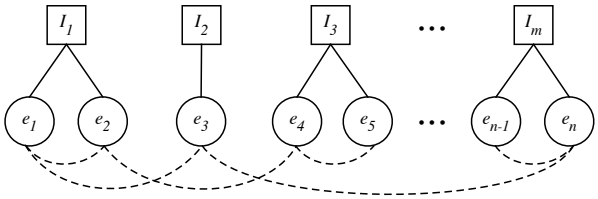

Figure 2: A simplified temporal structure of an article. There are m time intervals $I_1 \cdots I_m$ and n event mentions $e_1 \cdots e_n$. A solid edge indicates an association between an interval and an event mention, whereas a dash edge illustrates a temporal relation between two event mentions.

whether the former associates with the latter.

$$C_{E-T}(e_i, I_j) \rightarrow \{0, 1\},$$
$$\forall i, j, 1 \leq i \leq n, 1 \leq j \leq m, \quad (1)$$

where n and m are the number of event mentions and time intervals in an article, respectively.

C_{E-E}: is the E–E classifier that identifies the temporal relation between two event mentions. Given a pair of event mentions, the classifier predicts one of the four temporal relations between them: $\bar{b}$*efore*, $\bar{a}$*fter*, $\bar{o}$*verlap* and $\bar{n}$*o relation*. Specifically:

$$C_{E-E}(e_i, e_j) \rightarrow \{\bar{b}, \bar{a}, \bar{o}, \bar{n}\},$$
$$\forall i, j, 1 \leq i, j \leq n, i \neq j, \quad (2)$$

For training of the classifiers, we define a set of features following some previous work (Bethard et al., 2007; Chambers and Jurafsky, 2008; Yoshikawa et al., 2009), together with some additional features that we believe to be helpful for the interval-based representation. We describe the base features below and use † and ‡ to denote the features used for C_{E-T} and C_{E-E}, respectively. We use the term *temporal entity* (or *entity*, for short) to refer to either an event mention or a time interval.

Lexical Features: A set of lexical features related to the temporal entities: (i)†‡ the word, lemma and part-of-speech of the input event mentions and the context surrounding them, where the context is defined as a window of 2 words before and after the mention; (ii)† the modal verbs to the left and to the right of the event mention; (iii)‡ the temporal connectives between the event mentions[1].

[1]We define a list of temporal connectives including *before*, *after*, *since*, *when*, *meanwhile*, *lately*, etc.

Syntactic Features: (i)†‡ which entity appears first in the text; (ii)†‡ whether the two entities appear in the same sentence; (iii)†‡ the quantized number of sentences between the two entities[2]; (iv)†‡ whether the input event mentions are covered by prepositional phrases and what are the heads of the phrases; (v)†‡ if the entities are in the same sentence, what is their least common constituent on the syntactic parse tree; (vi)† whether there is any other temporal entity that is closer to one of the two entities.

Semantic Features‡: A set of semantic features, mostly related to the input event mentions: (i) whether the input event mentions have a common synonym from their synsets in WordNet (Fellbaum, 1998); (ii) whether the input event mentions have a common derivational form derived from WordNet.

Linguistic Features†‡: The tense and the aspect of the input event mentions. We use an in-house rule-based recognizer to extract these features.

Time Interval Features†: A set of features related to the input time interval: (i) whether the interval is implicit; (ii) if it is implicit, identify its interval type: "dct" $= [t^-_{DCT}, t^+_{DCT}]$, "past" $= (-\infty, t^-_{DCT}]$, "feature" $= [t^+_{DCT}, +\infty)$, and "entire" $= (-\infty, +\infty)$; (iii) the interval is before, after or overlapping with the DCT.

We note that unlike many previous work (Mani et al., 2006; Chambers and Jurafsky, 2008; Denis and Muller, 2011), our classifiers do not use any gold annotations of event attributes (event class, tense, aspect, modal and polarity) provided in the TimeBank corpus as features.

In our work, we use a regularized averaged Perceptron (Freund and Schapire, 1999) as our classification algorithm[3]. We used the one-vs.-all scheme to transform a set of binary classifiers into a multi-class classifier (for C_{E-E}). The raw prediction scores were converted into probability distribution using the Softmax function (Bishop 1996). If there are n classes and the raw score of class i is act_i, the posterior estimation for class i is:

$$\tilde{P}(i) = \frac{e^{act_i}}{\sum_{1 \leq j \leq n} e^{act_j}}$$

[2]We quantize the number of sentences between two entities to 0, 1, 2, less than 5 and greater than or equal to 5

[3]Other algorithm (e.g. SVM) gave comparable or worse results, so we only show the results from Averaged Perceptron.

3.2 Joint Inference for Event Timeline

To exploit the interaction among the temporal entities in an article, we optimize the predicted temporal structure, formed by predictions from C_{E-T} and C_{E-E}, w.r.t. a set of global constraints that enforce coherency on the final structure. We perform exact inference using Integer Linear Programming (ILP) as in (Roth and Yih, 2007; Clarke and Lapata, 2008). We use the Gurobi Optimizer[4] as a solver.

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ denote the set of time intervals extracted from an article, and let $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$ denote all event mentions in the same article. Let $\mathcal{EI} = \{(e_i, I_j) \in \mathcal{E} \times \mathcal{I} | e_i \in \mathcal{E}, I_j \in \mathcal{I}\}$ denote the set of all pairs of event mentions and time intervals. We also denote the set of event mention pairs by $\mathcal{EE} = \{(e_i, e_j) \in \mathcal{E} \times \mathcal{E} | e_i \in \mathcal{E}, e_j \in \mathcal{E}, i \neq j\}$. The prediction probability of an association of a pair $eI \in \mathcal{EI}$, given by classifier C_{E-T}, is denoted by $p_{\langle eI,1\rangle}$[5]. Now, let $\mathcal{R} = \{\bar{b}, \bar{a}, \bar{o}, \bar{n}\}$ be the set of temporal relations between two event mentions. The prediction probability of an event mention pair $ee \in \mathcal{EE}$ that takes temporal relation r, given by C_{E-E}, is denoted by $p_{\langle ee,r\rangle}$. Furthermore, we define $x_{\langle eI,1\rangle}$ to be a binary indicator variable that takes on the value 1 iff an association is predicted between e and I. Similarly, we define a binary indicator variable $y_{\langle ee,r\rangle}$ of a pair of event mentions ee that takes on the value 1 iff ee is predicted to hold the relation r.

The objective function is then defined as a linear combination of the prediction probabilities from the two local classifiers as follows:

$$\arg\max_{x,y} \Big[\lambda \sum_{eI \in \mathcal{EI}} p_{\langle eI,1\rangle} \cdot x_{\langle eI,1\rangle} + (1-\lambda) \sum_{ee \in \mathcal{EE}} \sum_{r \in \mathcal{R}} p_{\langle ee,r\rangle} \cdot y_{\langle ee,r\rangle}\Big] \quad (3)$$

subject to the following constraints:

$$x_{\langle eI,1\rangle} \in \{0,1\}, \quad \forall eI \in \mathcal{EI} \quad (4)$$

$$y_{\langle ee,r\rangle} \in \{0,1\}, \quad \forall ee \in \mathcal{EE}, r \in \mathcal{R} \quad (5)$$

$$\sum_{r \in \mathcal{R}} y_{\langle ee,r\rangle} = 1, \quad \forall ee \in \mathcal{EE} \quad (6)$$

We use the single parameter λ to balance the overall contribution of two components E-T and E-E. λ is determined through cross validation tuning on a development set. We use (4) and (5) to make sure $x_{\langle eI,1\rangle}$ and $y_{\langle ee,r\rangle}$ are binary values. The equality constraint (6) ensures that exactly one particular relation can be assigned to each event mention pair.

In addition, we also require that each event is associated with only one time interval. These constraints are encoded as follows:

$$\sum_{I \in \mathcal{I}} x_{\langle eI,1\rangle} = 1, \quad \forall e \in \mathcal{E} \quad (7)$$

Our model also enforces reflexivity and transitivity constraints on the relations among event mentions as follows:

$$y_{\langle e_i e_j, r\rangle} - y_{\langle e_j e_i, \hat{r}\rangle} = 0, \quad \forall e_i e_j = (e_i, e_j) \in \mathcal{EE}, i \neq j \quad (8)$$

$$y_{\langle e_i e_j, r_1\rangle} + y_{\langle e_j e_k, r_2\rangle} - y_{\langle e_i e_k, r_3\rangle} \leq 1, \quad \forall e_i e_j, e_j e_k, e_i e_k \in \mathcal{EE}, i \neq j \neq k \quad (9)$$

The equality constraints in (8) encode reflexive property of event-event relations, where the relation $\hat{r}$ denotes the inversion of the relation r. The set of possible $(r, \hat{r})$ pairs is defined as follows: $\{(\bar{b}, \bar{a}), (\bar{a}, \bar{b}), (\bar{o}, \bar{o}), (\bar{n}, \bar{n})\}$. Following the work of (Bramsen et al., 2006; Chambers and Jurafsky, 2008), we encode transitive closure of relations between event mentions with inequality constraints in (9), which states that if the pair (e_i, e_j) has a certain relation r_1, and the pair (e_j, e_k) has the relation r_2, then the relation r_3 must be satisfied between e_i and e_k. Examples of such triple (r_1, r_2, r_3) include ($\bar{b}$, $\bar{b}$, $\bar{b}$) and ($\bar{a}$, $\bar{a}$, $\bar{a}$).

Finally, to capture the interactions between our local pairwise classifiers we add the following constraints:

$$x_{\langle e_i I_k,1\rangle} + x_{\langle e_j I_l,1\rangle} - y_{\langle e_i e_j, \bar{b}\rangle} \leq 1, \quad \forall e_i I_k, e_j I_l \in \mathcal{EI}, \forall e_i e_j \in \mathcal{EE}, \quad I_k \text{ precedes } I_l, i \neq j, k \neq l \quad (10)$$

Intuitively, the inequality constraints in (10) specify that a temporal relation between two event mentions can be inferred from their respective associated

[4] http://gurobi.com/

[5] This value is complementary to the non-association probability, denoted by $p_{\langle eI,0\rangle} = 1 - p_{\langle eI,1\rangle}$

time intervals. Specifically, if two event mentions e_i and e_j are associated with two time intervals I_k and I_l respectively, and I_k precedes I_l in the timeline, then e_i must happen before e_j.

It is important to note that our interval-based formulation is more concise in terms of the number of variables and constraints needed in the ILP relative to time expression-based (or timepoint-based) formulations used in previous work (Chambers and Jurafsky, 2008). Specifically, in such timepoint-based formulations, the relation between each event mention and each time expression needs to be inferred, resulting in $|\mathcal{E}||\mathcal{T}||\mathcal{R}_\mathcal{T}|$ variables, where $|\mathcal{E}|$, $|\mathcal{T}|$, and $|\mathcal{R}_\mathcal{T}|$ are the numbers of event mentions, time points, and temporal relations respectively. In contrast, only $|\mathcal{E}||\mathcal{I}|$ variables are required in our formulation, where $|\mathcal{I}|$ is the number of intervals (since we extract intervals explicitly, $|\mathcal{I}|$ is roughly equal to $|\mathcal{T}|$). Furthermore, performing inference with the timepoint-based formulation would require $|\mathcal{E}||\mathcal{T}|$ equality constraints to enforce that each event mention can take only one relation in $\mathcal{R}_\mathcal{T}$ for a particular time point, whereas our interval-based model only requires $|\mathcal{E}|$ constraints, since each event is strictly associated with one interval (see Eqn. (7)). We justify the benefits of our formulation later in Sec. 5.4.

4 Incorporating Knowledge from Event Coreference

One of the key contributions of our work is using event coreference information to enhance the timeline construction performance. This is motivated by the following two principles:

> (P1) *All mentions of a unique event are associated with the same time interval, and overlap with each other.*
>
> (P2) *All mentions of an event have the same temporal relation with all mentions of another event.*

The example below, extracted from an article published on $03/11/2003$ in the Automatic Content Extraction (ACE), 2005, corpus[6] serves to illustrate the significance of event coreference to our task.

> *The world's most powerful fine art auction houses, Sotheby's and Christie's, have agreed to* [e_1^1 = **pay**] *40 million dollars to settle an international price-fixing scam, Sotheby's said. The* [e_2^1 = **payment**]*, if approved by the courts, would settle a slew of* [e_1^2 = **suits**] *by clients over auctions held between 1993 and 2000 outside the US. ... Sotheby's and Christie's will each* [e_3^1 = **pay**] *20 million dollars," said Sotheby's, which operates in 34 countries.*

In this example, there are 4 event mentions, whose trigger words are highlighted in bold face. The underlined text gives an explicit time interval: $I_1 =$ [1993-01-01 00:00:00, 2000-12-31 23:59:59] (we ignore 2 other intervals given by *1993* and *2000* to simplify the illustration). Now if we consider the event mention e_2^1, it actually belongs to the implicit future interval $I_2 =$ [2003-03-11 23:59:59, $+\infty$). Nevertheless, there is a reasonable chance that C_{E-T} associates it with I_1, given that they both appear in the same sentence, and there is no direct evident feature indicating the event will actually happen in the future. In such a situation, using a local classifier to identify the correct temporal association could be challenging.

Fortunately, precise knowledge from event coreference may help alleviate such a problem. The knowledge reveals that the 4 event mentions can be grouped into 2 distinct events: $E^1 = \{e_1^1, e_2^1, e_3^1\}$, $E^2 = \{e_1^2\}$. If C_{E-T} can make a strong prediction in associating the event mention e_1^1 (or e_3^1) to I_2, instead of I_1, the system will have a high chance to re-assign e_2^1 to I_2 based on principle (P1). Similarly, if C_{E-E} is effective in figuring out that some mention of event E^1 occurs *after* some mention of E^2, then all the mentions of E^1 would be predicted to occur *after* all mentions in E^2 according to (P2).

To incorporate knowledge from event coreference into our classifiers and the joint inference model, we use the following procedure: (1) performing classification with C_{E-T} and C_{E-E} on the data, (2) using the knowledge from event coreference to overwrite the prediction probabilities obtained by the two local classifiers in step (1), and (3) applying the joint inference model on the new prediction probabilities obtained from (2). We note that if we stop at step (2), we get the outputs of the local classifiers enhanced by event coreference knowledge.

To overwrite the classification probabilities using

[6] http://www.itl.nist.gov/iad/mig/tests/ace/2005/

event coreference knowledge, we propose two approaches as follows:

MaxScore: We define the probability between any mention $e \in E^i$ and an interval I as follows:

$$p_{\langle eI,1\rangle} = \max_{e' \in E^i} \tilde{P}(e', I) \quad (11)$$

where $\tilde{P}(e', I)$ is the classifier (C_{E-T}) probability for associating event mention e' to the time interval.

On the other hand, the probabilities for associating the set of temporal relations, $\mathcal{R}$, to each pair of mentions in $E^i \times E^j$, is given by the following pair:

$$\begin{aligned}(e^i, e^j)^* &= \underset{(e^{i'}, e^{j'}) \in E^i \times E^j, r \in \mathcal{R}}{\arg\max} \tilde{P}\big((e^{i'}, e^{j'}), r)\big) \\ p_{\langle ee,r\rangle} &= \tilde{P}\big((e^i, e^j)^*, r\big), \forall r \in \mathcal{R} \quad (12)\end{aligned}$$

In other words, over all possible event mention pairs and relations, we first pick the pair who globally obtains the highest probability for some relation. Next, we simply take the probability distribution of that event mention pair as the distribution over the relations, for the event pair.

SumScore: The probability between any mention $e \in E^i$ and an interval I is obtained by:

$$p_{\langle eI,1\rangle} = \frac{1}{|E^i|} \sum_{e' \in E^i} \tilde{P}(e', I) \quad (13)$$

To obtain the probability distribution over the set of temporal relations, $\mathcal{R}$, for any pair of mentions in $E^i \times E^j$, we used the following procedure:

$$\begin{aligned}r^* &= \underset{r \in \mathcal{R}}{\arg\max} \sum_{e^i \in E^i} \sum_{e^j \in E^j} \tilde{P}\big((e^i, e^j), r\big) \\ (e^i, e^j)^* &= \underset{(e^{i'}, e^{j'}) \in E^i \times E^j}{\arg\max} \tilde{P}\big((e^{i'}, e^{j'}), r^*\big) \\ p_{\langle ee,r\rangle} &= \tilde{P}\big((e^i, e^j)^*, r\big), \forall r \in \mathcal{R} \quad (14)\end{aligned}$$

In other words, given two groups of event mentions, we first compute the total score of each relation, and select the relation which has the highest score. Next from the list of pairs of event mentions from the two groups, we select the pair which has the relation r* with highest score compared to all other pairs. The probability distribution of this pair will be used as the probability distribution of all event mention pairs between the two events.

In both approaches, we assign the *overlap* relations to all pairs of event mentions in the same event with probability 1.0.

5 Experimental Study

We first describe the experimental data and then present and discuss the experimental results.

5.1 Data and Setup

Most previous works in temporal reasoning used the TimeBank corpus as a benchmark. The corpus contains a fairly diverse collection of annotated event mentions, without any specific focus on certain event types. According to the annotation guideline of the corpus, most of verbs, nominalizations, adjectives, predicative clauses and prepositional phrases can be tagged as events. However, in practice, when performing temporal reasoning about events in a given text, one is typically interested in significant and typed events, such as *Killing*, *Legislation*, *Election*. Furthermore, event mentions in TimeBank are annotated with neither event arguments nor event coreference information.

We noticed that the ACE 2005 corpus contains the annotation that we are interested in. The corpus consists of articles annotated with event mentions (with event triggers and arguments) and event coreference information. To create an experimental data set for our work, we selected from the corpus 20 newswire articles published in March 2003. To extract time intervals from the articles, we used the time interval extractor described in (Zhao et al., 2012) with minimal post-processing. Implicit intervals are also added according to Sec. 2.2. We then hired an annotator with expertise in the field to annotate the data with the following information: (i) event mention and time interval association, and (ii) the temporal relations between event mentions, including $\{\bar{b}, \bar{a}, \bar{o}\}$. The annotator was not required to annotate all pairs of event mentions, but as many as possible. Next, we saturated the relations based on the initial annotations as follows: (i) event mentions that had not been associated with any time intervals were assigned to the entire timeline interval $(-\infty, +\infty)$, and (ii) added inferred temporal relations between event mentions with reflectivity and transitivity. Table 1 shows the data statistics before and after saturation. There are totally 8312 event pairs from 20 documents, including *no relation* pairs. We note that in a separate experiment, we still evaluated C_{E-E} on the TimeBank corpus and got better performance

Data	#Intervals	#E-mentions	#E-T	#E-E
Initial	232	324	305	376
Saturated	232	324	324	5940

Table 1: The statistics of our experimental data set.

than a corresponding classifier in an existing work (see Sec. 5.4).

We conducted all experiments with 5-fold cross validation at the instance level on our data set after saturation. The global inference model was applied on a whole document. The results of the systems are reported in averaged precision, recall and F_1 score on the association performance, for C_{E-T}, and the temporal relations (we excluded the $\bar{n}$ relation, for C_{E-E}). We also measured the overall performance of the systems by computing the average of the performance of the classifiers.

5.2 A Baseline

We developed a baseline system that works as follows. It associates an event mention with the closest time interval found in the same sentence. If such an interval is not found, the baseline associates the mention with the closest time interval to the left. If the interval is again not found, the mention will be associated with the DCT interval. The baseline is based on the intuition of natural reading order: events that are mentioned earlier are likely to precede those mentioned later. For the temporal relation between a pair of event mentions, the baseline treats the event mention that appears earlier in the text as temporally happening before the other mention. The baseline performance is shown in the first group of results in Table 2.

5.3 Our Systems

For our systems, we first evaluated the performance of our local pairwise classifiers and the global inference model. The second group of results in Table 2 shows the systems' performance. Overall, the results show that our global inference model relatively outperformed the baseline and the local classifiers by 57.8% and 9.2% in F_1, respectively. We perform a bootstrap resampling significance test (Koehn, 2004) on the output predictions of the local classifiers with and without the inference model. The test shows that the overall improvement with the inference model is statistically significant ($p < 0.01$). This indicates the effectiveness of our joint inference model with global coherence constraints.

Next, we integrated event coreference knowledge into our systems (as described in Sec. 4) and evaluated their performance. Our experiments showed that the *SumScore* approach works better for C_{E-T}, while *MaxScore* is more suitable for C_{E-E}. Our observations showed that event mentions of an event may appear in close proximity with multiple time intervals in the text, making C_{E-T} produce high prediction scores for many event mention-interval pairs. This, consequently, confuses *MaxScore* on the best association of the event and the time intervals, whereas *SumScore* overcomes the problem by averaging out the association scores. On the other hand, C_{E-E} gets more benefit from *MaxScore* because C_{E-E} works better on pairs of event mentions that appear closely in the text, which activate more valuable learning features. We will report the results using the best approach of each classifier.

To evaluate our systems with event coreference knowledge, we first experimented our systems with gold event coreference as given by the ACE 2005 corpus. Table 2 shows the contribution of event coreference to our systems in the third group of the results. The results show that injecting knowledge from event coreference remarkably improved both the local classifiers and the joint inference model. Overall, the system that combined event coreference and the global inference model achieved the best performance, which significantly overtook all other compared systems. Specifically, it outperformed the baseline system, the local classifiers, and the joint inference model without event coreference with 80%, 25%, and 14% of relative improvement in F_1, respectively. It also consistently outperformed the local classifiers enhanced with event coreference. We note that the precision and recall of C_{E-T} in the joint inference model are the same because the inference model enforced each event mention to be associated with exactly one time interval. This is also true for the systems integrated with event coreference because our integration approaches assign only one time interval to an event mention.

We next move to experimenting with automatically learned event coreference systems. In this ex-

	Model	C_{E-T}			C_{E-E}			**Overall**		
		Prec.	*Rec.*	F_1	*Prec.*	*Rec.*	F_1	*Prec.*	*Rec.*	F_1
1	**Baseline**	33.29	33.29	33.29	20.86	32.81	25.03	27.06	33.05	29.16
	No Event Coref.									
2	Local classifiers	62.70	34.50	43.29	40.46	42.42	40.96	51.58	38.46	42.13
	Global inference	47.88	47.88	47.88	41.42	48.04	44.14	44.65	47.96	46.01
	With Gold Event Coref.									
3	Local classifiers	50.88	50.88	50.88	43.86	52.65	47.46	47.37	51.77	49.17
	Global inference	50.88	50.88	50.88	48.04	62.45	54.05	49.46	56.67	52.47
	With Learned Event Coref.									
4	Local classifiers	46.37	46.37	46.37	40.83	45.28	42.60	43.60	45.83	44.49
	Global inference	46.37	46.37	46.37	42.09	52.50	46.47	44.23	49.44	46.42

Table 2: Performance under various evaluation settings. All figures are averaged scores from 5-fold cross-validation experiments.

periment, we re-trained the event coreference system described in Chen et al. (2009) on all articles in the ACE 2005 corpus, excluding the 20 articles used in our data set. The performance of these systems are shown in the fourth group of the results in Table 2. The results show that by using a learned event coreference system, we achieved the same improvement trends as with gold event coreference. However, we did not obtain significant improvement when comparing with global inference without event coreference information. This result shows that the performance of an event coreference system can have a significant impact on the overall performance. While this suggests that a better event coreference system could potentially help the task more, it also opens the question whether event coreference can be benefited from our local classifiers through the use of a joint inference framework. We would like to leave this for future investigations.

5.4 Previous Work-Related Experiments

We also performed experiments using the same setting as in (Yoshikawa et al., 2009), which followed the guidelines of the TempEval challenges (Verhagen et al., 2007; Verhagen et al., 2010), on our saturated data. Several assumptions were made to simplify the task. For example, only main events in adjacent sentences are considered when identifying event-event relations. See (Yoshikawa et al., 2009) for more details. We performed 5-fold cross validation without event coreference. Overall, the system achieved 29.99 F_1 for the local classifiers and 34.69 when the global inference is used. These results are better than the baseline but underperform our full models where those simplification assumptions are not imposed, as shown in Table 2, indicating the importance of relaxing their assumptions in practice.

We also evaluated our C_{E-E} on the TimeBank corpus. We followed the settings of Chambers and Jurafsky (2008) to extract all event mention pairs that were annotated with *before* (or *ibefore*, "immediately before") and *after* (or *iafter*) relations in 183 news articles in the corpus. We trained and evaluated our C_{E-E} on these examples with the same feature set that we evaluated in our experiments above, with gold tense and aspect features but without event type. Following their work, we performed 10-fold cross validation. Our classifier achieved a micro-averaged accuracy of 73.45%, whereas Chambers and Jurafsky (2008) reported 66.8%. We next injected the knowledge of an event coreference system trained on the ACE2005 corpus into our C_{E-E}, and obtained a micro-averaged accuracy of 73.39%. It was not surprising that event coreference did not help in this dataset because: (i) different domains – the event coreference was trained on ACE 05 but applied on TimeBank, and (ii) different annotation guidelines on events in ACE 2005 and TimeBank.

Finally, we conducted an experiment that justifies the advantages of our interval-based inference model over a time point-based inference. To do this, we first converted our data in Table 1 from intervals to time points and infer the temporal relations between the annotated event mentions and the time points: *before*, *after*, *overlap*, and *unknown*. We modified the first component in the objective function in (3) to accommodate these temporal relations. We also made several changes to the constraints, including removing those in (7) since they are no longer required, and adding constraints that ensure

the relation between a time point and an event mention takes exactly one value. Proper changes were also made to other constraints in (10) to reflect the fact that time points are considered rather than intervals. We observed that experiment with such a formulation was unable to finish within 5 hours (we terminated the ILP inference after waiting for 5 hours), whereas our interval-based model finished the experiment with an average of 21 seconds per article.

6 Related Work

Research in temporal reasoning recently received much attention. Allen (1983) introduced an interval based temporal logic which has been used widely in the field. Recent efforts in building an annotated temporal corpus (Pustejovsky et al., 2003) has popularized the use of machine learning techniques for the task (Mani et al., 2006; Bethard et al., 2007). This corpus was later used (with simplifications) in two TempEval challenges (Verhagen et al., 2007; Verhagen et al., 2010). In these challenges, several temporal-related tasks were defined including the tasks of identifying the temporal relation between an event mention and a temporal expression in the same sentence, and recognizing temporal relations of pairs of event mentions in adjacent sentences. However, with several restrictions imposed to these tasks, the developed systems were not practical.

Recently, there has been much work attempting to leverage Allen's interval algebra of temporal relations to enforce global constraints on local predictions. The work of Tatu and Srikanth (2008) used global relational constraints to not only expand the training data but also identifies temporal inconsistencies to improve local classifiers. They used greedy search to select the most appropriate configuration of temporal relations among events and temporal expressions. For exact inferences, Bramsen et al. (2006), Chambers and Jurafsky (2008), Denis and Muller (2011), and Talukdar et al. (2012) formulated the temporal reasoning problem in an ILP. However, the inference models in their work were not a joint model involving multiple local classifiers but only one local classifier was involved in their objective functions.

The work of Yoshikawa et al. (2009) did formulate a joint inference model with Markov Logic Network (MLN). They, however, used the same setting as the TempEval challenges, thus only pairs of temporal entities in the same or adjacent sentences are considered. Our work, on the other hand, focuses on constructing an event timeline with time intervals, taking multiple local pairwise predictions into a joint inference model and removing the restrictions on the positions of the temporal entities. Furthermore, we propose for the first time to use event coreference and evaluate the importance of its role in the task of event timeline construction.

7 Conclusions and Future Work

We proposed an interval-based representation of the timeline of event mentions in an article. Our representation allowed us to formalize the joint inference model that can be solved efficiently, compared to a time point-based inference model, thus opening up the possibility of building more practical event temporal inference systems. Our inference model achieved significant improvement over the local classifiers. We also showed that event coreference can naturally support timeline construction, and good event coreference led to significant improvement in the system performance. Specifically, when such gold event coreference knowledge was injected into the model, a significant improvement in the overall performance could be obtained. While our experiments suggest that the temporal classifiers can potentially help enhance the performance of event coreference, in future work we would like to investigate into coupling event coreference with other components in a global inference framework.

Acknowledgments

The authors gratefully acknowledge the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract No. FA8750-09-C-0181, and the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053. The first author also thanks the Vietnam Education Foundation (VEF) for its sponsorship. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the VEF, DARPA, AFRL, ARL, or the US government.

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *ICSC*.

P. Bramsen, P. Deshpande, Y. K. Lee, and R. Barzilay. 2006. Inducing temporal graphs. In *EMNLP*.

N. Chambers and D. Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *EMNLP*.

Zheng Chen, Heng Ji, and Robert Haralick. 2009. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Workshop on Events in Emerging Text Types*.

J. Clarke and M. Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*.

Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *IJCAI*.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *ACL*.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK corpus. In *Corpus Linguistics*.

D. Roth and W. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. In *Introduction to Statistical Relational Learning*.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Coupled temporal scoping of relational facts. In *WSDM*.

Marta Tatu and Munirathnam Srikanth. 2008. Experiments with reasoning for temporal relations between events. In *COLING*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *SemEval-2007*.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *SemEval-2010*.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *ACL-IJCNLP*.

Ran Zhao, Quang Do, and Dan Roth. 2012. A robust shallow temporal reasoning system. In *NAACL-HLT Demo*.