Global Inference to Chinese Temporal Relation Extraction

Peifeng Li, Qiaoming Zhu, Guodong Zhou, Hongling Wang
School of Computer Science & Technology
Soochow University, Suzhou, 215006, China
{pfli, qmzhu, gdzhou, hlwang}@suda.edu.cn

Abstract

Previous studies on temporal relation extraction focus on mining sentence-level information or enforcing coherence on different temporal relation types among various event mentions in the same sentence or neighboring sentences, largely ignoring those discourse-level temporal relations in nonadjacent sentences. In this paper, we propose a discourse-level global inference model to mine those temporal relations between event mentions in document-level, especially in nonadjacent sentences. Moreover, we provide various kinds of discourse-level constraints, which derived from event semantics, to further improve our global inference model. Evaluation on a Chinese corpus justifies the effectiveness of our discourse-level global inference model over two strong baselines.

1 Introduction

Temporal relation extraction is to determine the temporal relationship (e.g., *Before* and *After*) holding among events. It has been drawing more and more attention due to the crucial importance of temporal information to various natural language processing (NLP) applications, such as language generation, information extraction, summarization, and question answering. The difficulty with this task is that temporal information about event mentions is sometimes not stated explicitly and one can only infer from their context. Currently, temporal relation extraction still remains a challenge in corpus construction and inference mechanism.

On one hand, although the TimeBank corpus (Pustejovsky et al., 2003), the commonly used corpus in previous studies, has largely promoted the development of temporal relation extraction, it only annotates a small subset of easily-identified event mention pairs. Moreover, it largely ignores almost all temporal relations between event mentions in nonadjacent sentences. These lead to fragmented relations and limit its applications to other NLP tasks, such as information extraction, and summarization. Finally, while constructing a fully-annotated corpus is expensive and time-consuming, many NLP tasks are normally interested in specific types of events. For example, a summarization or information extraction system on terrorism attacks may only concern with a few event types (e.g., *Attack*, *Die*, and *Injure*). Therefore, annotating an event-driven fully-annotated temporal relation corpus becomes a crucial issue to the success of real-life applications.

On the other hand, previous studies on temporal relation extraction focus on mining sentence-level information or enforcing coherence on different temporal relation types among various mentions in the same sentence or neighboring sentences, largely ignoring those discourse-level temporal relations in nonadjacent sentences. Specifically, only a few studies apply global inference models to exploit temporal relations in discourse level. Therefore, how to acquire discourse-level temporal information from those long-distance event mention pairs in nonadjacent sentences becomes another crucial issue to temporal relation extraction, especially for Chinese, as a discourse-driven language with a broad range of ellipsis and flexible sentence structures.

In this paper, we first annotate an event-driven fully-annotated Chinese temporal relation corpus, on the top of the ACE (Automatic Content Extraction) 2005 Chinese corpus. Then, we propose a discourse-

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1451–1460, Osaka, Japan, December 11-17 2016.

level global inference model to mine those temporal relations between event mentions in document-level (especially in nonadjacent sentences) with various kinds of discourse-level constraints, which derived from event semantics, to further improve the global inference model. Evaluation indicates the appropriateness of our event-driven fully-annotated Chinese corpus and justifies the effectiveness of our discourse-level global inference model over two strong baselines.

2 Related Work

In this section, we give a brief overview of temporal relation extraction from two aspects: corpus construction and inference mechanism.

2.1 Corpus Construction

Most of existing corpora for temporal relation extraction focus on English. As the commonly used corpus in temporal relation extraction, the TimeBank corpus (Pustejovsky et al., 2003) has been adopted in a series of TempEval competitions (Verhagen et al., 2007; Verhagen et al., 2010; Uz-Zaman et al., 2013), facilitating the development and evaluation of temporal relation extraction systems. The problems with the TimeBank corpus are that it only annotates a small subset of easily-identified event mention pairs and that it largely ignores those temporal relations between event mentions in nonadjacent sentences. These lead to fragmented relations and much limit its applications.

To overcome above problems, Do et al. (2012) produced an event-driven corpus on the ACE 2005 English corpus. However, "the annotator was not required to annotate all pairs of event mentions, but as many as possible", as stated in their paper. This makes the annotation inconsistent and difficult to follow. Recently, Cassidy et al. (2014) enriched the TimeBank-Dense corpus, on the top of TimeBank. Specifically, they approximated the completeness by labeling locally complete graphs over neighboring sentences.

In comparison, there are few corpora for Chinese temporal relation extraction. Li et al. (2004) annotated a Chinese corpus including 700 sentences. The TempEval-2 competition (Verhagen et al., 2010) provided 780 instances of Chinese temporal event relations. Obviously, both corpora are rather small and largely impede the research in Chinese temporal relation extraction. For example, no team participated in the TempEval-2 competition on Chinese temporal relation extraction.

2.2 Inference Mechanism

Due to the corpus limitation, previous studies on temporal relation extraction focus on inferring temporal relations between event mentions in the same sentence or neighboring sentences from English text, dominated by feature-based approaches. Mani et al. (2006) applied the temporal transitivity rule to greatly expand the corpus. Lapata and Lascarides (2006) introduced various kinds of syntactic and clause-ordering features to classify the temporal relationship. Chambers et al. (2007) used previously learned event attributes to classify the temporal relationship. Laokulrat et al. (2013), the best performing one in the TempEval-3 competition, applied various predicate-argument structure features from a deep syntactic parser to enhance their classifier. Mirza and Tonelli (2014) illustrated that simple features resulted in a better performance than sophisticated features. Chambers et al. (2014) proposed a sieve-based architecture to joint those different tasks of temporal relation extraction.

In comparison, few studies concern temporal relation extraction from Chinese text. Chen et al. (2008) used verbal attributes to identify temporal relations of verbs. Li et al. (2004) presented a classifier-based collaborative bootstrapping approach to analyze temporal relations in a small Chinese corpus.

While above studies focus on local information, a few studies sort to global inference, with focus on exploiting global information via various kinds of temporal logic reflexivity and transitivity constraints, using frameworks like Integer Linear Programming and Markov Logic Networks (Bramsen et al., 2006; Chambers and Jurafsky, 2008; Yoshikawa et al., 2009). However, their gains are rather small, largely due to the common disconnectedness in the sparsely annotated corpora (Chambers et al., 2014). To overcome this problem, Denis and Muller (2011) decomposed temporal entities into sub-graphs and enforced the coherence only within these substructures, while Do et al. (2012) proposed a joint event-event and event-time classification model to enforce various coreference constraints.

Different from previous studies, we build an event-driven fully-annotated Chinese corpus and propose

a discourse-level global inference model to extract temporal relations between event mentions in document level, especially in nonadjacent sentences. To our knowledge, this is the first attempt in discourse-level global inference for temporal relation extraction from an event-driven fully-annotated corpus.

3 Data Construction and Baseline

In this section, we present the construction of our Chinese temporal relation corpus and the learning-based baseline.

3.1 Data Construction

To address various problems in existing corpora, as described above, we build an event-driven fully-annotated Chinese temporal relation corpus, on the top of the ACE 2005 Chinese corpus with 8 predefined event types and 33 predefined event subtypes (e.g., *Die*, *Attack*, and *Transport*). That is, all other event mentions of non-predefined event types are ignored in our corpus.

Different from previous corpora, each document in our corpus is annotated with the temporal relations between the mentions of all the events relevant to concerned events in the document, with the constraint of event-relevant completeness. Besides, we focus on four temporal relations, i.e. *Before*, *After*, *Overlap* and *Unknown* (without relationship or with vague relationship). This is a simplification of the TimeBank corpus, which defines 14 temporal relations. Since differentiating 14 temporal relation types is too hard, even for a well-educated person, much work has been done to simplify the temporal relation types, e.g. 6 types (Mani et al., 2006; Chambers et al., 2007; Cassidy et al., 2014) and 4 types (Do et al., 2012; UzZaman et al., 2013 (Chinese subtask)). Our work is a typical practice of such tendency.

Specifically, 163 documents from the ACE 2005 Chinese corpus are selected as our experimental data, which contains 1166 event mentions. These documents are from three different data sources (i.e., Broadcast News, Newswire and WebLog), very different in various aspects, such as quality, length and style. Two postgraduates in computer science are involved in corpus annotation and the Kappa value between the two annotators is 0.70, similar to TimeBank's 0.71.

Table 1 shows 4 temporal relations and their occurrence frequencies in our event-driven fully-annotated corpus. The total number of event pairs in our corpus is three times larger than that of TimeBank. Since the ACE 2005 corpus is licensed, we cannot upload our annotated data. If anyone obtain the license of the ACE 2005 corpus, our corpus is free available for research purpose on request.

Type	Before	After	Overlap	Unknown	Total
#Number	7402	7402	4834	1494	21132

Table 1. The 4 temporal relations and their occurrence frequencies

We have implemented a tool to help the annotators to tag event relations easily and enforce the coherence in document level. Due to the reflexive property of event-event relationship, the annotators only need annotate half of the relations shown in Table 1. Besides, 7.1% of annotated event relations are *Unknown*, and this figure is much lower than that in TimeBank. The reason is that the relations between two ACE events of the predefined event types are relatively easy to be identified and this also verifies the relatively high Kappa value between the two annotators. In our corpus, the maximal size of the relations in a document is 625 (25 event mentions), while the minimal size is 2 (only 2 event mentions). If we ignore those *Unknown* relations, 32% of documents are not graph, but forest.

3.2 Baseline

Similar to the state-of-the-art system in temporal relation extraction, we employ a learning-based system as one of our baselines. As an event-event (E-E) classifier, this baseline predicts one of the four temporal relations, i.e. *Before(B)*, *After(A)*, *Overlap(O)*, and *Unknown(U)*, between two event mentions e_i and e_j as follows:

$$C_{E-E}(e_i, e_j) \rightarrow \{\underline{B}, \underline{A}, \underline{O}, \underline{U}\} \quad (1)$$

Besides those features adopted in English temporal relation extraction (e.g., D'Souza and Ng, 2014; Mirza and Tonelli, 2014), we also apply various kinds of Chinese-specific features to further boost the

performance of this baseline. Specifically, for each event mention pair <e_1, e_2> in a document, with trigger mentions t_1 and t_2 respectively, its feature set can be divided into 5 categories:

1) **Lexical features (14)**: the tokens of t_1 and t_2 (2); their POS tags (2); their preceding and succeeding words (4); the POS tags of their preceding and succeeding words (4); the hedge or negative word before t_1 or t_2 (2);
2) **Syntactic features (4)**: the dependency path between t_1 and t_2 (1); the governors of t_1 and t_2 (2); the constituent path between t_1 and t_2 (1);
3) **Event features (18)**: the tense, polarity, genericity, modality and event type of e_1 and e_2 (10); the agents (2), the patients (2), the times (2), and the places of e_1 and e_2 (2);
4) **Pairwise features (9)**: the conjunction between e_1 and e_2 (1); whether e_1 and e_2 are in the same sentence (1); whether e_1 and e_2 have the same tense (1), the same polarity (1), the same genericity (1), the same modality (1), the same event type (1), and the same *Time* argument (1); whether e_1 is before e_2 in the document (1);
5) **Semantic features (7)**: whether t_1 and t_2 are synonym (1); whether the agent of e_1 is the patient of e_2 (1); whether the agent of e_2 is the patient of e_1 (1); whether e_1 and e_2 have the same time (1), place (1), agent (1) or patient (1).

All the sentences in the corpus are divided into words using the word segmentation tool ICTCLAS. Besides, we use *Berkley Parser* and *Stanford Parser* to create the constituent and dependency parse trees respectively. The event features (e.g., trigger, event tense, event type, event arguments) are derived from the annotated data in the ACE 2005 Chinese corpus. After creating the training instances, we train four *one-vs-rest* classifiers using the Maximum Entropy tool *MaxEnt*.

4 Global Inference on Event Semantics

While existing approaches, as the baseline described above, focus on limited event mention pairs in the same sentence or neighboring sentences, our global inference model attempts to address those in non-adjacent sentences. In this section, we first present the discourse-level global inference model to temporal relation extraction and then introduce various kinds of discourse-level constraints to achieve global optimization on the temporal relations of event mention pairs in both nonadjacent and adjacent sentences.

4.1 Global Inference Model

To mine the interaction among events in a document, we optimize the predicted temporal graph, formed by prediction from $C_{E\text{-}E}$, with various kinds of discourse-level constraints derived from event semantics.

Let $E=\{e_1,e_2,\ldots,e_n\}$ denote the set of event mentions in a document, $\varepsilon=\{(e_i,e_j)\in E\times E|e_i,e_j\in E, i\neq j\}$ the set of event mention pairs, and $R=\{\underline{B},\underline{A},\underline{O},\underline{U}\}$ the set of temporal relations. Besides, let $P_{<i,j,r>}$ denote the prediction probability of (e_i,e_j) with relation r ($r\in R$), given by the event-event classifier $C_{E\text{-}E}$, and $x_{<i,j,r>}$ the binary indicator on the existence of relation r for (e_i,e_j). Following Roth and Yih (2004) and Li et al. (2013) in information extraction, we define the following log costs:

$$c_{<i,j,r>} = -\log(P_{<i,j,r>}) \tag{2}$$

$$\bar{c}_{<i,j,r>} = -\log(1-P_{<i,j,r>}) \tag{3}$$

Specifically, ILP (Integer Logical Programming), a global inference is employed to achieve global optimization with the following objective function to maximize over a document as follows:

$$\arg\min_{x} \sum_{(e_i,e_j)\in\varepsilon} \sum_{r\in R} (c_{<i,j,r>} \times x_{<i,j,r>} + (1-x_{<i,j,r>}) \times \bar{c}_{<i,j,r>}) \tag{4}$$

s.t.

$$x_{<i,j,r>} \in \{0,1\} \tag{5}$$

$$\sum_{r\in R} x_{<i,j,r>} = 1 \tag{6}$$

while binary constraint (5) ensures that $x_{<i,j,r>}$ is binary value and equality constraint (6) ensures that exactly only one temporal relation can be assigned to each event mention pair.

In addition, the reflexivity and transitivity constraints, as deployed in previous inference models

(Bramsen et al., 2006; Chambers and Jurafsky, 2008; Do et al., 2012), are also applied to our model as follows:

$$x_{<i,j,r>} - x_{<j,i,\bar{r}>} = 0 \qquad \forall r, \bar{r} \in R \tag{7}$$

$$x_{<i,j,r>} + x_{<j,k,r>} - x_{<i,k,r>} \leq 1 \qquad \forall r \in \{\underline{B}, \underline{A}, O\} \tag{8}$$

Here, **reflexivity constraint** (7) enforces the reflexive property of the event-event relationship, where relation $\bar{r}$ denotes inverse relation r with possible (r, $\bar{r}$) pairs { ($\underline{A}$,$\underline{B}$) , ($\underline{B}$,$\underline{A}$) , ($\underline{U}$,$\underline{U}$) }, and **transitivity constraint** (8) states that if both event mention pairs (e_i,e_j) and (e_j,e_k) have the same temporal relation r, temporal relation r must hold between e_i and e_k, with event mention e_j as a bridge to link e_i and e_k.

4.2 Discourse-level Constraints

Different from the TimeBank corpus which only annotates the temporal relations in the same sentence or neighboring sentences, our corpus is event-driven fully-annotated. That is, besides the temporal relations in the same sentence or neighboring sentences, our corpus also contains those in nonadjacent sentences, which occupy 56.3%. This poses the great necessity to address those temporal relations in nonadjacent sentences. Besides, although all the event types in the ACE corpus have a *Time* role, the statistics on our corpus shows that only 35.9% of event mentions have explicit *Time* arguments. This poses the great challenge to address those temporal relations in nonadjacent sentences due to the frequent lack of explicit *Time* arguments.

Motivated by the intuition that the intrinsic semantics of event mentions is helpful to reveal their temporal relations due to the semantic nature in the event definition, we propose various kinds of discourse-level constraints on time arguments, event relevance, event tense, discourse connective, and coreference to mine the temporal relations in both nonadjacent and adjacent sentences.

Argument *Time* constraint

Generally, an event can be expressed as "5W1H" (*Who*, *What*, *Whom*, *Where*, *When* and *How*). *When*, one of "5W", indicates the time an event happens. Naturally, this argument is the solid evidence to identify the temporal relation between two event mentions. For example, if the *Time* argument of one event mention e_1 is "今日" (today) and that of the other mention e_2 is "昨日" (yesterday), it is obvious that the relation between e_1 and e_2 is *After*.

For time arguments, we can obviously have the following constraint, stating that if *Time* argument at_i of event mention e_i is before *Time* argument at_j of event mention e_j, the temporal relation between e_i and e_j is $\underline{B}$, and if at_i is equal to at_j, or they have overlap part, the temporal relation between e_i and e_j is $\underline{O}$.

$$x_{<i,j,r>} = 1 \qquad \forall r \in \{\underline{B}, \underline{O}\} \wedge r = rel(at_i, at_j) \tag{9}$$

where function $rel(at_i,at_j)$ returns one of the four temporal relations between at_i and at_j. Due to the reflexivity constraint, it is unnecessary to enforce the constraint on *after* relation.

Since the ACE corpus uses *Timex2* to annotate all temporal expressions, the *Time* arguments need to be normalized. In this study, we first divide all time tags into two categories: time point and time duration. Then, we implement a simple rule-based tool based on the DCT (Document Create Time) to normalize all time points as "*year:month:day: hour:minute*" and all time durations as (*begintime, endtime*) where *begintime* and *endtime* are normalized as the style of time point. As a result, 92.7% of *Time* arguments are normalized correctly.

Event relevance constraint

In a discourse, most of event mentions are normally structured around a specific topic, which acts as a bone to link all the relevant event mentions together into a narration, via various kinds of event relations. Those semantics-based event relations, i.e. event relevance, are thus helpful to infer the temporal relations among event mentions. For example, if there is a causal relation between an *Attack* and a *Die* event mention, it is obvious to infer they have the *Before* temporal relation. That is, a *Die* event is always the result of an *Attack* event.

Specifically, we learn event relevance from the training set by counting the occurrence frequency $f_{<i,j,r>}$ for each event type pair (evt_i, evt_j) (e.g., (*Attack, Die)*) with relation r in the training set. To eliminate

the accidental factor in statistics, we modify the occurrence frequency of an event type pair to 0 when it only appears once in the training set.

Accordingly, we have the following constraint on event relevance, stating that if an event type pair (evt_i, evt_j) only has one occurrence frequency $f_{<i,j,r>}$(larger than 0), the temporal relations of all event mention pairs in the test set with event type pairs (evt_i, evt_j), are assigned with the temporal relation r according to constraint (10), and that if an event type pair (evt_i, evt_j) has two or three occurrence frequencies (i.e., larger than 0), the relations of all event mention pairs with event type pair (evt_i, evt_j), are enforced according to constraint (11).

$$x_{<i,j,r>} = 1 \qquad \forall f_{<i,j,r>} > 0 \wedge \sum_{r' \in SR_1} f_{<i,j,r'>} = 0 \tag{10}$$

$$\sum_{r \in SR_2} x_{<i,j,r>} = 1 \qquad \forall\, f_{<i,j,r>} > 0 \tag{11}$$

where SR_1 refers to the set of all the temporal relations except r and SR_2 refers to the set of all temporal relations whose occurrence frequencies are larger than 0.

Tense constraint

Event tense is also a helpful evidence to infer temporal relations. In the ACE 2005 corpus, each event mention has an annotated tense attribute, whose values are *Past(P)*, *Present(R)* and *Future(F)* and have been used in the baseline. For example, it is normal to infer the *Before* relation between two event mentions whose tense are *Past* and *Future* respectively. Accordingly, we can have the following constraint, stating that if the tense te_i of e_i is *Past* and that of e_j is *Present* or *Future*, the temporal relation of (e_i, e_j) is $\underline{B}$; 2), and that if te_i is *Present* and te_j is *Future*, the temporal relation of (e_i, e_j) is $\underline{B}$.

$$x_{<i,j,\underline{B}>} = 1 \qquad \forall te_i = P \wedge te_j \in \{R, F\} \vee te_i = R \wedge te_j = F \tag{12}$$

Connective constraint

In a discourse, the connective between two adjacent sentences or clauses can largely reveal their discourse relations. For example, the connective "because" illustrates the *Cause* relation. Likewise, the connective between two adjacent event mentions also explicitly unveils their temporal relation. For example, if the preceding event mention is the cause of the succeeding event mention, their temporal relation is *Before*. Besides, we find out that some verbs which represent the meaning of causality can indicate the temporal relation of an event mention pair. Take the following sentence as an example:

*E1: 这起炸弹攻击(EV1: Attack)事件**造成**了 2 个人死亡(EV2: Die)。(This bomb terror (EV1) **caused** two persons to death (EV2).)* *-From CBS20001120.1000.0823*

In sentence E1, the verb 造成 (cause) indicates that the temporal relation between the event mentions *EV1* and *EV2* is *Before*. Hence, we enumerate a set of Chinese verbs (e.g., 导致, 造成, 引起) whose meaning are "cause" and add them into our connective set *CS*. Besides, we find out that only causal and temporal connectives are helpful to infer temporal relations. Therefore, respective connectives are selected from Appendix B of the PDTB 2.0 annotation manual, which provides a list of classified explicit connectives. Finally, we divided all words in CS into two subsets CS_1 and CS_2, according to the statistics from the training set, where all words in CS_1 indicate that the preceding event mention occurs earlier than the succeeding one and all words in CS_2 indicate that the preceding event mention occurs later than the succeeding one.

Accordingly, we have the following constraint on discourse connective, stating that if there is a connective *con* (*con*$\in$*CS*) between two adjacent event mentions e_i and e_j in the same sentence or neighboring sentences, the temporal relation between e_i and e_j depends on whether *con* belongs to CS_1 or CS_2 as follows.

$$\begin{aligned} x_{<i,j,\underline{B}>} &= 1 \qquad \forall con \in CS_1 \\ x_{<i,j,\underline{A}>} &= 1 \qquad \forall con \in CS_2 \end{aligned} \tag{13}$$

Coreference constraint

An event may have more than one mention in a document and these mentions refer to the same event, called coreference events. Take the following two sentences as examples:

E2: 埃塞俄比亚与厄立特里亚 2 3 日在这里举行谈判(EV3: Meet)。(The talk (EV3) between Ethiopia and Eritrea will be held on 23rd.)

E3: 这次谈判(EV4: Meet)的目的是... (The goal of this talk (EV4) is ...) -From XIN20001024.2000.0141

It is obvious that two coreference event mentions EV3 and EV4 must have the same occurrence time and their relation is *Overlap*. Following Do et al. (2012), we also apply this constraint to our model and enforce the following constraint on event coreference, stating that if mentions e_i and e_j are coreferential event mentions, their temporal relation is *Overlap*.

$$x_{<i,j,O>} = 1 \qquad \forall\, cr(e_i, e_j) = true \tag{14}$$

where function *cr()* return true when e_i and e_j are coreferential. Besides, we use the tool described in Teng et. al. (2015) to construct those coreference event chains.

5 Experimentation

In this section, we first evaluate our model for Chinese temporal relation extraction and then report the experimental results on our event-driven fully-annotated Chinese temporal relation corpus.

5.1 Experimental Settings

All the experiments are done on the event-driven fully-annotated Chinese temporal relation corpus, annotated on the top of the ACE 2005 Chinese corpus, as described in Subsection 3.1. We conduct all evaluations with 5-fold cross-validation at document level and each fold contains about 4000 event mention pairs. Following previous studies on temporal relation extraction, we employ *Accuracy* as evaluation metric, which measures the percentage of correctly classified test instances. This metric is the same as micro F-score since each temporal relation of each event mention pair must belong to one of the four relations.

In this study, we use *lp_solve* as the ILP solver which implements the Branch-and-Bound algorithm. It takes less than 3 seconds on average to decode a document on a PC with 3.4Ghz Intel i7 CPU and 16GB memory.

5.2 Experimental Results

Performance comparison

To evaluate the performance of our discourse-level global inference model (DGIM), we compare it with two strong baselines. The first is a classifier-based system, mentioned in Subsection 3.2, originated from the top performing English systems (D'Souza and Ng, 2014; Mirza and Tonelli, 2014). The second (GIM) is a global inference model with the reflexivity and transitivity constraints following Bramsen et al. (2006), Chambers and Jurafsky (2007), and Do et al. (2012). It is worth to note that all annotated data in DGIM are also used in the first classifier-based system. Table 2 compares the performance of two baselines and our inference model DGIM.

Model	Accuracy (%)(Gold events)	Accuracy (%)(Auto events)
Baseline 1 ($C_{E\text{-}E}$)	62.17	36.21
Baseline 2 (GIM)	64.12	37.85
DGIM	68.36	40.92

Table 2. Performance comparison of different models

Table 2 shows that when all event mentions are known, i.e. with gold event mentions, DGIM significantly outperform the two baselines by 6.19% and 4.71% in accuracy respectively. All improvements from two baselines to DGIM are statistically significant ($p<0.000001$, McNemar's test, 2-tailed). Besides, GIM outperforms the classifier-based model by 2.05% in accuracy, indicating the limitation of

the conventional reflexivity and transitivity constraints in previous studies. In comparison, DGIM outperforms GIM by 4.24% in accuracy, indicating the effectiveness of our various discourse-level global constraints.

Table 2 also shows the performance comparison consistency when all the event mentions are automatically extracted, as described in (Li et al., 2013) with the F1-score of 68.2% and 53.7% in event trigger extraction and event argument extraction respectively.

Contributions of discourse-level constraints

Table 3 illustrates the contributions of different discourse-level constraints to our DGIM model with gold event mentions.

System	Accuracy (%)
Baseline 1	62.17
+Reflexivity (7)	+0.29
+Before/After Transitivity (8)	+0.54
+Argument Time (9)	+2.23
+Event relevance (10,11)	+1.26
+Tense (12)	+0.48
+Connective (13)	+0.74
+Coreference (Learned) (14)	+0.69
+Coreference (Gold) (14)	+0.86

Table 3. Contributions of different discourse-level constraints to temporal relation extraction

1) The conventional reflexivity and before/after transitivity constraints slightly improve the accuracy. This is not as effective as that on TimeBank, due to that those wrong probabilities produced by the event-event classifier will be incorrectly propagated to more temporal relations of event mention pairs since each document in our corpus is fully-annotated. Although the improvements of above two constraints are limited, they can interact with others to either improve the performance or reduce the time complexity. For example, the reflexivity constraint can simplify our discourse-level constraints, since if we have applied a constraint to an event pair, it is unnecessary to apply the opposite constraint to their inverses.
2) The argument *Time* constraint gains most with 2.23% in accuracy, while the tense constraint gains least among all constraints. Different from TimeBank, an event always has a *Time* role in the ACE 2005 corpus. If both event mentions have the *Time* arguments, we have a high confidence to determine their temporal relation. The error of the argument *Time* constraint mainly comes from those event mentions with a vague time (e.g., 最近 (recently), 日前 (a few days ago)).
3) Intuitively, tense can clearly identify the temporal relation of two event mentions if they have different tenses. However, our preliminary experiment shows that this constraint harms the accuracy if we apply it to the whole document. The reason is that the tenses annotated in the ACE 2005 corpus are relative ones based on the statement of a sentence itself. For example, although two *Transport* event mentions "他要**来**美国" (He will **come** to U.S.) and "他**来**到了美国" (He **arrive**d U.S.) have the *Future* and *Past* tenses respectively, they are coreferential events with different statement times. In this study, we only enforce this constraint on the sentence level.
4) The event relevance constraint gains an improvement of 1.26% in accuracy. This verifies that relevant events always occur in a regular order. In our experiments, we extract total 65 event type pairs to construct this constraint. For example, an *Arrest-Jail* event often occurs after an *Attack* event. Although this constraint contributes third, this is far from our expectation. Our error analysis shows that this constraint introduces lots of wrong predictions due to the lack of deep semantics, which is worth exploring in our future work.
5) Although the improvement of the connective constraint is not significant enough with a gain of 0.74% in accuracy, it achieves a high precision in predicting almost all event mention pairs enforced by this constraint, through discourse connective like "因为" (because), "后" (after) and "造成" (cause).
6) The conference constraint gains an improvement of 0.69% and 0.86% in accuracy with automatically learned (with F1-score of 61.7% using the tool described in Teng et. al. (2015)) and gold conference respectively. These figures are much smaller than those in Do et al. (2012) (2.33% for

learned 9.91% for gold). This is largely due to that although 40% of event mentions in our corpus are coreferential, the accuracy of the temporal relation among two coreferential event mentions, produced by the baseline classifier-based model, is already very high (86.2%). In comparison, the baseline in Do et al. (2012) only achieves ~40% in accuracy. Besides, Do et al. (2012) employed a much smaller corpus of only 20 documents, in comparison with 163 documents in our study.

Performance of different temporal relations

Table 4 shows the accuracy on four temporal relations with gold event mentions. From Table 4, the performance of temporal relations *Before*, *After* and *Overlap* is higher than that of *Unknown*, much due to the low recall of the *Unknown* relation, caused by its low percentage (7.1%). Compared to the two baselines, our DGIM improves the F1-scores for all temporal relations, with the highest improvement on the *Before* and *After* relations and the lowest improvement on the *Unknown* relation, much due to the fact that almost all discourse-level constraints focus on relations *Before*, *After* and *Overlap*.

Relation	Baseline 1	Baseline 2	DGIM
Before	63.43	65.30	72.21
After	63.44	65.30	72.21
Overlap	64.50	66.40	69.17
Unknown	45.60	47.20	47.44

Table 4 Accuracies (%) of four temporal relations

Analysis on adjacent or nonadjacent sentences

Table 5 shows the percentages and performance of event mention pairs in same sentence, adjacent sentences and nonadjacent sentences with gold event mentions. We can find out that 56.3% of event mention pairs are in nonadjacent sentences. Those event mention pairs in the same sentence achieve the highest accuracy while those in nonadjacent sentences gains least among all three types. Table 5 also proves that our DGIM outperforms two baselines in all three sentence levels significantly.

Distance	Rate(%)	Baseline 1	Baseline 2	DGIM
Same	18.6	68.13	69.13	72.25
Adjacent	25.1	64.54	65.47	69.09
Nonadjacent	56.3	59.15	61.87	66.75

Table 5. Accuracies (%) of event mention pairs in same sentence (Same), adjacent sentences (Adjacent) and nonadjacent sentences (Nonadjacent)

6 Conclusion

This paper first annotates an event-driven fully-annotated Chinese temporal relation corpus and then presents a novel discourse-level global inference model, enforced by various kinds of discourse-level constraints derived from event semantics, to recognize temporal relations of Chinese events in document-level, especially in nonadjacent sentences. Evaluation on an event-driven fully-annotated Chinese temporal relation corpus justifies the effectiveness of our discourse-level global inference model over two strong baselines.

Although our model focuses on Chinese, it can be naturally applied to other languages (e.g., English). Our future work will focus on how to introduce more linguistics-driven knowledge to boost our model and construct a joint modelling of temporal event relation extraction and event extraction on both Chinese and English.

Acknowledgements

The authors would like to thank three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant No. 61472265, No. 61402314 and No. 61331011, and partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

Reference

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. *In Proceedings of EMNLP 2006*, pages 189-198, Sydney, Australia.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. *In Proceedings of ACL 2014*, pages 501-506, Baltimore, MD, USA.

Nathanael Chambers, ShanWang and Dan Jurafsky. 2007. Classifying temporal relations between events. *In Proceedings of ACL 2007*, pages 173-176, Prague, Czech Republic.

Nathanael Chambers and Dan Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. *In Proceedings of EMNLP 2008*, pages 698-706, Singapore.

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.

Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto. 2008. Use of event types for temporal relation identification in Chinese text. *In Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing*, pages 31-38, Hyderabad, India.

Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. *In Proceeding of IJCAI 2011*, pages 1788-1793, Barcelona, Spain.

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. *In Proceedings of NAACL-HLT 2013*, pages 918-927, Atlanta, Georgia.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. *In Proceedings of EMNLP 2012*, pages 677-687, Jeju Island, Korea.

Natsuda Laokulrat, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. Uttime: Temporal relation classification using deep syntactic features. *In Proceedings of the 7th International Workshop on Semantic Evaluation*, pages 88-92, Atlanta, Georgia, USA.

Mirella Lapata and Alex Lascarides. 2006. Learning sentenceinternal temporal relations. *Journal of AI Research*, 27:85-117.

Wenjie Li, Kam-Fai Wong, Guihong Cao and Chunfa Yuan. 2004. Applying machine learning to Chinese temporal relation resolution. *In Proceedings of ACL 2004*, pages 582-588, Barcelona, Spain.

Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2013. Joint modeling of argument identification and role determination in Chinese event extraction with discourse-level information. *In Proceedings of IJCAI 2013*, pages 2120-2126, Beijing, China.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. *In Proceedings of ACL 2006*, pages 753-760, Sydney, Australia.

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

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, et al. 2003. The TimeBank corpus. *Corpus linguistics*, 647-656.

Dan Roth and wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. *In Proceedings of CoNLL 2004*, pages 1-8, Boston, MA, USA.

Jiayue Teng, Peifeng Li, Qiaoming Zhu. 2015. Chinese event co-reference resolution based on trigger semantics and combined features, *In Proceedings of Chinese Lexical Semantic Workshop (CLSW 2015)*, pages 494-503, Beijing, China.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. *In Proceedings of the 7th International Workshop on Semantic Evaluation*, pages 1-9, Atlanta, Georgia, USA.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75-80, Stroudsburg, PA, USA.

Marc Verhagen, Roser Saur´ı, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57-62, Stroudsburg, PA, USA.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with Markov Logic. *In Proceedings of the Joint Conference of ACL-AFNLP*, pages 405-413, Singapore.