TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events

Aakanksha Naik **Luke Breitfeller** **Carolyn Rose**
Language Technologies Institute, Carnegie Mellon University
{anaik, mbreitfe, cprose}@cs.cmu.edu

Abstract

Prior work on temporal relation classification has focused extensively on event pairs in the same or adjacent sentences (local), paying scant attention to discourse-level (global) pairs. This restricts the ability of systems to learn temporal links between global pairs, since reliance on local syntactic features suffices to achieve reasonable performance on existing datasets. However, systems should be capable of incorporating cues from document-level structure to assign temporal relations. In this work, we take a first step towards discourse-level temporal ordering by creating **TDDiscourse**, the first dataset focusing specifically on temporal links between event pairs which are more than one sentence apart. We create TDDiscourse by augmenting TimeBank-Dense, a corpus of English news articles, manually annotating global pairs that cannot be inferred automatically from existing annotations. Our annotations double the number of temporal links in TimeBank-Dense, while possessing several desirable properties such as focusing on long-distance pairs and not being automatically inferable. We adapt and benchmark the performance of three state-of-the-art models on TDDiscourse and observe that existing systems indeed find discourse-level temporal ordering harder.

1 Introduction

Temporal ordering of events is a crucial problem in automated text analysis. Systems capable of performing this task find widespread applicability in areas such as time-aware summarization, temporal information extraction or event timeline construction. Prior work has focused extensively on creating annotated corpora for temporal ordering, some notable efforts being the development of the TimeML annotation schema (Pustejovsky et al., 2003), TimeBank (Pustejovsky et al.) and TimeBank-Dense (Cassidy et al., 2014). However, most work has focused mainly on local ordering, i.e., events present in the same or adjacent sentences. This leads to a major drawback, also pointed out by Reimers et al. (2016). Low prevalence of global discourse-level temporal ordering annotation in existing datasets allows systems to achieve moderate performance simply using local syntactic cues. Having more global annotations would require systems to incorporate global consistency and assimilate features from document-level structure and flow to achieve high performance, thus presenting a more challenging task. In this work, we present TDDiscourse, a dataset focused on discourse-level temporal ordering.

We create TDDiscourse by augmenting TimeBank-Dense (Cassidy et al., 2014), a corpus of English news articles, with more long-distance event pair annotations. Our work makes the first attempt to *explicitly* annotate relations between event pairs that are more than one sentence apart, a more difficult annotation task than previous datasets. In addition to facing similar challenges as prior work (eg: hypothetical/negated events (Cassidy et al., 2014)), we tackle new *global discourse-level* issues such as incorporating event coreference and causality/prerequisite links arising from world knowledge. To handle these, we design a careful coding scheme that achieves high inter-annotator agreement (Cohen's Kappa of 0.69 on the test set). However, getting expert manual annotation for all possible long-distance event pairs is expensive. Moreover, it is possible to leverage annotations from existing datasets to automatically infer temporal relations for certain event pairs. To make optimal use of expert annotation, we develop a heuristic algorithm for automatic inference of temporal relations using EventTime (Reimers et al., 2016) and apply this

Proceedings of the SIGDial 2019 Conference, pages 239–249
Stockholm, Sweden, 11-13 September 2019.

to all documents.[1] We then randomly subsample the unannotated event pairs and source expert annotations for those. At 6150 pairs, our manually annotated subset is of the same size as TimeBank-Dense. Adding the automatic subset makes our dataset 7x larger (§6). Finally, we perform a principled comparison between manual and automatic pairs by annotating 3 test documents (107 manual and 110 automatic event pairs) with phenomena required to reason correctly about the pair. These annotations suggest that our manual subset exhibits a high proportion of global discourse-level phenomena such as reasoning about chains of events.

In addition to developing TDDiscourse, we adapt three state-of-the-art models on TimeBank-Dense for discourse-level temporal ordering and benchmark their performance on our data, separating scores on manual and automatic subsets. We observe that models perform worse on average on TDDiscourse, with none beating a majority class baseline on the manual subset. A manual analysis of model errors reveals key shortcomings of current temporal ordering techniques. We offer our dataset[2] as a challenging new resource for the temporal ordering community and hope that insights from our analysis will spark interest in the development of more global discourse-aware models.

2 Related Work

2.1 Prior Work on Temporal Annotation

The development of TimeML (Pustejovsky et al., 2003) and TimeBank (Pustejovsky et al.) marked the first attempt towards creating a corpus for temporal ordering of events. TimeML uses temporal links (TLINKs) (Setzer, 2002), to represent ordering. A TLINK expresses the temporal relation between two events. For example, an event *e1* can occur *before* another event *e2*. TimeBank is annotated using TLINKs, but the number of possible TLINKs in a document is large (quadratic in number of events). So annotation is restricted to a subset of TLINKs, leading to sparsity. To combat this, several works attempted to create denser corpora (Bramsen et al., 2006; Kolomiyets et al., 2012; Do et al., 2012; Cassidy et al., 2014), but still focused largely on local TLINKs.

Reimers et al. (2016) addressed high annotation cost by proposing a new scheme in which events were associated with explicit time expressions. Annotation effort now scaled linearly with number of events, making it feasible to annotate all of them. Using this scheme, they created EventTime, which had some discourse-level temporal annotation. However this dataset had one major drawback: events which could not be associated with a time expression were ignored. We observed that it may not always be possible to determine specific times for an event, but ordering it with respect to other events is often possible based on world knowledge. For example, consider the snippet: "Police discover body of *kidnapped* man. Police found the man's *dismembered* body wrapped in garbage bags". In this text, *dismembered* cannot be associated with a time. But the temporal relation between *dismembered* and *kidnapped* is clear because the kidnapping should have happened *before* dismembering. Based on this, we address the drawback in EventTime, by using TLINK-based annotation, which is expensive but allows more expressive power. Following TimeML, we augment TimeBank-Dense (Cassidy et al., 2014) with global discourse-level TLINKs. To optimize manual effort, we automatically generate all TLINKs that can be inferred from EventTime. Then, we manually annotate a large subset of missing TLINKs involving events not associated with specific dates.

Most recently, Ning et al. (2018b) proposed a new scheme, which labels TLINKs based only on event start time. This improved inter-annotator agreement allowing for crowdsourcing of long-distance annotations at lower cost. However, they focused only on verb events, whereas our work is broader in scope and poses no such restrictions.

2.2 Prior Temporal Ordering Systems

TimeBank and the TempEval tasks (Verhagen et al., 2007, 2010; UzZaman et al., 2013) spurred the development of many temporal ordering systems (UzZaman and Allen, 2010; Llorens et al., 2010; Strötgen and Gertz, 2010; Chang and Manning, 2012; Chambers, 2013; Bethard, 2013). More recently, TimeBank-Dense and EventTime prompted development of newer models (Chambers et al., 2014; Mirza and Tonelli, 2016; Cheng and Miyao, 2017; Reimers et al., 2018). Most systems built for TimeBank/ TimeBank-Dense focus

[1] We validate our algorithm by obtaining human annotations for a subset of 100 examples and observing agreement with the generated label in 99% cases

[2] https://github.com/aakanksha19/TDDiscourse

on TLINKs between events in the same or adjacent sentences, relying on local features rather than document-level structure, with some exceptions. Chambers and Jurafsky (2008); Denis and Muller (2011); Ning et al. (2017) introduce document-level consistency via integer linear programming constraints. Bramsen et al. (2006); Do et al. (2012) also incorporate document-level structure, but focus on different corpora. Reimers et al. (2018) develop a model for EventTime, which uses a decision tree of CNNs to associate each event from a document with a time. Several works have explored techniques to incorporate document-level cues such as event coreference (Do et al., 2012; Llorens et al., 2015) and causality (Do et al., 2012; Ning et al., 2018a) in temporal ordering systems. However, due to a lack of standard datasets focusing on global discourse-level links, most work has been evaluated on datasets of their own creation or standard datasets with mainly local TLINKs. This further stresses the need for a standardized benchmarking effort, which we address by evaluating adaptations of three state-of-the-art systems on our dataset (§8).

3 Constructing TDDiscourse

To emphasize the need for a global discourse-level focus in temporal ordering, we develop TDDiscourse, the first dataset which focuses *explicitly* on TLINK annotations between event pairs that are more than one sentence apart. To create TDDiscourse, we augment a subset of documents from TimeBank with global TLINKs. We use the same set of 36 documents as TimeBank-Dense (Cassidy et al., 2014) and EventTime (Reimers et al., 2016) to facilitate comparison with previous work. We also utilize the same set of temporal relations as TimeBank-Dense.[3] Table 1 gives a brief summary of these relations. To add global links, we use two approaches:

- **Manual annotation:** We manually label a subset of global TLINKs using document cues, world knowledge and causality (§4). To optimize human effort, we ensure that these TLINKs are not automatically inferable.
- **Automatic inference:** We use a heuristic algorithm to automatically label global TLINKs using EventTime (§5) annotations, to generate a large number of links at low cost.

Symbol	Relation
a	*e1* occurs **after** *e2*
b	*e1* occurs **before** *e2*
s	*e1* and *e2* are **simultaneous**
i	*e1* **includes** *e2*
ii	*e1* **is included** in *e2*

Table 1: Temporal relation set used in TDDiscourse. All relations are mutually exclusive.

4 Manual Annotation

In this phase, we ask experts[4] to label discourse-level TLINKs that cannot be inferred automatically.[5] Getting expert annotation for all missing TLINKs is expensive. Hence, we randomly subsample TLINKs not annotated by TimeBank-Dense or automatic inference. This subsample is as large as TimeBank-Dense, thus doubling the data size while making the overall task harder (see §8). Note that TLINKs annotated in this phase may involve events for which a specific time of occurrence cannot be determined, which were ignored in EventTime. We refer to this subset as **TDD-Man**.

Since TLINKs are not restricted to the same or adjacent sentences, our annotation task becomes harder, requiring cues from the entire document. Many TLINKs also require the use of causal links and world knowledge to label the relation. Based on our observations, we develop a coding scheme. To ensure high inter-annotator agreement, we refine our scheme over multiple rounds of annotation and discussion of disagreements.

4.1 Coding Scheme

Our scheme reduces the task of labeling a TLINK to a set of concrete decision steps:

1. Using textual cues
2. Using world knowledge
3. Using narrative ordering

A TLINK may be assigned a label at any step. If it cannot be assigned a label, it moves on to the next step. Information from previous steps is retained, making it possible to combine multiple sources of evidence. For example, textual cues may not suffice, but they can be used in conjunction with world knowledge to label a pair. We

[3] We discard the "vague" label since we do not require annotators to label all event pairs

[4] Expert annotators are the authors of the paper, with a background in computational linguistics

[5] The automatic inference algorithm is explained in §5

Snippet
Atlanta nineteen ninety-six. A bomb **blast shocks** the Olympic games. One person is **killed**. January nineteen ninety-seven. Atlanta again. This time a **bomb** at an abortion clinic. More people are **hurt**.
Event pair: *blast*, *hurt*
Relation: before
Textual cues: Event *blast* occurred in 1996. Event *hurt* occurred because of second bomb blast in 1997.

Table 2: Sample document-level textual cues used during temporal annotation

choose to organize our coding scheme as mentioned above, to make the process of gathering evidence about an event pair systematic, and ensure that experts do not miss important cues. The final step is guaranteed to assign a label. We choose not to allow annotators to leave event pairs unlabeled or label them "vague", to keep them from overusing this option. Owing to this decision, we need to develop mechanisms for handling TLINLKs containing events which have not actually occurred (eg: negated, hypothetical or conditional events). Drawing from prior work, we interpret these events using a *possible worlds* analysis, in which the event is treated as if it has occurred. We refer interested readers to (Chambers et al., 2014) for a more detailed discussion.

4.1.1 Using textual cues

In this step, we use document-level textual cues to label a TLINK. The cues used are similar to those used in previous datasets (Cassidy et al., 2014). Table 2 gives an example of the types of cues used.

A key textual cue we use here is event coreference. Event coreference has not been used for annotation because the occurrence of coreferent events in adjacent sentences is rare. However, this cue is crucial for global discourse-level annotation. Since TimeBank does not contain event coreference annotation, we develop a procedure to annotate our document subset. Our procedure is based on the ERE (Entities, Relations, and Events) scheme (Song et al., 2015), which cannot be directly used for TimeBank due to differing notions of what constitutes an event and different metadata. In our procedure, events are considered coreferent iff they share the following:

- Entities involved in the event
- Temporal attributes
- Location attributes
- Realis (whether event is real or hypothetical)

Events which are synonymous in context are also considered coreferent (for instance, in "...held an interview Monday. The segment covered...", *interview* and *segment* are synonymous). These attributes (barring temporal) are not provided in TimeBank and must be inferred. Often, an event may only have partial information about these attributes - here we use human judgment. Our definition of coreference is closer to the strict notion of "event identity" in Light ERE than the relaxed definition in Rich ERE.[6] To test our procedure, we select all "simultaneous" TLINKs from TimeBank-Dense to ensure that our sample contains a sizeable proportion of *possibly* coreferent event pairs. The corpus contains 179 "simultaneous" links, of which 93 are event pair TLINKs. Our first annotation pass achieves high agreement between two annotators, with a Kappa of 0.70. We refine our guidelines through an adjudication step, reaching perfect agreement on this sample. Post-adjudication guidelines are used to annotate event coreference for all documents. Resulting annotations are used as textual cues in our scheme. Based on textual cues, an appropriate label is assigned to a TLINK. Coreferent TLINKs are labeled "simultaneous". Unlabeled links move on to the next decision step.

4.1.2 Using world knowledge

This step uses real world knowledge to determine causal/prerequisite links which are used to label a TLINK. We consider both events in the TLINK and determine whether they possess one or both of the following:

- **Causal Link:** Two events have a causal link if the occurrence of one event results in the other event coming about. For example, in the sentence "The paper got wet when I spilled water on it", the event pair (spilled, wet) have a causal link.
- **Prerequisite Link:** Two events have a prerequisite link if one event *must* occur before the other can happen. For example, in the sentence "We cooked dinner and ate it", the event pair (cooked, ate) have a prerequisite link. Note that we use the knowledge that a meal must be cooked before it can be eaten, though it is not explicitly mentioned.

We examine the event pair in the context of the entire document to detect causal/prerequisite links, also allowing weak or transitive links. For in-

[6] Examples in the appendix

Rule	Label
TLINK=(A, B), A=P	Before
TLINK=(A, B), A=I	Includes
TLINK=(B, A), A=P	After
TLINK=(B, A), A=I	Is Included

Table 3: Labels assigned to event pairs based on event and TLINK metadata

stance, in the text "Diplomacy is making headway in resolving the UN's standoff with Iraq. One major sticking point has been Iraq's proposal...", *proposal* causes *standoff*, which is a prerequisite for *resolving*. Hence, the pair (proposal, resolving) is considered causal/prerequisite. Our assignment of causal/prerequisite links is unordered. For example, reverse event pairs (wet, spilled), (ate, cooked), and (resolving, proposal) are also considered causal/prerequisite. Link order is taken into consideration while assigning a temporal relation.

If two events contain a causal/prerequisite link, we identify the event in the pair that causes or is a prerequisite for the other. We call this event "A" and the other "B". For example, (spilled, wet) is expressed as (A, B), while (wet, spilled) is expressed as (B, A). To label the TLINK, we determine whether A is a point (P) or interval (I) event using existing date annotations from EventTime (Reimers et al., 2016). This helps us catch cases where A is a long-lasting interval and the time span for B is completely included in A. For instance, in "the war forced civilians to evacuate", (war, evacuate) has a causal/prerequisite link with *war* being event A. Though *war* caused *evacuation*, it is reasonable to expect that the war started *before* and ended *after* evacuation. If A is not present in EventTime (i.e it cannot be assigned a specific time), we use our judgment to determine event length. We then assign a label as per table 3. Unlabeled links are passed to the next step.

4.1.3 Using narrative ordering

This step uses a heuristic based on the intuition that events in a narrative are often presented in chronological order. To label a TLINK, we determine which event appeared first in the document. This event is called "A", and the other is "B". We then detect whether A is a point (P) or interval (I) from EventTime, falling back to our own judgment if it is not present. Finally, a label is assigned following table 3. This step is guaranteed to assign a label since every pair will have a narrative-based order.

Dataset	Kappa
TimeBank	0.71
TimeBank-Dense	0.56-0.64
TDD-Man	0.69

Table 4: Inter-annotator agreement (Cohen's Kappa) on temporal ordering datasets. Kappa scores for TDD-Man are reported on the test set containing 1500 links.

	a	b	s	i	ii
a	137	22	0	12	22
b	30	311	1	72	23
s	0	0	42	5	4
i	9	36	3	462	35
ii	12	32	0	21	209

Table 5: Relation agreement between annotators on the TDD-Man test set containing 1500 links.

4.2 Inter-annotator agreement

Our annotation scheme was developed over multiple rounds of coding and discussion between two experts. In each round, experts separately annotated a set of 10-15 TLINKs.[7] Cohen's Kappa was computed and disagreements were discussed. TLINKs were changed in every round to ensure exposure to diverse event pair types. Inter-annotator agreement in preliminary rounds ranged from 0.48-0.69. The final coding scheme resulted in an agreement of 0.69 on the test set. Table 4 shows that our agreement is comparable to prior work. Table 5 presents a class-wise distribution of agreements between pairs of annotators. Disagreements mainly include cases where one annotator chose after/before while the second chose includes/is included (64%). This indicates that determining precise end-points for an interval event is difficult, as corroborated by Ning et al. (2018b).

5 Automatic Inference

This approach uses automatic inference to derive new TLINKs at low cost from EventTime (Reimers et al., 2016), which assigns specific times to events. EventTime divides events into two types: SingleDay and MultiDay. SingleDay events are assigned dates, while MultiDay events are assigned intervals. Possible event pairs can be divided into three categories: SS (both events are SingleDay), SM (one event is SingleDay while the other is MultiDay) and MM (both events are MultiDay). Not all assigned dates and intervals are exact. EventTime relies heavily on under-specified

[7] chosen from documents in the development set

temporal expressions (such as "after1998-06-08"), making automatic inference non-trivial.

We follow separate algorithms to infer TLINKs for each pair type (SS, SM and MM). For SS pairs, both events are associated with dates, which may be expressed in one of four ways[8], resulting in 16 date combinations for SS links. We develop heuristics[9] for each combination, which generate a temporal relation based on date values. Our heuristics were developed with a focus on precision to avoid adding incorrect links. Often, a relation cannot be generated. For example, consider two events associated with the same date "after02-01-1999". We know that both events occur after 02-01-1999, but we cannot infer their order with respect to *each other*. In such cases, we do not label the pair. For SM pairs, one event is associated with a time interval having begin and end dates. Here we use the SS pair inference algorithm to generate relations between the SingleDay event date and the MultiDay event begin and end dates. These relations are compared to infer the label for the pair. For MM pairs, both events have begin and end dates. We infer relations between begin and end points using SS link inference and use these to infer the pair label. After inference, we perform temporal closure, according to Chambers et al. (2014). To evaluate validity of generated TLINKs, we randomly sample a subset of 100 TLINKs and ask three annotators [10] to determine the correctness of the labels. All annotators unanimously agree with the assigned label in 99% cases. We call this subset **TDD-Auto**.

6 Dataset Statistics

Our data construction pipeline produces the first dataset focused on temporal links between global discourse-level event pairs (**TDDiscourse**), consisting of two subsets **TDD-Man** and **TDD-Auto**. Table 6 presents train, dev and test set sizes for both subsets, Timebank-Dense as well as an augmented version of TimeBank-Dense with additional links inferred via temporal closure. Our complete dataset is 7x larger than both, indicating that our construction adds valuable new TLINKs. **TDD-Man** itself is as large as TimeBank-Dense and can be used in isolation, however incorporating **TDD-Auto** provides a large amount of training data making the task more amenable to deep neural net approaches.

Dataset	Train	Dev	Test
TB-Dense	4032	629	1427
TB-Dense + Closure	4399	722	1575
TDD-Man	4000	650	1500
TDD-Auto	32609	1435	4258

Table 6: Dataset sizes for TimeBank-Dense and our dataset. Note that we only count event-event TLINKs

Table 7 presents class distributions for TDD-Man and TDD-Auto test sets. Though there is a clear majority class, both sets are more balanced than TimeBank-Dense, in which 40% event pairs are labeled "vague". To evaluate the presence of long-distance TLINKs, we present the distribution of distance between event pairs from annotated TLINKs in table 8 which shows that nearly 53% TLINKs in our dataset comprise of event pairs which are more than 5 sentences apart. Further, to gain deeper insight into global discourse-level phenomena exhibited by our dataset, we augment 3 documents from the test set (107 manual and 110 automated event pairs) with additional annotations about phenomena required to label them correctly. We consider the following phenomena:

- **SingleSent (SS):** Textual cues from sentences containing the events suffice to predict the relation (irrespective of distance).
- **Chain Reasoning (CR):** Correct relation prediction requires reasoning about other events from the document.
- **Tense Indicator (TI):** For verb events, tense information indicates the correct relation.
- **Future Events (FE):** One or both events from the pair will occur in the future.
- **Hypothetical/ Negated (HN):** One or both events are hypothetical or negated.
- **Event Coreference (EC):** Event coreference resolution is needed to predict relation.
- **Causal/ Prereq (CP):** Causal/ prerequisite links must be identified to predict relation.
- **World Knowledge (WK):** Real world knowledge is needed to identify the relation.

Table 9 shows the distribution of these phenomena in TDD-Man and TDD-Auto. TDD-Man shows a higher percentage of difficult phenomena (CR, CP). On the other hand, TDD-Auto shows high prevalence of SS, indicating that local information

[8] MM-DD-YYYY, afterMM-DD-YYYY, beforeMM-DD-YYYY, afterMM-DD-YYYYbeforeMM-DD-YYYY (MM-DD-YYYY stands for a specific date value)

[9] Sample heuristics provided in the appendix

[10] Annotators were volunteers with no vested interest in the corpus

Dataset	a	b	s	i	ii
TB-Dense	0.18	0.22	0.02	0.05	0.06
TDD-Man	0.13	0.27	0.03	0.38	0.19
TDD-Auto	0.28	0.32	0.16	0.11	0.13

Table 7: Class distributions for our test sets and TimeBank-Dense. Note that the distribution for TimeBank-Dense does not sum to 1, since it includes a vague class.

Dataset	<5	<10	<15	<20	>20
TDD-Man	0.40	0.40	0.15	0.04	0.01
TDD-Auto	0.50	0.32	0.12	0.05	0.01

Table 8: Distribution of distance between events for all TLINKs in our test sets (in terms of #sentences)

may be sufficient to label many long-distance links in this subset correctly. This principled comparison of both subsets leads us to hypothesize that models which perform well on TimeBank-Dense, should achieve similar scores on TDD-Auto but perform much worse on TDD-Man.

7 Experiments

To statistically evaluate the difficulty of TDDiscourse, we adapt and benchmark three SOTA models on our data. Our results reveal interesting insights about model drawbacks, highlighting the need to shift focus to handling global discourse-level phenomena such as chain reasoning.

7.1 Adapting State-of-the-Art Models for Benchmarking

As most state-of-the-art temporal ordering models are built on datasets containing mainly local TLINKs, they are not well-equipped to handle global TLINKs. Hence, we adapt these models to ensure fair evaluation. We focus on the following:

CAEVO (Chambers et al., 2014): This system consists of specialized learners (sieves) which include heuristic rules and trained models. For each document, sieves run in decreasing order of precision. Decisions made by earlier sieves constrain following ones. This framework integrates transitive reasoning, but decisions made by earlier sieves cannot be overturned, causing error cascades. To extend CAEVO, we increase window sizes and remove the AllVague sieve.[11]

BiLSTM (Cheng and Miyao, 2017): Inspired by Xu et al. (2015), this model uses a BiLSTM classifier. For each pair, dependency paths from source and target events to the sentence root are fed to a BiLSTM. For events in adjacent sentences, source and target event sentences are assumed to be connected to a "common root". We follow the same framework to build a BiLSTM.

SP+ILP (Ning et al., 2017): CAEVO and BiLSTM make separate local decisions for each TLINK, which may result in global inconsistency. For example, for events A, B and C, if A occurs before B and B occurs before C, transitivity implies that A occurs before C. Models classifying each pair independently may assign a different relation to A-C. To correct this, Ning et al. (2017) proposed SP+ILP, which uses a structured perceptron with ILP constraints, explicitly enforcing global consistency. This model was trained on TimeBank-Dense which contains fewer TLINKs per document, making joint learning tractable with loose transitivity constraints. But loose transitivity is an issue for our data with 7x more TLINKs, since the number of constraints increases tremendously. To improve tractability, we define a stricter transitivity constraint. Let E, R and P be sets of events, temporal relations and event pairs respectively($P = \{(e_i, e_j) \in E \times E | e_i, e_j \in E, i \neq j\}$). We define an array of binary indicator variables y, where $y_{<r,i,j>}$ indicates whether the relation r holds between events e_i and e_j. Our objective function is defined as:

$$\arg\min_{y} \sum_{<e_i,e_j>\in P} \sum_{r\in R} -y_{<r,i,j>} \log p_{<r,i,j>} \quad (1)$$

subject to the following constraints:

$$y_{<r,i,j>} \in \{0,1\}, \forall (e_i, e_j) \in P, \forall r \in R \quad (2)$$

$$\sum_{r\in R} y_{<r,i,j>} = 1, \forall (e_i, e_j) \in P \quad (3)$$

Phenomenon	TDDMan	TDDAuto
SS	25.23%	90.91%
CR	58.88%	9.09%
TI	12.10%	46.36%
FE	36.45%	29.09%
HN	14.02%	19.09%
EC	16.82%	4.55%
CP	64.49%	29.09%
WK	16.82%	0.91%

Table 9: Distribution of various phenomena in the annotated test subset. These phenomena were labeled manually.

[11] since our data does not include the vague class. We also remove the WordNet sieve and add MLEventEventDiffSent. For more details on these sieves, we refer interested readers to Chambers et al. (2014)

System	TB-Dense			TDD-Auto			TDD-Man		
	P	R	F1	P	R	F1	P	R	F1
MAJOR	40.5	40.5	40.5	34.2	32.3	33.2	**37.8**	**36.3**	**37.1**
CAEVO	49.9	46.6	48.2	61.1	32.6	42.5	32.3	10.7	16.1
BiLSTM	63.9	38.9	48.4	**55.7**	**48.3**	**51.8**	24.9	23.8	24.3
SP	37.7	37.8	37.7	43.2	43.2	43.2	22.7	22.7	22.7
SP+ILP	**58.4**	**58.4**	**58.4**	46.4	45.9	46.1	23.9	23.8	23.8

Table 10: Performance of SOTA models on TB-Dense, TDD-Auto and TDD-Man. MAJOR represents a majority-class baseline. We report performance on non-vague event-event links for TB-Dense to ensure fair comparison.

$$y_{<r1,i,j>} + y_{<r2,j,k>} - y_{<r3,i,k>} \leq 1, \quad \forall (e_i, e_j), (e_j, e_k), (e_i, e_k) \in P, \forall (r1, r2, r3) \in TC \tag{4}$$

where $p_{<r,i,j>}$ is the probability that event pair (e_i, e_j) has label r. (2) ensures that indicator variables are binary, (3) forces event pairs to be assigned a unique label and (4) imposes transitivity. TC denotes the set of transitive relation triples.[12] Relation probabilities ($p_{<r,i,j>}$) come from the structured perceptron. In addition to this model, we also evaluate the structured perceptron (**SP**) in isolation, which lets us study the effect of introducing global consistency via ILP.

8 Results and Analysis

We benchmark 4 adapted SOTA models (**CAEVO**, **BiLSTM**, **SP** and **SP+ILP**) on **TDD-Auto** and **TDD-Man**. **SP** is a local perceptron-based classifier, while **SP+ILP** introduces transitivity via ILP into the perceptron. This . For tractability, we limit all models to using event pairs which are 15 or fewer sentences apart. This discards only 5% of our data (table 8). Table 10 presents the benchmarking results. We also benchmark models on TimeBank-Dense (**TB-Dense**) to demonstrate that our modifications do not affect performance on local TLINKs.

All models perform better than a majority class baseline on TDD-Auto. The BiLSTM and SP perform particularly well, achieving a higher F1 than TB-Dense, while CAEVO and SP+ILP show slight degradation in comparison to TB-Dense. This corroborates our hypothesis that many long-distance TLINKs in TDD-Auto can be handled with local information. However, all models show a significant drop on TDD-Man, with none outperforming a majority class baseline. Further analysis of model errors offers valuable insights into which phenomena are not handled by models, posing interesting challenges for future work.

Maintaining global consistency: Most SOTA models make separate local decisions for each pair and are not globally consistent. Adding global consistency improves the performance of a local classifier, as evinced by a 3-point F1 gain observed on adding ILP to SP. We validate this observation by performing a transitivity analysis of BiLSTM and SP+ILP on TDD-Auto. We go through all event triples (e_1, e_2, e_3). For each model, if (e_1, e_2), (e_2, e_3) and (e_1, e_3) are all assigned labels, we check whether labels are consistent. For example, e_1 after e_2, e_2 after e_3 and e_1 after e_3 is a consistent assignment. We observe that though the BiLSTM has higher F1, it maintains transitivity in 41.9% cases, while SP+ILP enforces transitivity in 53.6% cases, a 12% increase. We believe that incorporating such constraints into neural models can help, which we delegate to future work.

Incorporating real world knowledge: To examine the dismal performance of all models on TDD-Man, we manually look at 100 pairs on which all models made mistakes. 40% of these cases require real world knowledge. Some examples include determining that "military actions" refers to the same event as "air strikes" (strikes would have to be carried out by the military which cannot be inferred from text), or knowing that certain events (eg: "war") are long-term. No SOTA model currently has this ability.

Using event coreference and structure: Our analysis reveals another source of errors arising from models' inability to handle event coreference and event structure such as sub-events or aspectual predication, a grammatical device which focuses on different facets of event history (eg: using "begin" to indicate initiation) (Pustejovsky et al., 2003). This inability causes models to fail in 22% cases indicating that exploiting rich event structure information is a promising direction.

[12] ("before", "before", "before") form a transitive relation triple as A before B and B before C implies A before C

Dealing with hypothetical or negated events: We observe that SOTA models do not possess the ability to handle these, causing 31% of errors.

9 Conclusion and Future Work

In this work, we created TDDiscourse, the first dataset focused on global discourse-level temporal ordering. Our annotation scheme for TDDiscourse handled several issues which have not been explicitly addressed in prior work. We further adapted and benchmarked 3 SOTA models. All models, on average, performed worse on TDDiscourse, validating the difficulty of the task. Our error analysis reveals key phenomena not handled by current systems, such as hypothetical/negated events, event coreference, aspectual predication, real world knowledge and global consistency. Future work in temporal ordering must address these issues, and we suggest several avenues for exploration, such as a BiLSTM-ILP joint learning framework which has the advantage of combining representational power of neural models with key linguistic insights, and introducing event coreference information via ILP into a structured learning approach similar to Ning et al. (2017). Finally, we hope that our dataset offers a challenging testbed for the development of more global discourse-aware models for temporal ordering.

Acknowledgements

This work was supported by the University of Pittsburgh Medical Center (UPMC) and Abridge AI Inc through the Center for Machine Learning and Health at Carnegie Mellon University. It was also funded in part through NSF IIS 1723454. The authors would like to thank Lisa Carey Lohmueller, Shivani Poddar and Michael Miller Yoder for assisting in human evaluation of TDD-Auto, Evangelia Spiliopoulou for help in implementing parts of the SP+ILP system and the anonymous reviewers for their helpful feedback on this work.

References

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA. Association for Computational Linguistics.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198, Sydney, Australia. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA.

Nathanael Chambers. 2013. Navytime: Event and time ordering from raw text. Technical report, Naval Academy Annapolis MD.

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

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 2012, pages 3735–3740.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional lstm over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

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

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 88–97. Association for Computational Linguistics.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016. On the contribution of word embeddings to temporal relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2818–2828, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

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

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204, Berlin, Germany. Association for Computational Linguistics.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2018. Event time extraction with a decision tree of neural classifiers. *Transactions of the Association for Computational Linguistics*, 6:77–89.

Andrea Setzer. 2002. *Temporal information in newswire articles: an annotation scheme and corpus study.* Ph.D. thesis, University of Sheffield.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Naushad UzZaman and James Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283, Uppsala, Sweden. Association for Computational Linguistics.

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

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, 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, Uppsala, Sweden. Association for Computational Linguistics.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics.

Appendix

A Coreference Examples

- **Coreferent example:** In the example "their disputes have been **bedeviled** by a number of **disputes**", the event "disputes" is itself the entity enacting the event "bedeviled". The events take place over the same time period and location, and are both real events. Thus, we can conclude the events are coreferent.
- **Non-coreferent example:** In "lower rates have **helped** invigorate housing by **making** loans more affordable", though the events share an agent ("lower rates") and realis states, they act on different patient entities and thus are not coreferent.

B Sample heuristic rules from SS link inference procedure:

Assume S1 and S2 indicate the points associated with events 1 and 2 which are to be linked. Following subsections provide a brief sample of some of the heuristic rules we develop to infer the temporal link based on the values of S1 and S2.

B.1 S1 is of type MM-DD-YYYY and S2 is of type afterMM-DD-YYYY

- Get the relation (rel) between the date values from S1 and S2
- If rel is simultaneous or before, the SS link value is before
- Else skip this link

B.2 S1 is of type MM-DD-YYYY and S2 is of type beforeMM-DD-YYYY

- Get the relation (rel) between the date values from S1 and S2
- If rel is simultaneous or after, the SS link value is after
- Else skip this link

B.3 S1 is of type MM-DD-YYYY and S2 is of type afterMM-DD-YYYY beforeMM-DD-YYYY

- From S2, the date associated with after is named date1 and the date associated with before is named date2
- Get the relation (rel1) between date value from S1 and date1 from S2
- If rel1 is simultaneous or before, the SS link value is before
- Get the relation (rel2) between date value from S1 and date2 from S2
- If rel2 is simultaneous or after, the SS link value is after
- Else skip this link

We develop similar rules for the remaining 13 cases. We also develop rule-based inference procedures for SM and MM links. Please refer to the autogeneration code for the complete set of rules.