Modeling Document-level Causal Structures for Event Causal Relation Identification

Lei Gao, Prafulla Kumar Choubey, Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University
(sjtuprog, prafulla.choubey, huangrh)@tamu.edu

Abstract

We aim to comprehensively identify all the event causal relations in a document, both within a sentence and across sentences, which is important for reconstructing pivotal event structures. The challenges we identified are two: 1) event causal relations are sparse among all possible event pairs in a document, in addition, 2) few causal relations are explicitly stated. Both challenges are especially true for identifying causal relations between events across sentences. To address these challenges, we model rich aspects of document-level causal structures for achieving comprehensive causal relation identification. The causal structures include heavy involvements of document-level main events in causal relations as well as several types of fine-grained constraints that capture implications from certain sentential syntactic relations and discourse relations as well as interactions between event causal relations and event coreference relations. Our experimental results show that modeling the global and fine-grained aspects of causal structures using Integer Linear Programming (ILP) greatly improves the performance of causal relation identification, especially in identifying cross-sentence causal relations.

1 Introduction

Understanding causal relations between events in a document is an important step in text understanding and is beneficial to various NLP applications, such as information extraction, question answering and text summarization. Causal relations can occur between any two events in a document, both between events within a sentence and between events across sentences. In this paper, we aim to identify all the event causal relations in a document.

The main challenges for achieving comprehensive causal relation identification are that event causal relations are sparse among all the event pairs in a document and few event causal relations are explicitly stated. The challenges are especially true for identifying cross-sentence event causal relations and most of them have no clear causal indicators. To address these challenges, we model rich aspects of document-level causal structures, i.e., structural distributions of causal relations within a document, for achieving comprehensive causal relation identification in news articles.

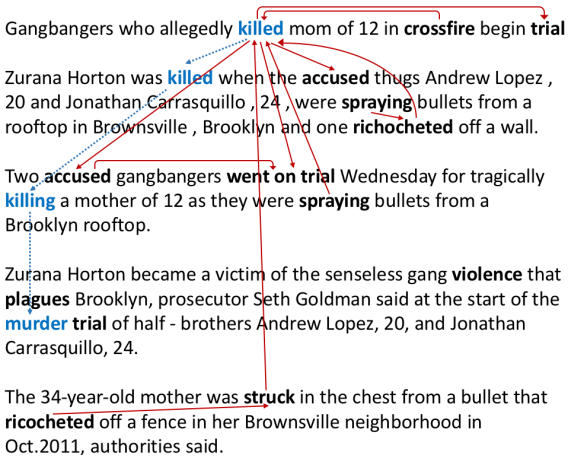

Figure 1: An Example of Main Event Causal Structure

Our key observation for improving causal relation identification is that causal relations, especially cross-sentence causal relations, tend to involve one or two main events of a document. The main events are the focus of a story, which are usually mentioned in the title of an article and have repeated mentions throughout the document. Intuitively, causal relations in a document are often used to explain why the main events happened as well as consequences of the main events. For example, as shown in figure 1, **killing** is the main event. The events **crossfire, spraying, richocheted, struck** are its preconditions, and **accuse, trial** are its consequences. Indeed, many causal

relations are related to the main event.

In addition to the global causal structures related to main events of a document, we model three types of fine-grained causal structures in order to accurately identify each individual causal relation. First, specific sentential syntactic relations may evoke causal relations between event pairs. For instance, adverbial clause modifier of a verb phrase explains its consequence, condition or purpose. Second, we model implications of a discourse relation between two text units (e.g., the *contingency* discourse relation) towards causal relations between events in the two text units. Third, we model interactions between event causal relations and event coreference relations. For example, coreferent event mentions should have the same causal relations; a causal relation and an identity relation should not co-exist between any two events.

We use Integer Linear Programming (ILP) to model these rich causal structures within a document by designing constraints and modifying the objective function to encourage causal relations akin to the observed causal structures and discourage the opposite. Our experimental results on the dataset EventStoryLine (Caselli and Vossen, 2017) show that modeling the global and fine-grained aspects of causal structures within a document greatly improves the performance of causal relation identification, especially in identifying cross-sentence causal relations.

2 Related Work

In the last decade or so, both unsupervised and supervised causal relation identification approaches have been proposed including linguistic patterns, statistical measures and supervised classifiers, primarily with the goal of acquiring event causality knowledge from a text corpus. The proposed approaches mainly rely on explicit contextual patterns (Girju; Hashimoto et al., 2014) or other causality cues (Riaz and Girju, 2010; Do et al., 2011), statistical associations between events (Beamer and Girju, 2009; Hu et al., 2017; Hu and Walker, 2017; Do et al., 2011; Hashimoto et al., 2014), and lexical semantics of events (Riaz and Girju, 2013, 2014b,a; Hashimoto et al., 2014).

An increasing amount of recent works focused on recognizing event causal relations within a document, but mostly limited to identifying intra-sentence causal relations with explicit causal indicators. Mirza et al. (2014) annotated event causal relations in the TempEval-3 corpus and created CausalTimeBank. Mirza and Tonelli (2014) stated that incorporating temporal information improved the performance of a causal relation classifier. Mirza and Tonelli (2016) built both a rule-based multi-sieve approach and a feature based classifier to recognize causal relations in CausalTimeBank. However, causal relations in CausalTimeBank are few and only explicitly stated intra-sentence causal relations were annotated. In addition, Mostafazadeh et al. (2016) annotated both temporal and causal relations in 320 short stories (five sentences in each story) taken from the ROCStories Corpus and indicated strong correlations between causal relations and temporal relations.

Lately, Caselli and Vossen (2017) created a corpus called EventStoryLine, which contains 258 documents and more than 5,000 causal relations. The EventStoryLine corpus is the largest dataset for causal relation identification till now with comprehensive event causal relations annotated, both intra-sentence and cross-sentence, which presents unique challenges for causal relation identification. Caselli and Vossen (2017) showed that only 117 annotated causal relations in this dataset are indicated by explicit causal cue phrases while the others are implicit. We conduct experiments on the EventStoryLine dataset. Distinguished from most of the previous approaches that identify one causal relation each time, we model coarse-grained and fine-grained document-level event causal structures and infer all the causal relations in a document.

Integer linear programming (ILP) approaches have been applied to predict a set of temporal relations or an event timeline in a document (Do et al., 2012; Teng et al., 2016; Ning et al., 2017). ILP has been used to improve causal relation identification (Do et al., 2011), but only with fine-grained constraints considering discourse relations between two text units. Our approach innovates on modeling other aspects of document-level causal structures, especially heavy involvements of main events in causal relations, that facilitate resolving multiple causal relations.

3 The EventStoryLine Corpus

Table 1 shows the statistics of the corpus EventStoryLine v0.9[1] (Caselli and Vossen, 2017).

Item	Size
Topics	22
Documents	258
Sentences	4,316
Event Mentions	5,334
Intra-sentence causal links	1,770
Cross-sentence causal links	3,855
The Total causal links	5,625
Explicit causal links	117

Table 1: EventStoryLine v0.9

Causal relations annotated in EventStoryLine are between two event mentions. Different causal relations are annotated in EventStoryLine, called "rising_action" and "falling_action", which indicate the directions of causal relations and intuitively correspond to "precondition" and "consequence" relations. Note that in this paper, we focus on identifying all the pairs of events in a document that are causally related, but not on classifying the direction of a causal relation though; specifically, we aim to recognize if there exists a causal relation between any two events *A* and *B* in a document, but we do not further distinguish if *A* causes *B* vs. *B* causes *A*.

On average, there is 1.2 event mentions in each sentence. There are 7,805 intra-sentence and 46,521 cross-sentence event mention pairs in total in the corpus, around 22% (1,770) and 8% (3,855) of them were annotated with a causal relation respectively. Out of the annotated causal links, only 117 Caselli and Vossen (2017) causal relations are indicated by explicit causal cue phrases while the others are implicit. In our experiments, we use the gold event mentions in EventStoryLine and exclude aspectual, causative, perception and reporting event mentions[2], most of which were not annotated with any causal relation according to Caselli and Vossen (2017).

4 The Feature Based Local Pairwise Classifiers

Intra- and cross-sentence causal relations are different by nature. For instance, dependency relations between words in a sentence may be more useful for detecting intra-sentence causal relations, than when used for detecting cross-sentence causal relations. Therefore, we train two separate logistic regression classifiers, one for intra-sentence causal link detection and the other for cross-sentence causal link detection.

We consider all event mention pairs within a sentence as training instances for the intra-sentence causal relation classifier. Then we pair event mentions from two sentences with one event mention from each sentence, which are used as training instances for the cross-sentence classifier. Note that training instances for both classifiers are unbalanced, with a POS:NEG ration of around 1:3 and 1:10 for intra- and cross-sentence cases respectively. We applied the "balanced" class weight option[3] in logistic regression classifiers to deal with the class imbalance problem.

We use the same set of features for training both classifiers, but we expect the two classifiers to assign different weights to features.

4.1 The Common Feature Set

Lexical Features: We implement rich lexical features to capture event word forms and similarities between two events, event modifiers and event arguments. First, we encode word and lemma for each token in two event phrases as features. Then we created various similarity features between two events.

- Similarities Based on Event Word Form Match. Three binary features indicating whether the lowercases of two event head words, two event head lemmas and two complete event phrases are exactly the same.
- Similarities Based on Wordnet. We first identify synsets for each event head word in Wordnet. Then for each pair of synsets, with one synset for each event head word, we calculate the Wup similarity (Wu and Palmer, 1994). We create numerical features using the average, minimal and maximal Wup similarities.
- Similarities Based on Word Embeddings. We apply l2 normalization on event head word embeddings, and then we calculate the Euclidean distance and Cosine distance between

[1] Statistics are calculated based on latest release https://github.com/tommasoc80/EventStoryLine

[2] 639 event mentions were excluded in this way.

[3] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

two word embeddings and use them as features. We use Glove Vectors (Pennington et al., 2014) for word embeddings.

- Similarities Based on Event Modifiers. We run the dependency parsing tool from the Stanford CoreNLP (Manning et al., 2014) and identify event modifiers as words that have a certain dependency relation[4] with an event head word. We measure the similarity between two events using the number of common modifiers and the number of common dependency relations that connect a modifier with an event head word.
- Similarities Based on Event Arguments. We consider entities that have a direct dependency relation with an event head word as its event arguments. We use the Stanford CoreNLP to identify entities and their types. We measure the similarity between two events using the number of common event arguments and the number of common entity types.

Causal Potential Features: As inspired by the causal potential metric proposed by (Beamer and Girju, 2009), we encode features based on the point-wise mutual information (PMI) score and the relative textual order between two events. We calculate the PMI score of two event words in EventStoryLine by using co-occurrences of two events in one sentence, and we use the score as a numerical feature.

Syntactic Features: We use dependency relations on the dependency path between two events. We use the basic dependencies extracted from StanfordCoreNLP (Manning et al., 2014). For cross-sentence event pairs, we consider the dependency path from each event to the root node in its own sentence in extracting dependency relations, following Cheng and Miyao (2017). In addition, we use Part Of Speech tags of two event head words as features.

4.2 Score Replacement

We observed that the cross-sentence causal relation classifier is usually not as capable as the intra-sentence classifier, probably due to less contextual evidence to rely on. Therefore, for cross-sentence event mention pairs that can be converted to intra-sentence cases through event coreference links, we use a heuristic method to improve causal relation prediction performance and replace the predictions from the cross-sentence classifier with the predictions from the intra-sentence classifier[5], by using system predicted event coreference links. Note that two events may have more than one pair of mentions, one mention for each event, co-occur within one sentence, we will use the highest score produced by the intra-sentence classifier over all the event mention pairs.

In addition, the score replacement procedure may change prediction scores of some intra-sentence event mention pairs as well. For instance, if one event mention has a coreferent mention within the same sentence that is closer to and is more clearly in a causal relation with the other event mention according to the intra-sentence classifier, and when paired up, the new event pair has received a higher score, then we will replace the score of the original event pair with the higher score. We implemented the within-document neural network based event coreference classifier as described in (Choubey and Huang, 2017a) and used the system to obtain event coreference links.

5 Modeling Causal Structures Using ILP

Our Integer Linear Programming (ILP) system performs document level global inference for resolving all the intra-sentence and inter-sentence event causal relations in a document. Let p_{ij} denotes confidence score from the corresponding local pairwise classifier for assigning a causal relation to the event pair (i, j). Let μ refer to the set of event mentions in a document, we formulate our basic ILP objective function with equation 1.

$$\Theta_{Basic} = \max \sum_{i \in \mu} \sum_{j \in \mu} [p_{ij} x_{ij} + \neg x_{ij}(1 - p_{ij})] \quad (1)$$

We then augment the objective function with new objectives (equation 2) and add constraints to induce causal structures, including heavy involvements of main events (Θ_M and Θ_F) in causal relations throughout the document, as well as fine-grained interactions of event causal relations with discourse relations (Θ_D), and event coreference

[4] Specifically, we consider 'nmod', 'amod', 'advmod', 'mark', 'aux', 'auxpass', 'expl', 'cc', 'cop', 'punct' to be modifiers.

[5] Note that we only conduct the score replacement when a score produced by the intra-sentence classifier is higher than the score produced by the cross-sentence classifier, which indicates that the intra-sentence classifier is more confident.

relations(Θ_C) as well as syntactic structure constraints (Θ_S) for identifying causal relations.

$$\Theta = \Theta_{Basic} + \Theta_M + \Theta_F + \Theta_D + \Theta_C + \Theta_S \quad (2)$$

5.1 Document Level Main Event Based Constraints

Main Event: Main events are central to the story in a document and tend to participate in multiple causal links. Similar to Choubey et al. (2018), we recognize main events based on characteristics of event coreference chains within a document. Specifically, we rank events based on the number of event mentions referring to an event, and choose the top two events as main events[6]. Then we add a new objective function (equation 3) and additional constraints to encourage causal links in event mention pairs containing a main event (equation 4) and discourage causal links in the remaining mention pairs (equation 5).

$$\Theta_M = \max \Big[\sum_{i \in \Lambda} [k_{m_1} m_1(i) + k_{m_2} m_2(i)] - \sum_{i \in \mu - \Lambda} [k_{n_1} n_1(i) + k_{n_2} n_2(i)] \Big] \quad (3)$$

$$\begin{aligned} \forall i \in \Lambda, \sum_{j \in \mu, d_i = d_j} x_{ij} \geq m_1(i) \\ \forall i \in \Lambda, \sum_{j \in \mu, d_i \neq d_j} x_{ij} \geq m_2(i) \end{aligned} \quad (4)$$

$$\begin{aligned} \forall i \in \mu - \Lambda, \sum_{j \in \mu - \Lambda, d_i = d_j} x_{ij} \leq n_1(i) \\ \forall i \in \mu - \Lambda, \sum_{j \in \mu - \Lambda, d_i \neq d_j} x_{ij} \leq n_2(i) \end{aligned} \quad (5)$$

In the above equations, Λ denotes the set of main event mentions, and d_i denotes the sentence number for event i. The independent variables $m_1(i)$ and $m_2(i)$ indicate the minimum number of intra- and cross-sentence causal relations that main events participate in. By maximizing $m_1(i)$ and $m_2(i)$ in the objective function Θ_M, our model favors main events to have more causal relations. Similarly, variables $n_1(i)$ and $n_2(i)$ in equation 5 are separately defined to set upper thresholds on the maximum number of intra- and cross-sentence causal relations without a main event. Unlike $m_1(i)$ and $m_2(i)$, we aim to minimize the variables $n_1(i)$ and $n_2(i)$ to restrict non-main events from participating in causal relations. Notice that we apply the constraints separately to intra- and cross-sentence mention pairs. This is primarily because main events are likely to participate in many more cross-sentence causal relations compared to intra-sentence cases. Furthermore, we observe that a main event may trigger several consequent events which themselves are causally related. However, causal relations involving only non-main events are less likely to show transitivity. Therefore, we add the constraint 6 to ensure non-transitivity among causal relations with no main event.

$$x_{ij} + x_{jk} + x_{ik} \leq 2 + 1_{i \in \Lambda} + 1_{j \in \Lambda} + 1_{k \in \Lambda} \quad (6)$$

Locality Constraints: Main events may not always have the largest coreference chain size, and the position of an event mention provides another strong heuristics for identifying the main event (Upadhyay and Roth, 2016). In addition, the first sentence often summarizes the central context of story and are likely to describe foreground events (Grimes, 1975) that have causal relation with multiple other events. Therefore, we add an objective function (equation 7) and additional constraints (equation 8) to encourage causal relations that contain an event from the first sentence.

$$\Theta_S = \max \sum_{i \in S} k_f b_1(i) - \sum_{i \in \mu} \sum_{j \in \mu} k_f l_{ij} \cdot |d_i - d_j| \quad (7)$$

$$\forall i \in F, \sum_{j \in \mu} x_{ij} \geq b_1(i) \quad (8a)$$

$$\forall < i, j > \in M, x_{ij} \leq l_{ij} + 1_{i \notin \{F\}} \wedge 1_{j \notin \{F\}} \quad (8b)$$

where, F represents all the events in first sentence, independent variable $b_1(i)$ indicates the minimum number of causal relations that an event in F participates in, M represents the set of event mention pairs that can be mapped to the same sentence and l_{ij} is a leakage variable that allows distant event mentions in F receiving a very high confidence value p_{ij} to have a causal relation. Particularly, we encourage causal links between two event mentions that are in nearby sentences or can

[6] If there is a tie between two event clusters with the same number of coreferential event mentions, we use the sum of confidence scores for pairs of coreferential event mentions in a cluster to break the tie. The confidence scores were assigned by the local pairwise coreference relation classifier.

be mapped to the same sentence using coreference links[7]. By maximizing the variables $b_1(i)$ and minimizing the term $l_{ij} \cdot |d_i - d_j|$, we encourage event mentions in F complying with certain constraints to have more causal relations.

5.2 Fine-grained Causal Structure Constraints

Syntactic Relations: Specific sentential syntactic relations may evoke causal relations between event pairs. First, adverbial clause modifier of a verb phrase explains its consequence, condition or purpose; Second, nominal events mentioned as subject in the main clause presents an assertional structure that delivers foreground (Grimes, 1975) information which may have causal associations with other events; Third, non-finite verb events that share arguments and complement the main event of a sentence are likely to have causal associations with the main event.

Therefore, we add an objective function (equation 9) and additional constraints (equation 10) to encourage causal relations that contain a nominal event as subject or verb event that modifies its parent with advcl or xcomp dependency relations. Here, S represents event mentions that possess one of the above syntactic structures, independent variable $b_2(i)$ indicates the minimum number of causal relations that an event in S participates in. Note that equation 10(b) was modified from 8(b) and allows discounted optimization (with l_{ij}) for events in S that are mappable to the same or nearby sentences.

$$\Theta_S = \max \sum_{i \in S} k_s b_2(i) \tag{9}$$

$$\forall i \in S, \sum_{j \in \mu} x_{ij} \geq b_2(i) \tag{10a}$$

$$\forall < i, j > \in M, x_{ij} \leq l_{ij} + 1_{i \notin \{F,S\}} \wedge 1_{j \notin \{F,S\}} \tag{10b}$$

Discourse Relations: Note that the implications of discourse relations between two text units towards causal relations between events in the two text units have been discussed in the previous work (Do et al., 2011). In this work, we consider three types of discourse relations[8]. First, two subtypes of the *contingency* discourse relation, namely *cause* and *condition*, strongly suggest that causal links exist between events in the two discourse units. On the contrary, the *comparison* discourse relation highlights semantic independence between two discourse units, thus inhibits causal relations between events described in them. Third, all causal relations are inherently temporal. An event that causes another event must necessarily occur before or temporally overlap with the latter. Thus, clauses having one of these temporal discourse relations may also favor causal relations between events in them. We model the above three dependencies between discourse relations and causation through constraints 11 and the objective function 12.

$$\begin{aligned} &\forall r = Contingency, \sum_{i \in arg_1} \sum_{j \in arg_2} x_{ij} \geq 1 \\ &\forall r = Comparison, \sum_{i \in arg_1} \sum_{j \in arg_2} x_{ij} \leq 0 \\ &\forall r = Temporal, \sum_{i \in arg_1} \sum_{j \in arg_2} x_{ij} \geq T(r) \end{aligned} \tag{11}$$

$$\Theta_D = \max \sum_{r = Temporal} k_t T(r) \tag{12}$$

Specifically, we enforce events in clauses with the contingency discourse relation to have at least one event pair with causal relation. Similarly, we inhibit a causal relation between any event pair in clauses with the comparison discourse relation. For events in clauses with a temporal discourse relation, we aim to maximize the number of causal relations without grounding it to any hard lower bound. Here, r denotes the discourse relation between two discourse arguments, arg_1 and arg_2, and *Temporal* refers to the set of temporal discourse relations. We use the pre-trained PDTB discourse parser (Lin et al., 2014) to obtain discourse relations in a document.

Event Coreference Relations: We model interactions between event causal relations and event coreference relations by adding constraints 13 and 14 and an objective function 15.

$$\forall i \equiv j, x_{ij} \leq c_3(i, j) \tag{13}$$

$$\begin{aligned} &\forall i \equiv j, x_{ik} + \neg x_{jk} \leq 1 + c_1(i, j, k) \\ &\forall i \equiv j, \neg x_{ik} + x_{jk} \leq 1 + c_2(i, j, k) \end{aligned} \tag{14}$$

[7] Two event mentions are mappable if their respective coreferential event mentions co-occur in at least one sentence.

[8] We use PDTB parser (Lin et al., 2014) to identify three discourse relations.

$$\Theta_C = \max \sum_{i \in \mu} \sum_{j \in \mu} [\sum_{k \in \mu} -k_c(c_1(i,j,k) + c_2(i,j,k))] - (1-k_c)(c_3(i,j)) \quad (15)$$

Here $\equiv$ represents the identity (coreference) relation. The constraint 13 ensures that causal relation and coreference relation are mutually exclusive, allowing some violations when $p_{i,j}$ is high. The constraints 14 along with the objective function 15 encourage coreferent event mentions to have a causal relation with the same other event. While this relation between causal and coreference relations is strictly true for gold standard data, we observed that these constraints make the system very sensitive to noise when using system predicted coreference links. Therefore, we added binary leakage variables $c_1(i,j,k)$, $c_2(i,j,k)$ and $c_3(i,j)$ to relax these constraints.By maximizing the negative of leakage variables, we allow our model to overcome this instability.

6 Evaluation

6.1 Experimental settings

There are 22 topics in the EventStoryLine corpus. We put them in order based on their topic IDs and use documents in the last two topics as the development set. We trained the ILP system using the rest 20 topics and tuned parameters based on the system performance on the development set. We report experimental results by conducting 5-fold cross validation on the rest 20 topics. For event causal relation identification, we report precision, recall, and F1-score.

The weighting parameters for constraints, including $k_{m_1}, k_{m_2}, k_{n_1}, k_{n_2}, k_f, k_t, k_c$ and k_s, were first pre-set to be a small number 0.1. We then conducted grid search and searched for the best value for each parameter over the range from 0.1 to 0.5 with a step size of 0.1. The best values for the parameters are $0.2, 0.1, 0.1, 0.5, 0.2, 0.1, 0.1, 0.2$ respectively.

6.2 Baseline Systems

We consider six baseline systems:

OP: a dummy model used in (Caselli and Vossen, 2017) that assigns a causal relation to every event mention pair.

Cheng and Miyao (2017): a dependency path based sequential neural network model that extensively models compositional meanings of the context between two event mentions for causal relation identification. This model was first used for identifying event temporal relations and has been shown effective in identifying both intra- and cross-sentence temporal relations.

Choubey and Huang (2017b): another dependency path based sequential neural network model that was first developed for identifying temporal relations between event mentions within a sentence. We make this model also work for cross sentence cases by merging the root nodes of two dependency trees associated with two separate sentences and extracting a dependency path connecting events across sentences.

So far, there is no well recognized effective approach for causal relation identification within a document. We applied the above two models for causal relation identification considering that causal relations are closely related with certain temporal relations and a causal event must occur before or overlap with the consequence event.

LR (Lexical): the same logistic regression classifier as our local pairwise classifier but using the lexical features only.

LR (Causal Potential): the same logistic regression classifier as our local pairwise classifier but using the causal potential features only.

LR (Full): our local pairwise classifier using the full set of features.

+ Score Replacement: our local pairwise classifier using the full set of features, with the heuristic score replacement procedure applied.

6.3 Experimental Results

The first section of table 2 shows the performance of baseline models on intra- and cross-sentence causal relation identification. The model **OP** labels each event mention pair as causal and suffers from low precisions[9], especially on identifying cross-sentence causal relations. The two dependency path based neural network model **(Cheng and Miyao, 2017; Choubey and Huang, 2017b)** do not perform effectively on identifying causal relations. The performance is especially poor on cross-sentence cases.

The model **LR (Lexical)** improved the precision of causal relation identification but suf-

[9]The reason it did not achieve the 100% recall is that we did not consider reporting, causative, perception or aspectual events.

	Intra-sentence			Cross-sentence			Intra + Cross		
Models	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Local Pairwise Models									
OP	22.5	**98.6**	36.6	8.4	**99.5**	15.6	10.5	**99.2**	19.0
Cheng and Miyao (2017)	34.0	41.5	37.4	13.5	30.3	18.7	17.6	33.9	23.2
Choubey and Huang (2017b)	32.7	44.9	37.8	11.3	29.5	16.4	15.5	34.3	21.4
LR (Lexical)	**38.7**	37.0	37.8	24.3	29.1	26.5	**28.2**	31.6	29.8
LR (Causal Potential)	28.2	61.2	38.6	10.7	74.6	18.7	12.9	70.4	21.8
LR (full)	37.6	41.4	39.4	23.8	33.6	27.9	27.4	36.1	31.2
+Score Replacement	37.0	45.2	**40.7**	**25.2**	48.1	**33.1**	27.9	47.2	**35.1**
Modeling Causal Structure using ILP									
+Main Event Constraints	38.1	47.6	42.3	31.5	45.4	37.2	33.4	46.1	38.7
+Locality Constraints	38.0	50.4	43.4	32.1	45.8	37.8	33.9	47.3	39.5
+Syntactic Constraints	37.2	54.8	44.3	32.1	48.6	38.7	33.7	50.6	40.4
+Discourse Constraints	37.4	**55.8**	**44.7**	32.2	**48.7**	38.8	33.8	**51.0**	40.6
+Coreference Constraints	**38.8**	52.4	44.6	**35.1**	48.2	**40.6**	**36.2**	49.5	**41.9**

Table 2: Performance of different models on causal relation identification

fers from low recall. In contrast, the model **LR (Causal Potential)** improved the recall but suffers from low precision. The model **LR (full)** with rich lexical, semantic and syntactic features achieved the best trade-off between precision and recall. **+ Score Replacement** significantly improves the recall and F1-score on identifying cross-sentence causal relations, which also slightly improves the recall of intra-sentence cases. But the precision of causal relation identification remains low, especially on cross-sentence cases.

The second section of table 2 shows the performance of our ILP model after gradually adding each type of constraints. **+Main Event Constraints** shows the performance of the ILP system with constraints encouraging causal relations involving a main event. By modeling this aspect of document-level causal structures, the precision of cross-sentence causal relation identification was clearly improved by around 6.3%. With a small loss on recall, the F1-score was improved by 4.1%. Modeling this document-level causal structure also improves both precision and recall on identifying intra-sentence causal relations, but with a relatively small margin. Compared to the local pairwise model **+ Score Replacement**, the overall F1-score improvement from using global main event constraints is statistically significant with $p<0.05$ (Dietterich, 1998). **+Locality Constraints** strengthens the effects of modeling main events and further improved the performance of both cross- and intra-sentence causal relation identification.

Next, adding sentential syntactic structure based constraints (**+Syntactic Constraints**) recovered additional intra-sentence causal relations and cross-sentence causal relations as well due to score replacement, and improved their recall by 4.4% and 2.8% respectively with little or no drop on precision. Then, after adding discourse constraints (**+Discourse Constraints**), both precision and recall on intra-sentence causal relation identification were slightly improved while the performance on cross-sentence causal relation identification remains roughly the same, this is mainly due to the fact that few cross-sentence discourse relations were identified by the discourse parser we used. Finally, after adding conference constraints (**+Coreference Constraints**), the precision of cross-sentence causal relation identification was increased by 2.9%, with a small loss on recall, the F1-score was improved by 1.8%. Unsurprisingly, the overall performance on intra-sentence causal relation identification was not affected much by coreference constraints since event coreference relations often involve events across sentences. Compared to the model considering global constraints only (the line **+ Locality Constraints**), the overall F1-score improvement from using fine-grained causal structure constraints is statistically significant with $p<0.01$.

To sum up, by modeling the global and fine-grained aspects of causal structures, the performance of both intra- and cross-sentence causal re-

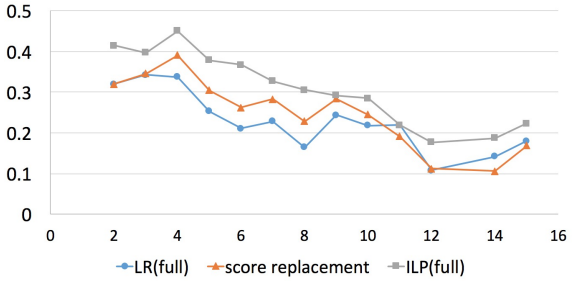

Figure 2: F1-scores on documents with different lengths. The x-axis indicates the number of sentences a document has. The y-axis indicates the macro average F1-score of causal relation identification.

lation identification was greatly improved by 3.9% and 7.5% in F1-score respectively. Compared to the local pairwise model **+ Score Replacement**, the overall F1-score improvement from using both global main event constraints and fine-grained causal structure constraints is statistically significant with $p<0.002$.

Impact of Document Lengths Figure 2 shows performance comparisons of three models on documents with different lengths. The first impression is that causal relation identification becomes harder when documents are longer. If we look into the figure, the score replacement heuristic improves the performance of causal relation identification on medium-sized documents, but not on short ($<$ 4 sentences) or long ($>$ 10 sentences) documents. This may either due to little event coreference information for use in short documents or event coreference information becoming too noisy in long documents. Compared to the mixed effects of the score replacement heuristic, the ILP system improved the performance of causal relation identification consistently in documents of any length, through modeling rich document-level causal structures.

7 Conclusions

We have presented an ILP system that collectively identifies all the causal relations within a document, both intra- and cross-sentence causal relations, by modeling the global and fine-grained aspects of causal structures. In the future, we will continue to enrich document-level causal structures, e.g., by considering segment-wise topic layouts and rhetorical discourse structures.

Acknowledgments

This work was partially supported by the National Science Foundation via NSF Award IIS-1755943. Disclaimer: the views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

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)*, volume 2, pages 1–6.

Prafulla Kumar Choubey and Ruihong Huang. 2017a. Event coreference resolution by iteratively unfolding inter-dependencies among events. *arXiv preprint arXiv:1707.07344*.

Prafulla Kumar Choubey and Ruihong Huang. 2017b. A sequential model for classifying temporal relations between intra-sentence events. *arXiv preprint arXiv:1707.07343*.

Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. 2018. Identifying the most dominant event in a news article by mining event coreference relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 340–345.

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

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Quang Xuan 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. Association for Computational Linguistics.

Roxana Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83.

Joseph Evans Grimes. 1975. *The thread of discourse*, volume 207. Walter de Gruyter.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 987–997.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn A Walker. 2017. Inference of fine-grained event causality from blogs and films. *arXiv preprint arXiv:1708.09453*.

Zhichao Hu and Marilyn A Walker. 2017. Inferring narrative causality between event pairs in films. *arXiv preprint arXiv:1708.09496*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th International Conference on Computational Linguistics*, pages 64–75. ACL.

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

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 361–368. IEEE.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *SIGDIAL Conference*, pages 21–30.

Mehwish Riaz and Roxana Girju. 2014a. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 161–170.

Mehwish Riaz and Roxana Girju. 2014b. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57.

Jiayue Teng, Peifeng Li, Qiaoming Zhu, and Weiyi Ge. 2016. Joint event co-reference resolution and temporal relation identification. In *Workshop on Chinese Lexical Semantics*, pages 426–433. Springer.

Christos Upadhyay, Shyam @articledietterich1998approximate, title=Approximate statistical tests for comparing supervised classification learning algorithms, author=Dietterich, Thomas G, journal=Neural computation, volume=10, number=7, pages=1895–1923, year=1998, publisher=MIT Press and Christodoulopoulos and Dan Roth. 2016. "making the news": Identifying noteworthy events in news articles. In *Proceedings of the Fourth Workshop on Events*, pages 1–7.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.