Minimally Supervised Event Causality Identification

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

Abstract

This paper develops a minimally supervised approach, based on focused distributional similarity methods and discourse connectives, for identifying of causality relations between events in context. While it has been shown that distributional similarity can help identifying causality, we observe that discourse connectives and the particular discourse relation they evoke in context provide additional information towards determining causality between events. We show that combining discourse relation predictions and distributional similarity methods in a global inference procedure provides additional improvements towards determining event causality.

1 Introduction

An important part of text understanding arises from understanding the semantics of events described in the narrative, such as identifying the events that are mentioned and how they are related semantically. For instance, when given a sentence "The police arrested him because he killed someone.", humans understand that there are two events, triggered by the words "arrested" and "killed", and that there is a causality relationship between these two events. Besides being an important component of discourse understanding, automatically identifying causal relations between events is important for various natural language processing (NLP) applications such as question answering, etc. In this work, we automatically detect and extract causal relations between events in text.

Despite its importance, prior work on event causality extraction in context in the NLP literature is relatively sparse. In (Girju, 2003), the author used noun-verb-noun lexico-syntactic patterns to learn that "mosquitoes cause malaria", where the *cause* and *effect* mentions are nominals and not necessarily event evoking words. In (Sun et al., 2007), the authors focused on detecting causality between search query pairs in temporal query logs. (Beamer and Girju, 2009) tried to detect causal relations between verbs in a corpus of screen plays, but limited themselves to consecutive, or adjacent verb pairs. In (Riaz and Girju, 2010), the authors first cluster sentences into topic-specific scenarios, and then focus on building a dataset of causal text spans, where each span is headed by a verb. Thus, their focus was not on identifying causal relations between events in a given text document.

In this paper, given a text document, we first identify events and their associated arguments. We then identify causality or relatedness relations between event pairs. To do this, we develop a minimally supervised approach using focused distributional similarity methods, such as co-occurrence counts of events collected automatically from an unannotated corpus, to measure and predict existence of causality relations between event pairs. Then, we build on the observation that discourse connectives and the particular discourse relation they evoke in context provide additional information towards determining causality between events. For instance, in the example sentence provided at the beginning of this section, the words "arrested" and "killed" probably have a relatively high apriori likelihood of being ca-

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 294–303,
Edinburgh, Scotland, UK, July 27–31, 2011.

sually related. However, knowing that the connective "because" evokes a contingency discourse relation between the text spans "The police arrested him" and "he killed someone" provides further evidence towards predicting causality. The contributions of this paper are summarized below:

- Our focus is on identifying causality between event pairs in context. Since events are often triggered by either verbs (e.g. "attack") or nouns (e.g. "explosion"), we allow for detection of causality between verb-verb, verb-noun, and noun-noun triggered event pairs. To the best of our knowledge, this formulation of the task is novel.
- We developed a minimally supervised approach for the task using focused distributional similarity methods that are automatically collected from an unannotated corpus. We show that our approach achieves better performance than two approaches: one based on a frequently used metric that measures association, and another based on the effect-control-dependency (ECD) metric described in a prior work (Riaz and Girju, 2010).
- We leverage on the interactions between event causality prediction and discourse relations prediction. We combine these knowledge sources through a global inference procedure, which we formalize via an Integer Linear Programming (ILP) framework as a constraint optimization problem (Roth and Yih, 2004). This allows us to easily define appropriate constraints to ensure that the causality and discourse predictions are coherent with each other, thereby improving the performance of causality identification.

2 Event Causality

In this work, we define an event as an action or occurrence that happens with associated participants or arguments. Formally, we define an event e as: $p(a_1, a_2, \ldots, a_n)$, where the predicate p is the word that triggers the presence of e in text, and $a_1, a_2, \ldots, a_n$ are the arguments associated with e. Examples of predicates could be *verbs* such as "attacked", "employs", *nouns* such as "explosion", "protest", etc., and examples of the arguments of "attacked" could be its *subject* and *object* nouns.

To measure the causality association between a pair of events e_i and e_j (in general, e_i and e_j could be extracted from the same or different documents), we should use information gathered about their predicates and arguments. A simple approach would be to directly calculate the pointwise mutual information (PMI)[1] between $p^i(a_1^i, a_2^i, \ldots, a_n^i)$ and $p^j(a_1^j, a_2^j, \ldots, a_m^j)$. However, this leads to very sparse counts as the predicate p^i with its list of arguments $a_1^i, \ldots, a_n^i$ would rarely co-occur (within some reasonable context distance) with predicate p^j and its entire list of arguments $a_1^j, \ldots, a_m^j$. Hence, in this work, we measure causality association using three separate components and focused distributional similarity methods collected about event pairs as described in the rest of this section.

2.1 Cause-Effect Association

We measure the causality or cause-effect association (CEA) between two events e_i and e_j using the following equation:

$$CEA(e_i, e_j) = s_{pp}(e_i, e_j) + s_{pa}(e_i, e_j) + s_{aa}(e_i, e_j) \quad (1)$$

where s_{pp} measures the association between event predicates, s_{pa} measures the association between the predicate of an event and the arguments of the other event, and s_{aa} measures the association between event arguments. In our work, we regard each event e as being triggered and rooted at a predicate p.

2.1.1 Predicate-Predicate Association

We define s_{pp} as follows:

$$s_{pp}(e_i, e_j) = PMI(p^i, p^j) \times \max(u^i, u^j) \times IDF(p^i, p^j) \times Dist(p^i, p^j) \quad (2)$$

which takes into account the PMI between predicates p^i and p^j of events e_i and e_j respectively, as well as various other pieces of information. In Suppes' *Probabilistic theory of Casuality* (Suppes, 1970), he highlighted that event e is a possible cause of event e', if e' happens more frequently with e than

[1] PMI is frequently used to measure association between variables.

by itself, i.e. $P(e'|e) > P(e')$. This can be easily rewritten as $\frac{P(e,e')}{P(e)P(e')} > 1$, similar to the definition of PMI:

$$PMI(e, e') = log\frac{P(e, e')}{P(e)P(e')}$$

which is only positive when $\frac{P(e,e')}{P(e)P(e')} > 1$.

Next, we build on the intuition that event predicates appearing in a large number of documents are probably not important or discriminative. Thus, we penalize these predicates when calculating s_{pp} by adopting the inverse document frequency (idf):

$$IDF(p^i, p^j) = idf(p^i) \times idf(p^j) \times idf(p^i, p^j),$$

where $idf(p) = log\frac{D}{1+N}$, D is the total number of documents in the collection and N is the number of documents that p occurs in.

We also award event pairs that are closer together, while penalizing event pairs that are further apart in texts, by incorporating the distance measure of Leacock and Chodorow (1998), which was originally used to measure similarity between concepts:

$$Dist(p^i, p^j) = -log\frac{|sent(p^i) - sent(p^j)| + 1}{2 \times ws},$$

where $sent(p)$ gives the sentence number (index) in which p occurs and ws indicates the window-size (of sentences) used. If p^i and p^j are drawn from the same sentence, the numerator of the above fraction will return 1. In our work, we set ws to 3 and thus, if p^i occurs in sentence k, the furthest sentence that p^j will be drawn from, is sentence $k + 2$.

The final component of Equation 2, $\max(u^i, u^j)$, takes into account whether predicates (events) p^i and p^j appear most frequently with each other. u^i and u^j are defined as follows:

$$u^i = \frac{P(p^i, p^j)}{\max_k[P(p^i, p^k)] - P(p^i, p^j) + \epsilon}$$
$$u^j = \frac{P(p^i, p^j)}{\max_k[P(p^k, p^j)] - P(p^i, p^j) + \epsilon},$$

where we set $\epsilon = 0.01$ to avoid zeros in the denominators. u^i will be maximized if there is no other predicate p^k having a higher co-occurrence probability with p^i, i.e. $p^k = p^j$. u^j is treated similarly.

2.1.2 Predicate-Argument and Argument-Argument Association

We define s_{pa} as follows:

$$s_{pa}(e_i, e_j) = \frac{1}{|A_{e_j}|}\sum_{a \in A_{e_j}} PMI(p^i, a) + \frac{1}{|A_{e_i}|}\sum_{a \in A_{e_i}} PMI(p^j, a), \quad (3)$$

where A_{e_i} and A_{e_j} are the sets of arguments of e_i and e_j respectively.

Finally, we define s_{aa} as follows:

$$s_{aa}(e_i, e_j) = \frac{1}{|A_{e_i}||A_{e_j}|}\sum_{a \in A_{e_i}}\sum_{a' \in A_{e_j}} PMI(a, a') \quad (4)$$

Together, s_{pa} and s_{aa} provide additional contexts and robustness (in addition to s_{pp}) for measuring the cause-effect association between events e_i and e_j.

Our formulation of CEA is inspired by the ECD metric defined in (Riaz and Girju, 2010):

$$ECD(a, b) = \max(v, w) \times -log\frac{dis(a, b)}{2 \times maxDistance}, \quad (5)$$

where

$$v = \frac{P(a, b)}{P(b) - P(a, b) + \epsilon} \times \frac{P(a, b)}{\max_t[P(a, b_t)] - P(a, b) + \epsilon}$$
$$w = \frac{P(a, b)}{P(a) - P(a, b) + \epsilon} \times \frac{P(a, b)}{\max_t[P(a_t, b)] - P(a, b) + \epsilon},$$

where ECD(a,b) measures the causality between two events a and b (headed by verbs), and the second component in the ECD equation is similar to $Dist(p^i, p^j)$. In our experiments, we will evaluate the performance of ECD against our proposed approach.

So far, our definitions in this section are generic and allow for any list of event argument types. In this work, we focus on two argument types: agent (subject) and patient (object), which are typical core arguments of any event. We describe how we extract event predicates and their associated arguments in the section below.

3 Verbal and Nominal Predicates

We consider that events are not only triggered by verbs but also by nouns. For a verb (verbal predicate), we extract its subject and object from its associated dependency parse. On the other hand, since

events are also frequently triggered by nominal predicates, it is important to identify an appropriate list of event triggering nouns. In our work, we gathered such a list using the following approach:

- We first gather a list of deverbal nouns from the set of most frequently occurring (in the Gigaword corpus) 3,000 verbal predicate types. For each verb type v, we go through all its WordNet[2] senses and gather all its derivationally related nouns $\mathcal{N}_v$ [3].

- From $\mathcal{N}_v$, we heuristically remove nouns that are less than three characters in length. We also remove nouns whose first three characters are different from the first three characters of v. For each of the remaining nouns in $\mathcal{N}_v$, we measured its Levenstein (edit) distance from v and keep the noun(s) with the minimum distance. When multiple nouns have the same minimum distance from v, we keep all of them.

- To further prune the list of nouns, we next removed all nouns ending in "er", "or", or "ee", as these nouns typically refer to a person, e.g. "writer", "doctor", "employee". We also remove nouns that are not hyponyms (children) of the first WordNet sense of the noun "event"[4].

- Since we are concerned with nouns denoting events, FrameNet (Ruppenhofer et al., 2010) (FN) is a good resource for mining such nouns. FN consists of frames denoting situations and events. As part of the FN resource, each FN frame consists of a list of lexical units (mainly verbs and nouns) representing the semantics of the frame. Various frame-to-frame relations are also defined (in particular the *inheritance* relation). Hence, we gathered all the children frames of the FN frame "Event". From these children frames, we then gathered all their noun lexical units (words) and add them to our list of nouns. Finally, we also add a few nouns denoting natural disaster from Wikipedia[5].

Using the above approach, we gathered a list of about 2,000 noun types. This current approach is heuristics based which we intend to improve in the future, and any such improvements should subsequently improve the performance of our causality identification approach.

Event triggering deverbal nouns could have associated arguments (for instance, acting as subject, object of the deverbal noun). To extract these arguments, we followed the approach of (Gurevich et al., 2008). Briefly, the approach uses linguistic patterns to extract subjects and objects for deverbal nouns, using information from dependency parses. For more details, we refer the reader to (Gurevich et al., 2008).

4 Discourse and Causality

Discourse connectives are important for relating different text spans, helping us to understand a piece of text in relation to its context:

[The police arrested him] <u>because</u> [he killed someone].

In the example sentence above, the discourse connective ("because") and the discourse relation it evokes (in this case, the *Cause* relation) allows readers to relate its two associated text spans, "The police arrested him" and "he killed someone". Also, notice that the verbs "arrested" and "killed", which *cross* the two text spans, are causally related. To aid in extracting causal relations, we leverage on the identification of discourse relations to provide additional contextual information.

To identify discourse relations, we use the Penn Discourse Treebank (PDTB) (Prasad et al., 2007), which contains annotations of discourse relations in context. The annotations are done over the Wall Street Journal corpus and the PDTB adopts a predicate-argument view of discourse relations. A discourse connective (e.g. because) takes two text spans as its arguments. In the rest of this section, we briefly describe the discourse relations in PDTB and highlight how we might leverage them to aid in determining event causality.

[2] http://wordnet.princeton.edu/

[3] The WordNet resource provides derivational information on words that are in different syntactic (i.e. part-of-speech) categories, but having the same root (lemma) form and that are semantically related.

[4] The first WordNet sense of the noun "event" has the meaning: "something that happens at a given place and time"

[5] http://en.wikipedia.org/wiki/Natural_disaster

Coarse-grained relations	Fine-grained relations
Comparison	Concession, Contrast, Pragmatic-concession, Pragmatic-contrast
Contingency	Cause, Condition, Pragmatic-cause, Pragmatic-condition
Expansion	Alternative, Conjunction, Exception, Instantiation, List, Restatement
Temporal	Asynchronous, Synchronous

Table 1: Coarse-grained and fine-grained discourse relations.

4.1 Discourse Relations

PDTB contains annotations for four coarse-grained discourse relation types, as shown in the left column of Table 1. Each of these are further refined into several fine-grained discourse relations, as shown in the right column of the table.[6] Next, we briefly describe these relations, highlighting those that could potentially help to determine event causality.

Comparison A *Comparison* discourse relation between two text spans highlights prominent differences between the situations described in the text spans. An example sentence is:

Contrast: [According to the survey, *x*% of Chinese Internet users prefer Google] whereas [*y*% prefer Baidu].

According to the PDTB annotation manual (Prasad et al., 2007), the truth of both spans is independent of the established discourse relation. This means that the text spans are not causally related and thus, the existence of a *Comparison* relation should imply that there is no causality relation across the two text spans.

Contingency A *Contingency* relation between two text spans indicates that the situation described in one text span causally influences the situation in the other. An example sentence is:

Cause: [The first priority is search and rescue] because [many people are trapped under the rubble].

Existence of a *Contingency* relation potentially implies that there exists at least one causal event pair crossing the two text spans. The PDTB annotation manual states that while the *Cause* and *Condition* discourse relations indicate casual influence in their text spans, there is no causal influence in the text spans of the *Pragmatic-cause* and *Pragmatic-condition* relations. For instance, *Pragmatic-condition* indicates that one span provides the context in which the description of the situation in the other span is relevant; for example:

Pragmatic-condition: If [you are thirsty], [there's beer in the fridge].

Hence, there is a need to also identify fine-grained discourse relations.

Expansion Connectives evoking *Expansion* discourse relations expand the discourse, such as by providing additional information, illustrating alternative situations, etc. An example sentence is:

Conjunction: [Over the past decade, *x* women were killed] and [*y* went missing].

Most of the *Expansion* fine-grained relations (except for *Conjunction*, which could connect arbitrary pieces of text spans) should not contain causality relations across its text spans.

Temporal These indicate that the situations described in the text spans are related temporally. An example sentence is:

Synchrony: [He was sitting at his home] when [the whole world started to shake].

Temporal precedence of the (cause) event over the (effect) event is a necessary, but not sufficient requisite for causality. Hence by itself, *Temporal* relations are probably not discriminative enough for determining event causality.

4.2 Discourse Relation Extraction System

Our work follows the approach and features described in the state-of-the-art Ruby-based discourse system of (Lin et al., 2010), to build an in-house Java-based discourse relation extraction system. Our system identifies explicit connectives in text, predict their discourse relations, as well as their associated text spans. Similar to (Lin et al., 2010), we achieved a competitive performance of slightly over 80% F1-score in identifying fine-grained relations for explicit connectives. Our system is developed using the Learning Based Java modeling lan-

[6] PDTB further refines these fine-grained relations into a final third level of relations, but we do not use them in this work.

guage (LBJ) (Rizzolo and Roth, 2010) and will be made available soon. Due to space constraints, we refer interested readers to (Lin et al., 2010) for details on the features, etc.

In the example sentences given thus far in this section, all the connectives were explicit, as they appear in the texts. PDTB also provides annotations for implicit connectives, which we do not use in this work. Identifying implicit connectives is a harder task and incorporating these is a possible future work.

5 Joint Inference for Causality Extraction

To exploit the interactions between event pair causality extraction and discourse relation identification, we define appropriate constraints between them, which can be enforced through the Constrained Conditional Models framework (aka ILP for NLP) (Roth and Yih, 2007; Chang et al., 2008). In doing this, the predictions of CEA (Section 2.1) and the discourse system are forced to cohere with each other. More importantly, this should improve the performance of using only CEA to extract causal event pairs. To the best of our knowledge, this approach for causality extraction is novel.

5.1 CEA & Discourse: Implementation Details

Let $\mathcal{E}$ denote the set of event mentions in a document. Let $\mathcal{EP} = \{(e_i, e_j) \in \mathcal{E} \times \mathcal{E} \mid e_i \in \mathcal{E}, e_j \in \mathcal{E}, i < j, |sent(e_i) - sent(e_j)| \leq 2\}$ denote the set of event mention pairs in the document, where $sent(e)$ gives the sentence number in which event e occurs. Note that in this work, we only extract event pairs that are at most two sentences apart. Next, we define $\mathcal{L}_{ER}$ = {"causal", "¬ causal"} to be the set of event relation labels that an event pair $ep \in \mathcal{EP}$ can be associated with.

Note that the CEA metric as defined in Section 2.1 simply gives a score without it being bounded to be between 0 and 1.0. However, to use the CEA score as part of the inference process, we require that it be bounded and thus can be used as a binary prediction, that is, predicting an event pair as *causal* or ¬*causal*. To enable this, we use a few development documents to automatically find a threshold CEA score that separates scores indicating *causal* vs ¬*causal*. Based on this threshold, the original CEA scores are then rescaled to fall within 0 to 1.0. More details on this are in Section 6.2.

Let $\mathcal{C}$ denote the set of connective mentions in a document. We slightly modify our discourse system as follows. We define $\mathcal{L}_{DR}$ to be the set of discourse relations. We initially add all the fine-grained discourse relations listed in Table 1 to $\mathcal{L}_{DR}$. In the PDTB corpus, some connective examples are labeled with just a coarse-grained relation, without further specifying a fine-grained relation. To accommodate these examples, we add the coarse-grained relations *Comparison*, *Expansion*, and *Temporal* to $\mathcal{L}_{DR}$. We omit the coarse-grained *Contingency* relation from $\mathcal{L}_{DR}$, as we want to separate *Cause* and *Condition* from *Pragmatic-cause* and *Pragmatic-condition*. This discards very few examples as only a very small number of connective examples are simply labeled with a *Contingency* label without further specifying a fine-grained label. We then retrained our discourse system to predict labels in $\mathcal{L}_{DR}$.

5.2 Constraints

We now describe the constraints used to support joint inference, based on the predictions of the CEA metric and the discourse classifier. Let $s_c(dr)$ be the probability that connective c is predicated to be of discourse relation dr, based on the output of our discourse classifier. Let $s_{ep}(er)$ be the CEA prediction score (rescaled to range in [0,1]) that event pair ep takes on the *causal* or ¬*causal* label er. Let $x_{\langle c,dr\rangle}$ be a binary indicator variable which takes on the value 1 iff c is labeled with the discourse relation dr. Similarly, let $y_{\langle ep,er\rangle}$ be a binary variable which takes on the value 1 iff ep is labeled as er. We then define our objective function as follows:

$$\max \Big[|\mathcal{L}_{DR}| \sum_{c \in \mathcal{C}} \sum_{dr \in \mathcal{L}_{DR}} s_c(dr) \cdot x_{\langle c,dr\rangle} + |\mathcal{L}_{ER}| \sum_{ep \in \mathcal{EP}} \sum_{er \in \mathcal{L}_{ER}} s_{ep}(er) \cdot y_{\langle ep,er\rangle} \Big] \quad (6)$$

subject to the following constraints:

$$\sum_{dr \in \mathcal{L}_{DR}} x_{\langle c,dr\rangle} = 1 \quad \forall c \in \mathcal{C} \quad (7)$$

$$\sum_{er \in \mathcal{L}_{ER}} y_{\langle ep,er\rangle} = 1 \quad \forall ep \in \mathcal{EP} \quad (8)$$

$$x_{\langle c,dr\rangle} \in \{0,1\} \quad \forall c \in \mathcal{C}, dr \in \mathcal{L}_{DR} \quad (9)$$

$$y_{\langle ep,er\rangle} \in \{0,1\} \quad \forall ep \in \mathcal{EP}, er \in \mathcal{L}_{ER} \quad (10)$$

Equation (7) requires that each connective c can only be assigned one discourse relation. Equation (8) requires that each event pair ep can only be *causal* or ¬*causal*. Equations (9) and (10) indicate that $x_{\langle c,dr\rangle}$ and $y_{\langle ep,er\rangle}$ are binary variables.

To capture the relationship between event pair causality and discourse relations, we use the following constraints:

$$x_{\langle c,\text{“}Cause\text{”}\rangle} \leq \sum_{ep\in\mathcal{EP}_c} y_{\langle ep,\text{“}causal\text{”}\rangle} \quad (11)$$

$$x_{\langle c,\text{“}Condition\text{”}\rangle} \leq \sum_{ep\in\mathcal{EP}_c} y_{\langle ep,\text{“}causal\text{”}\rangle}, \quad (12)$$

where both equations are defined $\forall c \in \mathcal{C}$. $\mathcal{EP}_c$ is defined to be the set of event pairs that cross the two text spans associated with c. For instance, if the first text span of c contains two event mentions e_i, e_j, and there is one event mention e_k in the second text span of c, then $\mathcal{EP}_c = \{(e_i, e_k), (e_j, e_k)\}$. Finally, the logical form of Equation (11) can be written as: $x_{\langle c,\text{“}Cause\text{”}\rangle} \Rightarrow y_{\langle ep_i,\text{“}causal\text{”}\rangle} \vee \ldots \vee y_{\langle ep_j,\text{“}causal\text{”}\rangle}$, where $ep_i, \ldots, ep_j$ are elements in $\mathcal{EP}_c$. This states that if we assign the *Cause* discourse label to c, then at least one of $ep_i, \ldots, ep_j$ must be assigned as *causal*. The interpretation of Equation (12) is similar.

We use two more constraints to capture the interactions between event causality and discourse relations. First, we defined $\mathcal{C}_{ep}$ as the set of connectives c enclosing each event of ep in each of its text spans, i.e.: one of the text spans of c contain one of the event in ep, while the other text span of c contain the other event in ep. Next, based on the discourse relations in Section 4.1, we propose that when an event pair ep is judged to be *causal*, then the connective c that encloses it should be evoking one of the discourse relations in $\mathcal{L}_{DR_a}$ = {“Cause”, “Condition”, “Temporal”, “Asynchronous”, “Synchrony”, “Conjunction”}. We capture this using the following constraint:

$$y_{\langle ep,\text{“}causal\text{”}\rangle} \leq \sum_{dr_a\in\mathcal{L}_{DR_a}} x_{\langle c,dr_a\rangle} \quad \forall c \in \mathcal{C}_{ep} \quad (13)$$

The logical form of Equation (13) can be written as: $y_{\langle ep,\text{“}causal\text{”}\rangle} \Rightarrow x_{\langle c,\text{“}Cause\text{”}\rangle} \vee x_{\langle c,\text{“}Condition\text{”}\rangle} \ldots \vee x_{\langle c,\text{“}Conjunction\text{”}\rangle}$. This states that if we assign ep as *causal*, then we must assign to c one of the labels in $\mathcal{L}_{DR_a}$.

Finally, we propose that for any connectives evoking discourse relations $\mathcal{L}_{DR_b}$ = {“Comparison”, “Concession”, “Contrast”, “Pragmatic-concession”, “Pragmatic-contrast”, “Expansion”, “Alternative”, “Exception”, “Instantiation”, “List”, “Restatement”}, any event pair(s) that it encloses should be ¬*causal*. We capture this using the following constraint:

$$x_{\langle c,dr_b\rangle} \leq y_{\langle ep,\text{“}\neg causal\text{”}\rangle} \quad \forall\ dr_b \in \mathcal{L}_{DR_b},\ ep \in \mathcal{EP}_c, \quad (14)$$

where the logical form of Equation (14) can be written as: $x_{\langle c,dr_b\rangle} \Rightarrow y_{\langle ep,\text{“}\neg causal\text{”}\rangle}$.

6 Experiments

6.1 Experimental Settings

To collect the distributional statistics for measuring CEA as defined in Equation (1), we applied part-of-speech tagging, lemmatization, and dependency parsing (Marneffe et al., 2006) on about 760K documents in the English Gigaword corpus (LDC catalog number LDC2003T05).

We are not aware of any benchmark corpus for evaluating event causality extraction in contexts. Hence, we created an evaluation corpus using the following process: Using news articles collected from CNN[7] during the first three months of 2010, we randomly selected 20 articles (documents) as evaluation data, and 5 documents as development data.

Two annotators annotated the documents for causal event pairs, using two simple notions for causality: the Cause event should temporally precede the Effect event, and the Effect event occurs because the Cause event occurs. However, sometimes it is debatable whether two events are involved in a causal relation, or whether they are simply involved in an uninteresting temporal relation. Hence, we allowed annotations of C to indicate causality, and R to indicate relatedness (for situations when the existence of causality is debatable). The annotators will simply identify and annotate the *C* or *R* relations between predicates of event pairs. Event arguments are not explicitly annotated, although the annotators are free to look at the entire document text while making their annotation decisions. Finally, they are free

[7] http://www.cnn.com

System	Rec%	Pre%	F1%
PMI_{pp}	26.6	20.8	23.3
ECD_{pp} &$PMI_{pa,aa}$	40.9	23.5	29.9
CEA	62.2	28.0	38.6
CEA+Discourse	65.1	30.7	41.7

Table 2: Performance of baseline systems and our approaches on extracting *Causal* event relations.

System	Rec%	Pre%	F1%
PMI_{pp}	27.8	24.9	26.2
ECD_{pp} &$PMI_{pa,aa}$	42.4	28.5	34.1
CEA	63.1	33.7	43.9
CEA+Discourse	65.3	36.5	46.9

Table 3: Performance of the systems on extracting *Causal* and *Related* event relations.

to annotate relations between predicates that have any number of sentences in between and are not restricted to a fixed sentence window-size.

After adjudication, we obtained a total of 492 $C+R$ relation annotations, and 414 C relation annotations on the evaluation documents. On the development documents, we obtained 92 $C+R$ and 71 C relation annotations. The annotators overlapped on 10 evaluation documents. On these documents, the first (second) annotator annotated 215 (199) $C+R$ relations, agreeing on 166 of these relations. Together, they annotated 248 distinct relations. Using this number, their agreement ratio would be 0.67 $(166/248)$. The corresponding agreement ratio for C relations is 0.58. These numbers highlight that causality identification is a difficult task, as there could be as many as N^2 event pairs in a document (N is the number of events in the document). We plan to make this annotated dataset available soon.[8]

6.2 Evaluation

As mentioned in Section 5.1, to enable translating (the unbounded) CEA scores into binary *causal*, ¬*causal* predictions, we need to rescale or calibrate these scores to range in [0,1]. To do this, we first rank all the CEA scores of all event pairs in the development documents. Most of these event pairs will be ¬*causal*. Based on the relation annotations in these development documents, we scanned through this ranked list of scores to locate the CEA score *t* that gives the highest F1-score (on the development documents) when used as a threshold between *causal* vs ¬*causal* decisions. We then ranked all the CEA scores of all event pairs gathered from the 760K Gigaword documents, discretized all scores higher than *t* into B bins, and all scores lower than t into B bins. Together, these $2B$ bins represent the range [0,1]. We used $B = 500$. Thus, consecutive bins represent a difference of 0.001 in calibrated scores.

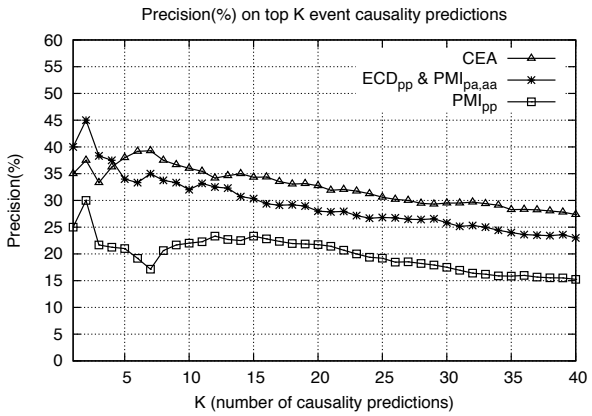

Figure 1: Precision of the top K causality C predictions.

To measure the causality between a pair of events e_i and e_j, a simple baseline is to calculate $PMI(p^i, p^j)$. Using a similar thresholding and calibration process to translate $PMI(p^i, p^j)$ scores into binary causality decisions, we obtained a F1 score of 23.1 when measured over the causality C relations, as shown in the row PMI_{pp} of Table 2.

As mentioned in Section 2.1.2, Riaz and Girju (2010) proposed the ECD metric to measure causality between two events. Thus, as a point of comparison, we replaced s_{pp} of Equation (1) with $ECD(a, b)$ of Equation (5), substituting $a = p^i$ and $b = p^j$. After thresholding and calibrating the scores of this approach, we obtained a F1-score of 29.7, as shown in the row ECD_{pp}&$PMI_{pa,aa}$ of Table 2.

Next, we evaluated our proposed CEA approach and obtained a F1-score of 38.6, as shown in the row CEA of Table 2. Thus, our proposed approach obtained significantly better performance than the PMI baseline and the ECD approach. Next, we performed joint inference with the discourse relation predictions as described in Section 5 and obtained

[8] http://cogcomp.cs.illinois.edu/page/publication_view/663

an improved F1-score of 41.7. We note that we obtained improvements in both recall and precision. This means that with the aid of discourse relations, we are able to recover more causal relations, as well as reduce false-positive predictions.

Constraint Equations (11) and (12) help to recover causal relations. For improvements in precision, as stated in the last paragraph of Section 5.2, identifying other discourse relations such as "Comparison", "Contrast", etc., provides counter-evidence to causality. Together with constraint Equation (14), this helps to eliminate false-positive event pairs as classified by CEA and contributes towards CEA+Discourse having a higher precision than CEA.

The corresponding results for extracting both causality and relatedness $C + R$ relations are given in Table 3. For these experiments, the aim was for a more relaxed evaluation and we simply collapsed C and R into a single label.

Finally, we also measured the precision of the top K causality C predictions, showing the precision trends in Figure 1. As shown, CEA in general achieves higher precision when compared to PMI_{pp} and $ECD_{pp}\&PMI_{pa,aa}$. The trends for $C + R$ predictions are similar.

Thus far, we had included both verbal and nominal predicates in our evaluation. When we repeat the experiments for $ECD_{pp}\&PMI_{pa,aa}$ and CEA on just verbal predicates, we obtained the respective F1-scores of 31.8 and 38.3 on causality relations. The corresponding F1-scores for casuality and relatedness relations are 35.7 and 43.3. These absolute F1-scores are similar to those in Tables 2 and 3, differing by 1-2%.

7 Analysis

We randomly selected 50 false-positive predictions and 50 false-negative *causality* relations to analyze the mistakes made by CEA.

Among the false-positives (precision errors), the most frequent error type (56% of the errors) is that CEA simply assigns a high score to event pairs that are not causal; more knowledge sources are required to support better predictions in these cases. The next largest group of error (22%) involves events containing pronouns (e.g. "he", "it") as arguments. Applying coreference to replace these pronouns with their canonical entity strings or labeling them with semantic class information might be useful.

Among the false-negatives (recall errors), 23% of the errors are due to CEA simply assigning a low score to causal event pairs and more contextual knowledge seems necessary for better predictions. 19% of the recall errors arises from causal event pairs involving nominal predicates that are not in our list of event evoking noun types (described in Section 3). A related 17% of recall errors involves nominal predicates without any argument. For these, less information is available for CEA to make predictions. The remaining group (15% of errors) involves events containing pronouns as arguments.

8 Related Work

Although prior work in event causality extraction in context is relatively sparse, there are many prior works concerning other semantic aspects of event extraction. Ji and Grishman (2008) extracts event mentions (belonging to a predefined list of target event types) and their associated arguments. In other prior work (Chen et al., 2009; Bejan and Harabagiu, 2010), the authors focused on identifying another type of event pair semantic relation: event coreference. Chambers and Jurafsky (2008; 2009) chain events sharing a common (protagonist) participant. They defined events as verbs and given an existing chain of events, they predict the next likely event involving the protagonist. This is different from our task of detecting causality between arbitrary event pairs that might or might not share common arguments. Also, we defined events more broadly, as those that are triggered by either verbs or nouns. Finally, although our proposed CEA metric has resemblance the ECD metric in (Riaz and Girju, 2010), our task is different from theirs and our work differs in many aspects. They focused on building a dataset of causal text spans, whereas we focused on identifying causal relations between events in a given text document. They considered text spans headed by verbs while we considered events triggered by both verbs and nouns. Moreover, we combined event causality prediction and discourse relation prediction through a global inference procedure to further improve the performance of event causality prediction.

9 Conclusion

In this paper, using general tools such as the dependency and discourse parsers which are not trained specifically towards our target task, and a minimal set of development documents for threshold tuning, we developed a minimally supervised approach to identify causality relations between events in context. We also showed how to incorporate discourse relation predictions to aid event causality predictions through a global inference procedure. There are several interesting directions for future work, including the incorporation of other knowledge sources such as coreference and semantic class predictions, which were shown to be potentially important in our error analysis. We could also use discourse relations to aid in extracting other semantic relations between events.

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments and suggestions. University of Illinois at Urbana-Champaign gratefully acknowledges the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract No. FA8750-09-C-0181. The first author thanks the Vietnam Education Foundation (VEF) for its sponsorship. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the VEF, DARPA, AFRL, or the US government.

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *CICLING*.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *ACL*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL-HLT*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL*.

Ming-Wei Chang, Lev Ratinov, Nicholas Rizzolo, and Dan Roth. 2008. Learning and inference with constraints. In *AAAI*.

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

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *ACL workshop on Multilingual Summarization and Question Answering*.

Olga Gurevich, Richard Crouch, Tracy Holloway King, and Valeria de Paiva. 2008. Deverbal nouns in knowledge representation. *Journal of Logic and Computation*, 18, June.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through unsupervised cross-document inference. In *ACL*.

Claudia Leacock and Martin Chodorow, 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*. MIT Press.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. Technical report. http://www.comp.nus.edu.sg/ linzihen/publications/tech2010.pdf.

Marie-catherine De Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The penn discourse treebank 2.0 annotation manual. Technical report. http://www.seas.upenn.edu/ pdtb/PDTBAPI/pdtb-annotation-manual.pdf.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *ICSC*.

N. Rizzolo and D. Roth. 2010. Learning based java for rapid development of nlp systems. In *LREC*.

Dan Roth and Wen Tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *CoNLL*.

Dan Roth and Wen Tau Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. http://framenet.icsi.berkeley.edu.

Yizhou Sun, Ning Liu, Kunqing Xie, Shuicheng Yan, Benyu Zhang, and Zheng Chen. 2007. Causal relation of queries from temporal logs. In *WWW*.

Patrick Suppes. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.