# Toward a Better Understanding of Causality between Verbal Events: Extraction and Analysis of the Causal Power of Verb-Verb Associations

**Mehwish Riaz and Roxana Girju**
Department of Computer Science and Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{mriaz2, girju}@illinois.edu

## Abstract

The identification of causal relations between verbal events is important for achieving natural language understanding. However, the problem has proven notoriously difficult since it is not clear which types of knowledge are necessary to solve this challenging problem close to human level performance. Instead of employing a large set of features proved useful in other NLP tasks, we split the problem in smaller sub problems. Since verbs play a very important role in causal relations, in this paper we harness, explore, and evaluate the predictive power of causal associations of verb-verb pairs. More specifically, we propose a set of knowledge-rich metrics to learn the likelihood of causal relations between verbs. Employing these metrics, we automatically generate a knowledge base ($KB_c$) which identifies three categories of verb pairs: Strongly Causal, Ambiguous, and Strongly Non-causal. The knowledge base is evaluated empirically. The results show that our metrics perform significantly better than the state-of-the-art on the task of detecting causal verbal events.

## 1 Introduction

The identification of semantic relations between events is a mandatory component of natural language understanding. In this paper, we focus on the identification of causal relations between events represented by verbs. Following Riaz and Girju (2010), we define a verbal event $e_{v_i}$ as "[subject$_{v_i}$] $v_i$ [object$_{v_i}$]", where the subject and object of the verb may or may not be explicitly present in an instance. Consider the following examples:

1. Yoga **builds** stamina because you **maintain** your poses for a certain period of time. (CAUSE ($e_{maintain}, e_{build}$))
2. The monster storm Katrina **raged** ashore along the Gulf Coast Monday morning. There were early reports of buildings **collapsing** along the coast. (CAUSE ($e_{rage}, e_{collapse}$))

In example 1, the two bold events are causally connected by an explicit and unambiguous discourse marker (*because*). However, in English, not all discourse markers unambiguously identify causality (Prasad et al., 2008) - for example, Bethard and Martin (2008) proposed a corpus of 1000 causal and non-causal event pairs conjoined by the marker *and*. Even more, causal relations can be encoded by implicit contexts - i.e., those where no discourse marker is present (example 2). Despite the recent achievements obtained in discourse processing, it is still unclear what types of knowledge can contribute most towards detecting causality in both explicit and implicit contexts (Sporleder and Lascarides, 2008). The complexity of the task of detecting causality between events stems from the fact that there are many factors involved, such as contextual features of an instance (e.g., lexical items, tenses of verbs, arguments of verbs, etc.), semantic and pragmatic features of events, background knowledge, world knowledge, common sense, etc. Prior approaches have employed contextual features of an instance to identify causality between events or discourse segments (Bethard and Martin, 2008; Pitler and Nenkova, 2009; Pitler et al., 2009). Although contextual features provide important knowledge about sentence(s) in which events appear, humans also make use of other information such as background knowledge to comprehend causality. For instance, in example 2 we use knowledge about the causal association between verbal entities **rage** and **collapse** to label it with causality.

This research is motivated by the need to extract and analyze other type of knowledge necessary for the identification of causal relations between verbal events. We start from the fact that verbs are the



*Proceedings of the SIGDIAL 2013 Conference*, pages 21–30,
Metz, France, 22-24 August 2013. 

main components of language to express events and semantic relations between events. Thus, in order to identify and extract causal relations between events (denoted by $(e_{v_i}, e_{v_j})$), it is critical for a model to employ knowledge about the tendency of a verb pair $(v_i, v_j)$ to encode causation. For example, the pair (kill, arrest) has a high tendency to encode a cause relation irrespective of the context in which it is used, thereby a good indicator of causality. The state-of-the-art resources on verb semantics, such as WordNet, VerbNet, PropBank, FrameNet, etc. (Miller, 1990; Kipper et al., 2000; Kingsbury et al., 2002; Baker et al., 1998), provide information about the semantic classes, thematic roles and selectional restrictions of verbs. Among these, WordNet is the only resource which provides information about the cause relation between verbs, but it has very limited coverage. For VerbOcean, a semi-automatically generated resource, Chklovski and Pantel (2004) have used explicit lexical patterns (e.g., "verb * by verb") as means of mining enablement (cause-effect) relations between verbs. Such approaches help detecting causality with high precision but suffer from limited coverage due to the highly implicit nature of language. Moreover, such resources do not provide any information about the likelihood of a causal relation in verb pairs - e.g., (kill, arrest) has a high tendency to encode cause relation as compared with the pair (build, maintain). The pair (build, maintain) seems ambiguous because it can encode both cause and non-cause relations depending on the context, as shown by examples 1 and 3. Thus, causality detection models should employ knowledge about which verb pairs are strongly causal (non-causal) in nature and for which pairs the context plays an important role to signal causality.

3. Republicans had not cut the funds for **maintaining** the levee and **building** up the ecological protections. (NON-CAUSE)

We propose a fully automated procedure to learn the likelihood of causal relations in verb pairs. In this process, we create three categories of verb pairs: Strongly Causal ($S_c$), Ambiguous ($A_c$) and Strongly Non-causal ($S_{\neg c}$). The result is a knowledge base ($KB_c$) of causal associations of verbs. In $KB_c$, the category $S_c$ ($S_{\neg c}$) contains the verb pairs which have the greatest (least) likelihood to encode a causal relation, respectively. However, the category $A_c$ contains ambiguous verb pairs which have the likelihood to encode both causal and non-causal relations. The information about such causal associations provides a rich knowledge source to causality detection models.

The main contributions of our research are as follows:

- We propose a set of novel metrics (i.e., Explicit Causal Association (ECA), Implicit Causal Association (ICA) and Boosted Causal Association (BCA)) to identify the likelihood of verb pairs to encode causality. Our metrics exploit the information available from a large number of unlabeled explicit and implicit instances of verb pairs for this purpose.
- We introduce an automated procedure to build a training corpus of causal and non-causal event pairs. This prevents us from the trouble of annotating a large number of event pairs for cause and non-cause relations. Our metrics make use of supervision from the training corpus to identify causality in verb pairs. We also provide a mechanism to determine causal verb pairs which remain undiscovered due to the issue of training data sparseness.
- We revisit recent approaches employing distributional similarity methods to predict causality between events (Riaz and Girju, 2010; Do et al., 2011). The state-of-the-art metric Cause-Effect Association (CEA) (Do et al., 2011) identifies causality mainly based on probabilities of verb-verb, verb-argument, and argument-argument pairs. In comparison with CEA, our metrics perform significantly better by improving the prior knowledge about the causal associations from CEA's components.

After a brief review of related work in next section, we describe our approach for acquisition of training corpus in section 3. The model for the extraction of causal associations is presented in section 4, followed by the evaluation and discussion in section 5 and conclusion in section 6.

## 2 Related Work

Causality has long been studied from various perspectives by philosophers, data-mining researchers and computer scientists (Menzies, 2008; Woodward, 2008; Suppes, 1970; Silverstein et al., 2000; Pearl, 2000).

In NLP, the problem of detecting causality between events is a very challenging but less researched topic. Previously, researchers have stud-

ied this task by focusing on supervised classification models for both verbal and nominal events (Girju, 2003; Bethard and Martin, 2008). Bethard and Martin (2008), for example, have focused mainly on the contextual features available in test instances of verbal event pairs to predict causality. They have relied on a small scale dataset of 1000 instances (697 training and 303 test) for this task. Unlike above models, recently some researchers have employed unsupervised causality detection metrics and minimal supervision for this task. For example, Riaz and Girju (2010) have proposed an unsupervised metric Effect-Control Dependency (ECD) to determine causality between events in news scenarios. Following their model, Do et al. (2011) introduced an improved metric CEA which uses PMI and some components of ECD to predict the causal relation in verbal and nominal event pairs in a text document. They also proposed a minimally supervised method using explicit discourse markers. For example, they used ILP framework to assign a non-causal relation to all the event pairs appearing in two discourse segments connected by a non-causal marker. They evaluated their model on a set of 20 documents, a highly skewed evaluation set with around 2-3% causal instances and 58% human inter-annotator agreement on cause-effect relations. On verbal events, they reported 38.3% F-score with CEA and 1-2% improvement using minimally supervised method. As compared with above mentioned metrics, we introduce knowledge rich association measures which employ supervision from the automatically generated training corpus to learn causality.

Several other NLP researchers have studied related topics e.g., identifying events, building of temporal chain of events sharing a common protagonist (participant), predicting future events and identifying hidden links in news articles to build a coherent chain (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Radinsky and Horvitz, 2013; Shahaf and Guestrin, 2010). Unlike these tasks, our focus is on identifying causality between events.

## 3 Acquisition of Training Corpus

In this section, we propose a fully automated procedure to build a training corpus of event pairs which encode cause and non-cause relations. This training corpus is used in our model to identify the likelihood of cause relations in verb pairs. As discussed earlier, previous researchers have worked with a small scale dataset of annotated event pairs. The current task requires us to use a large training corpus to learn the pervasive relation of causality and the manual generation of such corpus is a laborious task. Therefore, we decided to depend on the unambiguous discourse markers *because* and *but* to automatically collect training instances of cause and non-cause event pairs, respectively. For example, the marker *because* in the instance 1 of section 1 encodes a cause relation between the events $e_{build}$ and $e_{maintain}$. Some researchers have utilized unambiguous discourse markers to acquire training instances of semantic relations between discourse segments (Marcu and Echihabi, 2001; Sporleder and Lascarides, 2008). However, the process is not simple for the current problem since it is not always clear how to create a causal instance of an event pair. Consider the following meta instance $I$:

$I$ : <s>/$m_1 \ldots v_1 \ldots v_2 \ldots v_k \ldots$ *because* $\ldots v_{k+1} \ldots v_{k+2}, \ldots, v_r, \ldots m_2$/</s>.

It is composed of main verbs ($v_1$, $v_2$, ..., $v_r$), discourse markers ($m_1$, $m_2$), and sentence boundaries (<s>, </s>). Here, we assume that the discourse markers or the sentence boundaries whichever appear first in $I$ represent the boundaries of discourse segments for the marker *because* (appendix A contains a table of notations used in this paper). In $I$, there are $k$ and $r-k$ main verbs appearing before and after *because*, respectively. The problem here is to determine the event pair encoding causality out of $k \times (r-k)$ choices. Here, we consider that the most dependent pair among all choices in $I$ is the best candidate to encode causality.

In this work, we propose the following function $f(I)$ to pick the most dependent pair:

$$f(I) = \underset{(v_i \prec m_c, v_j \succ m_c)}{\arg\max} CD(v_i, v_j) \times PS_I(v_i, v_j) \quad (1)$$

Here, $i$ ($j$) refers to all verbs that appear before (after) the causal marker (i.e., $m_c$) *because* in $I$. CD (equation 2) is a component of predicate-predicate association of CEA (Do et al., 2011) to determine causal dependency of a pair $(v_i, v_j)$. Do et al. (2011) used the score CD to determine causality in an unsupervised fashion but here we employ this to build a training corpus of causal event pairs.

$$CD(v_i, v_j) = PMI(v_i, v_j) \times max(v_i, v_j) \times IDF(v_i, v_j) \quad (2)$$

The functions PMI, max and IDF depend on co-occurrence probabilities and idf scores to determine causal dependency. Due to space limitations, for details we refer the reader to Do et al. (2011).

Next, we define a novel penalization factor $PS_I$ for the verbs of a pair appearing at greater distance from the causal marker *because*. For example, this assumes the verbs in the pair $(v_2, v_{k+2})$ are less likely to be in a cause relation as compared with $(v_k, v_{k+1})$ in $I$. We came up with this idea because our initial experiments revealed that the causal instances obtained by penalizing CD with $PS_I$ provide better training for our model as compared to using only CD for this purpose. The similar behavior of reduction in the likelihood of causality with respect to increase in distance between two events was observed by Riaz and Girju (2010).

$$PS_I(v_i, v_j) = -\log \frac{pos(v_i) + pos(v_j)}{2.0 \times (C(v_p) + C(v_q))} \quad (3)$$

Here, $C(v_p)$ ($C(v_q)$) is the count of the main verbs appearing before (after) *because*, respectively. The distance of the verb is measured in terms of its position (i.e., $pos(v_i)$) with respect to *because*. The position is 1 for the verb closest to *because* and 2 for the verb next to the closest verb. $PS_I$ has maximum value for $(v_k, v_{k+1})$ and it reduces for other pairs with verbs at greater distance from *because* in instance $I$.

In order to extract non-causal event pairs, we utilized instances with two discourse segments conjoined by the marker *but* which represents comparison (non-causal) relation. Any event pair collected from the two discourse segments in non-causal relation encodes non-causality. Therefore, we depend on selecting the closest verb pair from the instances of form $I$ with marker *but* instead of *because*.

In this paper, we present the results produced using a training corpus of 240K instances (50% for each class) from the English Gigaword Corpus. In order to prepare this corpus, we identified discourse markers (i.e., $m_1$, $m_2$), if available, before and after *because/but* in each instance $I$ and assumed that only those markers which have discourse usage in $I$ define boundaries of discourse segments of *because/but*. We used the list of 100 explicit discourse markers provided by Prasad et al. (2008) and the supervised approach of Pitler and Nenkova (2009) to detect markers and the discourse versus non-discourse usage of these markers. We use this training corpus to identify causation for both explicit and implicit instances of event pairs. Using this training corpus, a model tends to give higher causal weights to those instances in which events are connected by the explicit causal marker *because* as compared to implicit instances of causation. Thus, to provide fair supervision to both explicit and implicit instances of event pairs, we remove the cue words *because* and *but* which were used to automatically label the training instances.

## 4 Causal Associations of Verb Pairs

In this section, we explain our approach to learn the likelihood of causal relations in verb pairs by exploiting information available from both explicit and implicit instances of these pairs. We extracted around $12,000$ documents from the English Gigaword corpus to collect instances of verb pairs from single sentences (intra-sentential) and adjacent sentences (inter-sentential) of text. In this set, we added instances from $3,000$ articles on news stories "Hurricane Katrina" and the "Iraq war". These articles were collected and used to identify causal relations in news scenarios by Riaz and Girju (2010). We used these collections because natural disaster and war-related news articles are rich in causal events and chains of such events. In order to identify the causal associations with high confidence, we decided to apply our model on those verb pairs which have at least 30 instances in the above mentioned documents. We acquired $10,455$ such verb pairs. The set of intra- and inter-sentential instances of these verb pairs is referred to as the development set for our model.

### 4.1 Explicit Causal Association (ECA)

In order to find the likelihood of a verb pair to encode causal relations, we define our novel metric Explicit Causal Association (ECA) as follows:

$$ECA(v_i, v_j) = \frac{1}{\mid VP \mid} \sum_{I_{(v_i, v_j)} \in VP} (CD(v_i, v_j) \times C_I) \quad (4)$$

where $VP$ is the set of intra- and inter-sentential instances (denoted by $I(v_i, v_j)$) of the verb pair $(v_i, v_j)$, CD determines the causal dependency of the verb pair in unsupervised fashion (equation 2), and $C_I$ finds the tendency of instance $I$ of $(v_i, v_j)$ to belong to the cause class as compared to the non-cause class using training corpus of event pairs. The goal of ECA is to combine the unsupervised causal dependency (i.e., CD) with the supervised score of instance I of belonging to cause

class than the non-cause one (i.e., $C_I$). Here, CD represents the prior knowledge about the causal association based on co-occurrence probabilities and idf scores (equation 2). It can discover lots of false positives because the co-occurrence probabilities can fail to differentiate causality from any other type of correlation. Therefore, we improve this prior knowledge with the help of supervision from the training corpus containing instances of both cause and non-cause relations. The global decision of the causal association is made by taking the average of scores on all the instances containing that verb pair. Notice that CD can also be moved out from the summation function in equation 4.

We define the function $C_I$ as follows:

$$C_I = \sum_{k=1}^{n} log(\frac{P(f_k \mid c)}{P(f_k \mid \neg c)}) \quad (5)$$

Here, the notations c and ¬ c represent cause and non-cause class, respectively. The notation $f_k$ represents the feature of an instance $I$. In this work, we use some language features of events and context of an instance $I$ which are defined later in this section. P($f_k \mid c$) and P($f_k \mid \neg c$) are the smoothed probabilities of feature $f_k$ given the cause and non-cause training instances. The value of $C_I$ is positive only when the instance $I$ has more tendency to encode a cause relation than a non-cause one. To avoid negative values, we map $C_I$ scores to the range $[0, 1]$ using $\frac{C_I - C_{min}}{C_{max} - C_{min}}$ where $C_{min}$ ($C_{max}$) is the minimum (maximum) value of $C_I$ obtained on our development set, respectively. Also, we add a small value $\epsilon$ to $C_I$ to avoid 0 value. Similarly, to avoid negative scores of PMI in equation 2 we can map it to the range [0,1].

We present below the features for the calculation of $C_I$ . We use lexical, syntatic and semantic features on verbs and verb phrases of both events of a pair. These features include words, lemmas, part-of-speech tags, all senses from WordNet for the verbs and the lexical items of verb phrases. These features were introduced by Bethard and Martin (2008) (for an in-depth description of these features see Bethard and Martin (2008)). Next, we describe the set of features which are the contributions of this research.

1. **Verbs Arguments:** Words, lemma, part-of-speech tags and all senses from WordNet for subject and object of verbs of both events.
2. **Verbs and Arguments Pairs:** For this feature, we take the cross product of both events of a pair ($e_{v_i}$,$e_{v_j}$) where $e_{v_i}$ = [subject$_{v_i}$] $v_i$ [object$_{v_i}$] and $e_{v_j}$ = [subject$_{v_j}$] $v_j$ [object$_{v_j}$]. Some examples of this feature are (subject$_{v_i}$,subject$_{v_j}$), (subject$_{v_i}$,$v_j$), (subject$_{v_i}$,object$_{v_j}$), etc. In this work, we use unordered pairs as features (i.e., ($v_i$,$v_j$)) is same as ($v_j$,$v_i$)) because the temporal order of events is unknown for the unlabeled development set instances. In future, this feature can be improved by adding temporal information.

The next three features are taken from the minimum relevant context ($min_{context}$) of a verb pair which we define as follows. $min_{context}$ of a pair ($v_i$, $v_j$) in an intra-sentential instance is <s>/$m_1 \ldots v_i \ldots v_j \ldots m_2$/</s> – i.e., words between the discourse markers (i.e., $m_1$, $m_2$) or sentence boundaries (i.e., <s>, </s>) whichever appear first in the sentence. The $min_{context}$ for the pair ($v_i$, $v_j$) in an inter-sentential is given below:

<s> / $m_1 \ldots v_i \ldots m_2$ / </s>
<s> / $m_1 \ldots v_j \ldots m_2$ / </s>

3. **Context Words:** Lemmas of all words from $min_{context}$. This feature captures words other than two events.
4. **Context Main Verbs:** All main verbs and their lemmas from $min_{context}$. It collects information about all verbs that appear with the causal and non-causal event pair.
5. **Context Main Verb Pairs:** The pairs of main verbs from $min_{context}$. The lemmas are taken from the feature "Context Main Verbs" and then the pairs on these lemmas are used as this feature. For example, for lemmas of verbs (i.e., $v_1, v_2, \ldots, v_k$), pairs (i.e., ($v_1, v_2$), ($v_1, v_k$), etc.) are used for this feature. This feature is used to get information about the interesting causal chains of verbs that may appear in causal instances.

We propose next a novel metric ICA to avoid the problem of training data sparsity.

### 4.2 Implicit Causal Association (ICA)

In order to determine the causal associations using ECA, we depend on explicit cause and non-cause training instances for supervision. However, it is possible that some strongly causal verb pairs may frequently appear in implicit causal contexts. Therefore, the causality of such pairs can remain uncaptured by ECA which merely relies on explicit training instances. For example, a pair (fall,

break) seems strongly causal, but it does not appear often in our explicit training corpus due to training data sparsity. Thus, in order to handle this problem, we propose a new metric called ICA. This metric makes use of functions for the identification of roles of events in a cause relation. After briefly describing the roles of events in causal relations below, we continue with the description of ICA.

#### 4.2.1 Roles of Events in Cause Relation

Each of the two events in a cause relation can be assigned either cause or effect role. For example for the following training instance, the verb appearing after *because* represents cause event and the verb before *because* represents effect event.

1. Yoga *builds* stamina because you **maintain** your poses for a certain period of time. (**Role:** $r_C$)
2. Yoga **builds** stamina because you *maintain* your poses for a certain period of time. (**Role:** $r_E$)

The notation $r_C$ and $r_E$ represents the classes of cause and effect role of events, respectively. We use core features of events to determine the likelihood of their roles in causation. These features include lemma, part-of-speech tag, all senses from WordNet of both verbs and their arguments (i.e., subject and object). Next, we use these features to handle training data sparseness.

#### 4.2.2 Handling of Training Data Sparsity

To deal with the problem of training data sparsity, we define the metric ICA as follows:

$$ICA(v_i, v_j) = \frac{1}{\mid VP \mid} \sum_{I_{(v_i,v_j)} \in VP} (CD(v_i, v_j) \times C_I \times ERM_{(e_{v_i}, e_{v_j})}) \quad (6)$$

where CD and $C_I$ are defined earlier and ERM determines the likelihood of roles of the events in the cause relation. We remind the reader that CD is the unsupervised causal dependency of verb pair and $C_I$ is the likelihood of instance $I$ of the verb pair to belong to the cause class than the non-cause one using full set of features from section 4.1.

Events Roles Matching ($ERM_{(e_{v_i}, e_{v_j})}$) (equations 7 and 8) is the negative log-likelihood of events $e_{v_i}$ and $e_{v_j}$ appearing as cause or effect role determined using the explicit causal instances of the training corpus and the core features of events defined in section 4.2.1.

$$ERM_{(e_{v_i}, e_{v_j})} = -1.0 \times \max(S(e_{v_i}, r_C) + S(e_{v_j}, r_E), S(e_{v_i}, r_E) + S(e_{v_j}, r_C)) \quad (7)$$

$$S(e_{v_i}, r_C) = \sum_{k=1}^{n} log(P(f_k \mid r_C)) \quad (8)$$

$$S(e_{v_j}, r_E) = \sum_{k=1}^{n} log(P(f_k \mid r_E))$$

Here, $S(e_{v_i}, r_C)$ is the score of $e_{v_i}$ being the cause event and $S(e_{v_j}, r_E)$ is the score of $e_{v_j}$ being the effect event. These scores are computed using smoothed probabilities – i.e., $P(f_k \mid r_C)$ and $P(f_k \mid r_E)$. Similarly, $S(e_{v_i}, r_E)$ and $S(e_{v_j}, r_C)$ are calculated and max is taken. The high value of ERM represents low matching of an event pair (verbs and their arguments) in the explicit causal instances of the training corpus. The high value of ERM of an event pair can have one of the following two interpretations: (A) it is a non-causal event pair, or (B) it is a causal event pair but this pair and the pairs which are semantically closer to it hardly appear in explicit causal contexts. In the metric ICA, $C_I \times CD(v_i, v_j)$ is used as a guiding score to interpret ERM as follows:

1. If $C_I \times CD(v_i, v_j)$ has high score then the value of ERM is not penalized by this guiding score because ERM's value can be interpreted using (B) above.
2. If $C_I \times CD(v_i, v_j)$ has low score then the value of ERM is penalized by this guiding score because $(e_{v_i}, e_{v_j})$ can be a non-causal pair according to the interpretation (A) above.

ICA is a boosting factor to determine causal verb pairs which remain undiscovered because of training data sparseness. We also define a Boosted Causal Association (BCA) metric by adding ICA to original ECA metric as follows:

$$BCA(v_i, v_j) = \frac{1}{\mid VP \mid} \sum_{I_{(v_i,v_j)} \in VP} (CD(v_i, v_j) \times C_I + CD(v_i, v_j) \times C_I \times ERM_{(e_{v_i}, e_{v_j})}) \quad (9)$$

To build the knowledge base of causal associations ($KB_c$), we generate a ranked list of all verb pairs based on the likelihood of causality encoded by these pairs. Here, we assume that verb pairs are uniformally distributed across three categories - i.e., top one-third and bottom one-third ranked verb pairs belong to Strongly Causal ($S_c$) and Strongly Non-Causal ($S_{\neg c}$) categories and rest of the pairs are considered Ambiguous ($A_c$). Following our assumption, we evaluate this categorization in next section, but in future researchers can perform empirical study of how to automatically cluster verb pairs into three or more categories with respect to causation.

## 5 Evaluation and Discussion

In this section, we present our evaluation of knowledge base to identify causality between verbal events. Specifically we performed experiments to evaluate (1) the ranking of verb pairs based on their likelihood of encoding causality, and (2) the quality of the three categories of verb pairs in $KB_c$ (i.e., $S_c$, $A_c$ and $S_{\neg c}$). For this purpose, we collected two test sets. For each test set, we randomly selected $50$ verb pairs from the list of $10,455$ verb pairs in $KB_c$. For each verb pair, we selected randomly 3 intra- and 3 inter-sentential instances from the English Gigaword corpus and the "Hurricane Katrina" and "Iraq war" articles. In order to keep the development set different from the test sets, we automatically traversed the development set to determine if any test instance is available in it. In case of finding any such test instance, we removed it from the development set to perform evaluation on unseen test instances. Two annotators were asked to provide Cause or Non-Cause labels for each instance. They were provided with annotation guidelines from the manipulation theory of causality (Woodward, 2008). Given these guidelines have been successfully used by Riaz and Girju (2010), we use them here as well. For ease of annotation, we randomly selected inter-sentential instances such that the length of each sentence is at most $40$ words.

The human inter-annotator agreement achieved on Test-set$_1$ (Test-set$_2$) is $90\%$ ($88.3\%$) and the agreement on the cause class is $70\%$ ($62.7\%$), respectively. The kappa score on Test-set$_1$ (Test-set$_2$) is $0.75$ ($0.69$), respectively. The Test-set$_1$ (Test-set$_2$) contains $25\%$ ($22\%$) causal instances, respectively.

We employed Spearman's rank correlation coefficient (equation 10) to compare the ranked list of verb pairs based on the scores of our metrics and the rank given by the human annotators. The score $P$ ranges from $+1$ to $-1$ where $+1$ and $-1$ show strong and negative correlation, respectively.

$$P = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \quad (10)$$

Here, $n$ is the total number of verb pairs in the test set, $x_i$ is the human annotation rank and $y_i$ is the metric's rank of verb pair $i$ of the test set. The values of $x_i$ and $y_i$ are determined as follows. For each verb pair, C$_h$ is calculated which is the number of cause labels given by both human annotators out of 6 instances of a verb pair. The pairs are ranked in descending order according to the score C$_h$ s.t. the top scored pair(s) gets rank $50$ and the next to the top pair(s) gets rank $49$ and so on. Similarly, ranks are given to the verb pairs according to the metric's scores. This way of evaluation was also used by Beamer and Girju (2009) for temporally ordered adjacent verb pairs. But here, we are working with verb pairs appearing in any temporal order in both intra- and inter-sentential instances.

| Metric | CEA | ECA | ICA | BCA |
|---|---|---|---|---|
| Test-set$_1$ | -0.077 | 0.144 | 0.427 | 0.435 |
| Test-set$_2$ | 0.167 | 0.217 | 0.353 | 0.338 |

Table 1: The Spearman's rank correlation coefficient for the metrics CEA, ECA, ICA and BCA.

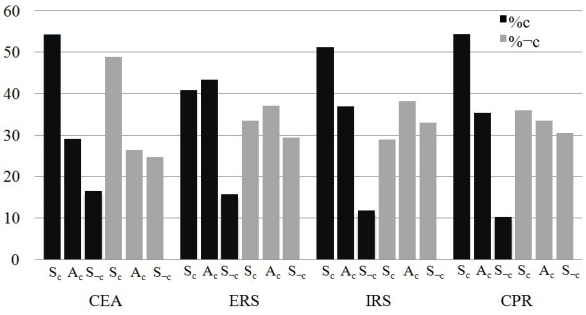


Figure 1: The percentage of causal ($\%c$) and non-causal ($\%\neg c$) test instances in $S_c$, $A_c$ and $S_{\neg c}$ generated by the metrics CEA, ECA, ICA and BCA.

We used ECA, ICA and BCA scores to generate the ranked list of all verb pairs. In this work, we also used the state-of-the-art causality identifier CEA (Do et al., 2011) as baseline metric. For each verb pair, we computed the likelihood of causality by taking the average of CEA scores on all instances of that pair in the development set.

The results with Spearman's rank correlation coefficient in Table 1 show that CEA is not very capable of matching the human ranked list of pairs as compared with our metrics (i.e., ECA, ICA and BCA). Specifically, the difference is significant for Test-set$_1$ where the correlation coefficient with CEA goes below $0$. This behavior of CEA makes sense because it is unsupervised and requires more knowledge to perform well. As compared with ECA, both ICA and BCA perform significantly better to match human ranking. The Spearman's score gain by BCA on Test-set$_1$ is of about $30$ ($52$) points over ECA (CEA) and the gain by ICA on Test-set$_2$ is of about $13$ ($18$) points over ECA (CEA), respectively.

In order to explain the behavior of our metrics

more clearly, we performed an evaluation of three categories of verb pairs as follows. We generated three categories of verb pairs using our metrics and CEA. We combined two test sets to show the percentage of total causal and non-causal instances of verb pairs that lie in $S_c$, $A_c$ and $S_{\neg c}$ using following procedure. If a verb pair belongs to $S_c$ and has 3 causal and 2 non-causal instances after human agreement, then these 5 instances are considered members of $S_c$. This step is performed for all verb pairs in the test set. After this the percentage of total causal and non-causal test instances are calculated for each category (see Figure 1).

Figure 1 reveals that ICA, BCA and CEA are successful in pulling more causal instances in $S_c$ as compared to ECA. But, CEA has a hard time distinguishing cause from non-cause instances because it also brings the highest percentage of non-causal instances in $S_c$. The reason is the dependence of CEA on PMI scores of pairs of verbs and arguments to make decision for causality where PMI is not good enough to distinguish a simple correlation from an asymmetric relation of causality. However, ICA and BCA work better by placing less non-causal instances in $S_c$ as compared with CEA. ICA and BCA also work better because by pulling more causal instances in $S_c$ and $A_c$, these metrics are keeping least percentage of causal instances in $S_{\neg c}$. Also, ICA and BCA bring more causal instances in $S_c$ as compared with ECA by handling training data sparseness.

Another important line of research is the construction of a classifier on top of the component of knowledge base for the classes of cause and non-cause relations. This allows us to evaluate our model in terms of standard evaluation measures - i.e., precision, recall and F-score. These measures can also be used to compare our model with supervised classifier depending merely on shallow contextual features with no information from the knowledge base. Due to space limitations, we plan to present such classifiers and evaluation in the future.

### 5.1 Analysis

In this work, we have focused on determining the predictive power of knowledge of causal associations of verb pairs to identify causality between events. Our results reveal that our best metrics (i.e., ICA and BCA) bring desired behavior of keeping least percentage of total causal instances in category $S_{\neg c}$. However, there is need to build a classifier on top of knowledge base which can help detection of non-causal instances for verb pairs lie in $S_c$ and $A_c$. Here, we state some brief details of our test set which can help building such classifier in future. An important aspect to consider is the highly skewed nature of real distribution of test set. There are only $23.69\%$ causal instances in the test set and majority of these instances (i.e., $56.7\%$) are intra-sentential instances. Therefore, a classifier should have mechanism to decide why inter-sentential instances of event pair are non-causal most of the time. For example, some inter-sentential events may not even be directly relevant at first place because they appear in different sentences. Another critical point to consider is the encoding of non-causal instances by strongly causal verb pairs. For example, we asked one of the annotators to identify strongly causal verb pairs out of 100 verb pairs of the test set. There are 22 such verb pairs determined by our annotator and each of these pairs contain $43\%$ causal instances on the average. There are many factors (e.g., temporal information, arguments of verbs) which can make an instance of strongly causal verb pair non-causal. For example, (call, respond) may encode causality only if $e_{call}$ temporally precedes $e_{respond}$ as demonstrated by the following instances.

1. Deputies spotted the truck parked at the home of the suspect's father and **called** for assistance. The Border Patrol agents and others **responded**. (CAUSE)
2. Prime Minister of Israel promptly **responded** to the widespread unrest in the West Bank and Gaza, saying that he would **call** a timeout to rethink Israel's commitment to peace talks. (NON-CAUSE)

In future, the above issues need to be addressed to improve performance for the current task.

## 6 Conclusion

In this research, we have developed a knowledge base ($KB_c$[1]) of causal associations of verb pairs to detect causality. This resource provides the causal associations in terms of three categories of verb pairs (i.e., Strongly Causal, Ambiguous and Strongly Non-Causal). We have proposed a set of knowledge rich metrics to learn these associations. Our analysis of results reveals the biases of different metrics and brings important insights into the future research directions to address the challenge of detecting causality between verbal events.

[1]We will make the resource available.

## References


Collin F. Baker, Charles J. Fillmore and John B. Lowe. 1998. The Berkeley FrameNet project. *In proceedings of COLING-ACL. Montreal, Canada*.

Brandon Beamer and Roxana Girju. 2009. Using a Bigram Event Model to Predict Causal Potential. *In proceedings of Computational Linguistics and intelligent Text Processing (CICLING), 2009*.

Steven Bethard and James H. Martin. 2008. Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations. *In proceedings of ACL-08: HLT*.

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

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

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. *In proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain.*

Quang X. Do, Yee S. Chen and Dan Roth. 2011. Minimally Supervised Event Causality Identication. *In proceedings of EMNLP-2011.*

Roxana Girju. 2003. Automatic detection of causal relations for Question Answering. *Association for Computational Linguistics ACL, Workshop on Multilingual Summarization and Question Answering Machine Learning and Beyond 2003*.

Paul Kingsbury, Martha Palmer and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. *In proceedings of HLT-2002. San Diego, California*.

Karin Kipper, Hoa T. Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. *In proceedings of AAAI-2000. Austin, TX*.

Daniel Marcu and Abdessamad Echihabi. 2001. An unsupervised approach to recognizing discourse relations. *In proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL).*

Peter Menzies. 2008. Counterfactual theories of causation. *Online Encyclopedia of Philosophy, 2008.*

George A. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography, 3(4).*

Judea Pearl. 2000. *Causality*. Cambridge University Press.

Emily Pitler, Annie Louis and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. *In proceedings of ACL-IJCNLP, 2009.*

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. *In proceedings of ACL-IJCNLP, 2009.*

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, Bonnie Webber. 2010. The penn discourse treebank 2.0. *In proceedings of LREC 2008*.

Kira Radinsky and Eric Horvitz. 2013. Mining the Web to Predict Future Events. *In proceedings of sixth ACM international conference on Web search and data mining, WSDM '13.*

Mehwish Riaz and Roxana Girju. 2010. Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision. *In proceedings of the IEEE 4th International Conference on Semantic Computing (ICSC)*.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the Dots Between News Articles. *In proceedings of Knowledge Discovery and Data Mining KDD 2010.*

Craig Silverstein, Sergey Brin, Rajeev Motwani and Jeff Ullman. 2000. Scalable Techniques for Mining Causal Structures. *Data Mining and Knowledge Discovery, 2000, 4(2−3):163−192.*

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Journal of Natural Language Engineering Volume 14 Issue 3, July 2008 Pages 369−416*.

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

James Woodward. 2008. Causation and Manipulation. *Online Encyclopedia of Philosophy, 2008.*

## Appendix A. Notations

This appendix presents the details of important notations used in this paper.

| Notation | Equation(s) | Explanation |
|---|---|---|
| $e_{v_i}$ | 6, 7, 8, 9 | Verbal event represented by the verb $v_i$ |
| $KB_c$ | – | Knowledge base of causal associations of verb pairs |
| $S_c$ | – | Strongly Causal category of verb pairs |
| $A_c$ | – | Ambiguous category of verb pairs |
| $S_{\neg c}$ | – | Strongly Non-Causal category of verb pairs |
| $m_i$ | – | Discourse marker |
| $m_c$ | 1 | Causal marker (e.g., *because*) |
| $f(I)$ | 1 | Function to select the most dependent pair from two discourse segments conjoined with causal marker |
| $CD(v_i, v_j)$ | 1, 2, 4, 6, 9 | Causal dependency of the verb pair $(v_i, v_j)$ |
| $PS_I(v_i, v_j)$ | 1, 3 | Penalization factor for the verbs of the pair $(v_i, v_j)$ with respect to their distance from the causal marker |
| $pos(v_i)$ | 3 | Distance of verb in terms of its position with respect to causal marker |
| $C(v_p)$ | 3 | Count of main verbs appearing before causal marker |
| $C(v_q)$ | 3 | Count of main verbs appearing after causal marker |
| $ECA(v_i, v_j)$ | 4 | Explicit Causal Association of the verb pair $(v_i, v_j)$ |
| $VP$ | 4, 6, 9 | Set of intra- and inter-sentential instances of a verb pair |
| $I(v_i, v_j)$ | 4, 6, 9 | Instance of the verb pair $(v_i, v_j)$ |
| $C_I$ | 4, 5, 6, 9 | Tendency of the instance I to belong to cause class than the non-cause one |
| $c$ | 5 | Cause class |
| $\neg c$ | 5 | Non-cause class |
| $C_{min}$ | – | Minimum value of $C_I$ obtained on the development set |
| $C_{max}$ | – | Maximum value of $C_I$ obtained on the development set |
| $r_C$ | 7, 8 | Class of cause role |
| $r_E$ | 7, 8 | Class of effect role |
| $ICA(v_i, v_j)$ | 6 | Implicit Causal Association of the verb pair $(v_i, v_j)$ |
| $ERM(e_{v_i}, e_{v_j})$ | 6, 7 | Events Roles Matching (ERM) determines the negative log-likelihood of events to belong to class of cause or effect role |
| $S(e_{v_i}, r_C)$ | 8 | Score of $e_{v_i}$ to belong to the class of cause role |
| $S(e_{v_j}, r_E)$ | 8 | Score of $e_{v_j}$ to belong to the class of effect role |
| $P(f_k\|.)$ | 5, 8 | Probability of feature $f_k$ given some class |
| $BCA(v_i, v_j)$ | 9 | Boosted Causal Association of the verb pair $(v_i, v_j)$ |

Table 2: Details of notations.