2010 IEEE Fourth International Conference on Semantic Computing

Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision

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

***Abstract*—Contingency discourse relations play an important role in natural language understanding. In this paper we propose an unsupervised learning model to automatically identify contingency relationships between scenario-specific events in web news articles (on the Iraq war and on hurricane Katrina). The model generates ranked contingency relationships by identifying appropriate candidate event pairs for each scenario of a particular domain. Scenario-specific events, contributing towards the same objectives in a domain, are likely to be dependent on each other, and thus form good candidates for contingency relationships. In order to evaluate the ranked contingency relationships, we rely on the manipulation theory of causation and a comparison of precision-recall performance curves. We also perform various tests which bring insights into how people perceive causality. For example, our findings show that the larger the distance between two events, the more likely it becomes for the annotators to identify them as non-causal.**

***Keywords*-causality; contingency; scenario; topics;**

I. Introduction

Unlike computers, people are very good at perceiving and inferring the causal, reason, purpose and explanation relationships between events in a discourse context. Detecting such relations helps them make sense of the constantly changing flow of events in their daily activities and interactions. Thus, causal reasoning enables people to find meaningful order in events, which in turn helps them plan and even predict the future [13].

In linguistics, these relations form a class known as *contingency discourse relations* (cause-consequence, argument-claim, instrument-goal, purpose and reason/explanation) which are different from additive relations (list, opposition, exception, enumeration, temporal, and concession) [18], [14]. Since they are very related semantically, contingency relations can be identified as the class of *causal relations* in a broader sense [7]. Examples of such relations are:

(1) Teachers *are not going* to work today because they *are on strike* (explanation).

(2) The company *makes* and *repairs* cell phones (temporal).

Thus, two events are contingent if the occurrence of one event enables the occurrence of the other – i.e., they are linked by one of the contingency relations in the dicourse context.

Identifying contingency relations is very important for a number of natural language understanding tasks, such as textual entailment and explanation question answering [6]. In the PASCAL Recognizing Textual Entailment (RTE) Challenge [5], for example, a computer system is presented with a series of yes-no questions concerning whether one English sentence entails another. The semantic entailment relation is defined in a broad sense, whether the meaning of a given text snippet (text T) logically entails that of another (hypothesis H) or whether they paraphrase each other. For example, "Google files for its long awaited IPO" entails that "Google goes public". Explanation question answering deals with questions asking for particular explanations, like "Why did the Dallas-based Southwest airlines cancel more than 250 flights last week?".

In what concerns causality, in natural language processing the focus has been mainly on causal knowledge extraction. Most of the proposed approaches depend either on small lists of predefined linguistic patterns employed in supervised learning models [6], [4], or on special data sets [21], [1].

In this paper we propose a novel unsupervised approach to automatically identify contingency relationships between events in web news articles both within (intra-sentential) and between sentences (inter-sentential) without relying on a deep processing of contextual information. Our approach focuses on a simple context which consists of two events, which if contingent, represent the Cause (independent) event (**a**) and the Effect (dependent) event (**b**), such that **a** $\rightarrow$ **b** (i.e., **a** and **b** encode a contingency relation either directly or indirectly). Here events are "$[Subj_e]$ $verb_e$ $[Obj_e]$" instances, where the subject or the object can be missing.

Our contribution is as follows:

1) Our approach is totally unsupervised and depends neither on predefined linguistic patterns nor on special datasets where the events are already temporally ordered. Approaches relying on linguistic patterns aim at higher precision but have low recall. Our model does not have this limitation. Moreover, this approach saves us the trouble of annotating large text collections, a prerequisite for training supervised models.

978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.19

2) We introduce two novel measures – Effect-Control-Dependency (ECD) and Effect-Control-Ratio (ECR) – which not only find dependencies but also identify the Cause and the Effect roles without relying on hard-to-build temporal classifiers.
3) This is a flexible and feasible approach which brings new insights into how much a knowledge-poor, statistical approach can help in the automatic identification of contingency relationships in text.
4) We also perform various tests which bring insights into the cognitive aspects of this challenging task - i.e., how people perceive such broader causal relationships in context. Specifically, our findings show that the larger the distance between two events, the more likely it becomes for the annotators to identify them as non-causal.

We test our model on two domain-specific data sets collected from the web: one on the Iraq war and one on hurricane Katrina. For each collection, the model identifies scenario-specific event pairs that are potential candidates for contingency relations. We rely here on the hypothesis that natural language events contributing to one particular scenario tend to be strongly dependent on each other, and thus make good candidates for this task (Examples of such contingent/causal event pairs are shown in Table I).

The paper is structured as follows. In the next section we present relevant previous work, while in Section 3 we describe the model and introduce its components. The system evaluation is presented in Section 4, followed by discussions and conclusion in Section 5.

Collection:	Hurricane Katrina
Scenario:	Hurricane Katrina disaster and damage.
Example:	Katrina {hit} Florida late last week. Since Friday, Dallas-based Southwest airlines {canceled} more than 250 flights.
Cont. Rel:	"Katrina {hit} Florida"→"Dallas-based Southwest airlines {canceled} more than 250 flights"
Type:	Inter-sentential contingency relationship.
Collection:	Iraq War
Scenario:	US accusations and the UN inspections.
Example:	Bush {criticized} UN for {being ineffective}.
Cont. Rel:	"UN {being ineffective}"→"Bush {criticized} UN"
Type:	Intra-sentential contingency relationship.

Table I
EXAMPLES OF CONTINGENCY RELATIONS. "CONT. REL" REFERS TO CONTINGENCY RELATIONSHIP BETWEEN EVENTS.

II. PREVIOUS WORK

Causality has been a popular subject of study in philosophy, logic, linguistics, data-mining, and economics [8], [9], [19], [24], [15]. In philosophy and logic, for example, one of the most influential theories is the manipulation theory of causality [24]. This theory identifies two necessary conditions for causality: (1) the temporal precedence of the Cause (**a**) over the Effect (**b**), and (2) their causal dependency (i.e., the effect can occur only with the Cause). Since this theory was proven to provide an easy and objective notion of causality on some language tasks [1], we also employ it here for annotation and evaluation purposes.

In natural language processing, the focus has been mainly on the analysis of causal lexico-syntactic patterns (e.g. "*mosquitoes cause malaria*" – "NP-Cause verb NP-Effect"), which are employed most of the time in supervised learning models [6], [4]. The patterns are identified either manually or semi-automatically and most of them are ambiguous, making the system difficult to port to different domains. Girju (2003)'s pattern-based approach, for example, achieves a precision of 73.91%.

Causality can also be expressed implicitly with no causal markers, especially at the discourse level. One such example is "Katrina *hit* Florida late last week. Since Friday, Dallas-based Southwest airlines *canceled* more than 250 flights." In such contexts, the causal discourse relations need to be inferred which is quite a challenging task [8], [20], [16], [17]. In this paper we focus on both explicit and implicit contingency relationships at the sentence level, but also between sentences. We try to approximate the context by identifying scenario-specific strongly related events which can form good candidates for causal relationships in a domain.

Other approaches [1], [21] do not rely on cue phrases, but make use of statistical methods applied on special data sets. Beamer & Girju (2009), for example use a statistical measure, Causal Potential, on a text corpus of screen plays where the verb events are already temporally ordered. They report a good correlation of the results with the human judgments (a Spearman's rank correlation of 0.497). Sun et al. (2007) have proposed another measure, the Event Causality Test (ECT), to discover causal relationships between events in search queries extracted from temporal query logs. The model achieves an accumulated precision from 32% to 21% for instances ranked 1 to 99.

In this paper we introduce two novel statistical measures (ECD and ECR) to capture contingency relations rather than use general dependency measures such as Pearson's correlation coefficient and Chi-square. These measures have been employed in previous data mining approaches to learn causal chains of events in structured datasets (e.g., census data) [19].

III. APPROACH

In this section we propose a three-layer unsupervised statistical system (shown in Figure 1) which automatically identifies contingency information in domain-specific text collections. These layers are (1) Identifying Topic-Specific

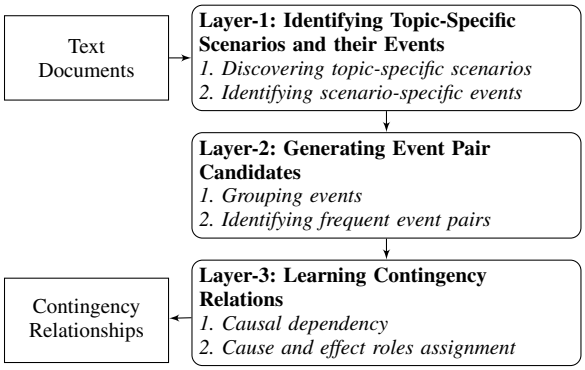

Figure 1. The Contingency Learning System Architecture.

Scenarios and their Events; (2) Generating event pair candidates; and (3) Learning Contingency Relations. The data and system components are presented in the next subsections.

A. Data

In this research we employ two text collections to identify contingency relationships between events. In particular, we crawled a set of 447 news articles on hurricane Katrina[1] (189,840 word-tokens and 14,996 word-types) and 556 news articles on the Iraq war[2] (304,481 word-tokens and 20,629 word-types) from various news archives websites.

B. Identifying Topic-Specific Scenarios and their Events

This module clusters the input text units (i.e., sentences) according to their probability distributions into topic-specific scenarios. The idea is that a single text document can contain multiple topics, and thus can identify multiple scenarios (e.g., a news article about the Iraq war can refer to the "*US accusations and the UN inspection*", "*post-war developments*", etc). Our intuition is that the events describing a particular topic-specific scenario tend to be dependent on each other. For example, events such as **(a)** "*Iraq might have developed chemical weapons*" and **(b)** "*the UN team inspecting Iraqi scientists*" belong to the scenario "*US accusations and the UN inspection*" where the occurrence of event **b** is dependent on the occurrence of event **a**.

In order to learn strongly dependent scenario events, we rely on topic modeling [2], [12] assuming that semantically dependent events will have higher generative probabilities conditioned on a particular topic-specific scenario. We employ an advanced topic model, the Pachinko Allocation Topic Model (PAM) [12] for discovering topic-specific scenario events. This model captures not only correlations between words to determine topics but also identifies relationships between topics.

[1]http://websearch.archive.org/katrina/list.html/

[2]http://www.comw.org/warreport/

Since statistical models such as PAM require large data sets as input, we run it on words rather than events because they are less frequent than simple words. We then extract the events corresponding to each identified scenario as explained next.

1) Discovering Topic-Specific Scenarios: Using as input a text collection, PAM generates an n-level Directed Acyclic Graph (DAG) – a fully connected tree structure, with topics at intermediate levels. The leaves are clusters of words. This structure helps finding correlations between topics as well as correlations between words given a topic. Here we are interested only in the words corresponding to topics (leaves) and leave the correlations between topics for future work. The topics learned by PAM represent topic-specific scenarios. Table II shows the 10 most representative words for each of the three topic-specific scenarios identified for the Iraq war collection.

Topic-Specific Scenarios		
Topic-1	Topic-2	Topic-3
America	Officials	Free
Iran	Intelligence	Enemy
Control	Mass	Marine
Help	U.N	Happen
Freedom	Resolution	Police
Occupation	Question	Israel
Economic	Disarm	Corps
Democratic	Accuse	Troops
Elections	Uranium	Tactics
Opportunity	Inspection	Air

Table II
THE THREE TOPIC-SPECIFIC SCENARIOS FOR THE IRAQ WAR COLLECTION.

2) Identifying Scenario-specific Events: We identify first the sentences which correspond to the scenario clusters and then from these sentences, we recover the events contributing to a particular scenario. We represent each cluster as a vector $\vec{v}$ of words. Each word has a weight p_w (the word's assignment probability given a scenario cluster $\vec{v}$ discovered by PAM). Each sentence $\vec{s}$ (the weight for each word in $\vec{s}$ is its probability of occurrence in sentence) is assigned to a cluster $\vec{v}$ based on the standard normalized cosine similarity score (N is the vocabulary size).

$$Cosine - Sim(\vec{s}, \vec{v}) = \frac{\vec{s} \cdot \vec{v}}{\sqrt{\Sigma_{i=1}^{N} s_i^2} \sqrt{\Sigma_{i=1}^{N} v_i^2}}.$$

The model assigns sentence $\vec{s}$ to cluster $\vec{v}_i$ (i is the total number of scenario clusters) with which it has the maximum cosine measure. In case of a tie, the model assigns the sentence to all clusters with the same maximum cosine measure.

From these scenario-specific sentences we identify their events (i.e., "$[Subj_e]$ $verb_e$ $[Obj_e]$" instances). We rely here on a semantic role labeler, SWiRL [23], to identify the subject (A_0) and the object (A_1). Table I shows some

of the scenario-specific sentences and their events from the hurricane Katrina and the Iraq war collections.

C. Generating Event Pair Candidates

This module generates event pair candidates following the steps described below.

1) Grouping Events: Similar events identified for each scenario by the previous module need to be grouped together. For example, the instances *"The UN council suspects Iraq"* and *"The UN Security Council suspects Iraq"* are referring to the same event in the scenario "*US accusations and the UN inspection*". The grouping is done using the following basic clustering procedure based on the naïve lexical similarity between events:

Procedure: Grouping events
Input: Events e_1,e_2,...,e_n (e_i = <[$Subj_{e_i}$] $verb_{e_i}$ [Obj_{e_i}]>)
Output: Event Groups G_1, G_2,..., G_m ($m \leq n$)

[1.] Initially place every event e_i into its own group (G_i = $\{e_i\}$)
[2.] For each event $e_i \in G_i$:
  For each $G_{k \neq i}$ where $G_{k.Lemma(verb)}$= $G_{i.Lemma(verb)}$:
    Calculate average cosine-similarity(e_i, G_k) (for each event e_j in G_k find cosine-similarity (e_i, e_j) and take the average).
  Add event e_i to G_k s.t. the average cosine-similarity(e_i, G_k) is maximum (also above some threshold value) and discard original G_i. In case of tie put event e_i randomly in any of the groups on tie.
[3.] Return the resulting event groups G_1, G_2,...,G_m.

2) Identifying Frequent Event Pairs: Once similar events are grouped together, frequent candidate event group pairs (G_i, G_j) are generated based on the FP-Growth algorithm [10] with minimum support of 5. These are the event group pairs which appear in at least 5 documents (i.e., news articles).

FP-Growth is a very popular algorithm in data mining [10], usually used to generate frequent combinations of items to learn associations between them. Frequent combinations of items are those appearing at least n number of times in a database where n is *minimum support*. For example, transaction database records might contain information about what items people are frequently purchasing together . Transaction records may contain frequent combination of $(laptop, harddrive)$ with minimum support n – i.e., n records contain this combination. This shows that people tend to buy hard drives when they frequently purchase laptops or vice-versa depending on which conditional probabilities $P(harddrive|laptop)$ or $P(laptop|harddrive)$ is higher. In order to identify such frequent combinations of items from a database by using the FP-Growth algorithm, one has to specify the minimum support n. Silverstein et al (2000) have also used this algorithm to mine frequent combinations of items as a preliminary step in identifying causal associations between census variables and text words. Here we are using the FP-Growth algorithm to generate frequent event group pairs which appear in at least 5 documents (i.e., news articles). In our approach, each document D_i contains a set of events $e_1, e_2, .., e_n$. Since some of these events are similar and thus, belong to the same event group, we generalize the representation by replacing the events with the groups they are part of, as shown in the Grouping events procedure. Next, we apply FP-Growth which generates event group pairs (G_i, G_j) with minimum support of 5. One such example of frequent event group pair instances is ("*US suspects Iraq*", "*Iraq develops chemical weapons*") occurring at least 5 times (minimum support) in the dataset assigned to the scenario "*US accusations and the UN inspection*".

D. Learning Contingency Relations

This module has two objectives: (1) to determine if the events of a frequent event pair (**a**,**b**) are contingent (i.e., encode a contingency relation), and if yes, (2) to assign the Cause and the Effect roles to such events. These steps are presented in detail below.

1) Causal Dependency: In order to identify if two events are contingent, we propose a novel statistical measure (*Effect-Control-Dependency* – ECD), to measure contingency between two variables **a** and **b**. For this, we need to take into account the following issues:

(1) One important contingency condition is the temporal precedence of the causing event over the effect. However, our data is not temporally ordered. Instead of relying on temporal classifiers which are hard to build [3], we also introduce here a statistical measure Effect-Control-Ratio (ECR) derived from ECD to identify the Cause and the Effect roles once causality is decided between them.

(2) The Causing event can appear independently with other events, while the Effect event is expected to have a high likelihood of occurrence in the presence of the causing event (this condition is similar to the causality notion proposed by Suppes, 1970 [22]).

(3) Since we work with domain- or topic-specific collections, we assume that in a scenario specific to a domain or topic, highly dependent frequent event pairs should be given priority over less frequent ones because they are more important than less frequent pairs.

(4) The Contingent events can be in direct or indirect relationships. We hypothesize here that the larger the distance between the events, the lower their degree of dependency.

The proposed measure of contingency (ECD(a,b)) is defined as follows:

$$max(\frac{P(a,b)}{P(b)-P(a,b)+\gamma} * \frac{P(a,b)}{\max_t P(a,b_t)-P(a,b)+\gamma}, \frac{P(a,b)}{P(a)-P(a,b)+\gamma} * \frac{P(a,b)}{\max_t P(a_t,b)-P(a,b)+\gamma})$$

This measure computes the maximum contingency score based on which event is the Cause and which is the Effect.

ECD is an improvement over Point-Wise Mutual Information (PMI), a measure frequently used to capture dependencies between variables. Two events **a** and **b** are strongly dependent when at least one of them appears more frequently in the presence of the other than alone (i.e., $P(b) - P(a,b) < P(a,b)$ or $P(a) - P(a,b) < P(a,b)$, or both). Unlike PMI, ECD captures not only the dependency between two events, but also its importance (relevancy) in a given scenario. PMI has the disadvantage of giving higher weights to strongly dependent but rare events. This problem is addressed by ECD which gives higher scores to more frequent pairs.

The first fraction of the first argument assumes that if **b** is the Effect event then its probability given **a** is greater than when **a** does not occur. Thus, if this condition is true then the value of the fraction will be greater than when this condition is false. This fraction determines the dependency between events and also gives higher score to more frequent (important) pairs than less frequent (unimportant) pairs. The second fraction indicates that **a** can be the cause of other events as well. However, if it is very strongly correlated with the Effect event **b**, then it will appear most often with **b** than with any other event. If this condition is true then the second fraction will have a higher score than when it is false. We use γ (a small value > 0, say 0.01) in both fractions to avoid ∞ score for the two pairs when they are not equally important in a particular scenario.

An example of such event pairs **(a,b)** is given below:

Pair-1 = ("US accused Iraq of developing chemical weapons" [92], "UN inspected Iraqi scientists" [51]) [50]

Pair-2 = ("UN Security Council held an emergency session" [55], "Security Council closed emergency session" [6]) [5]

Here both pairs are identified as causal (**a** $\rightarrow$ **b**) by the manipulation theory – i.e., if the US had not accused Iraq of developing chemical weapons, one can necessarily infer that UN would not have inspected Iraqi scientists regarding this. The numbers in brackets indicate the frequency of events and pairs. It can be noticed that the events are strongly dependent since in each example **b** has a higher probability of occurrence with **a** than alone. Here PMI gives a higher score to Pair-2 because it is a less frequent pair with lower frequency events. The ECD test gives Pair-1 a higher score than to Pair-2 because the denominator of the first fraction (P(b) - P(a,b) + γ) in the ECD's first argument is the same for both pairs. Thus, P(a,b) will decide which dependent pair is more important.

In order to discover more direct and stronger causal relationships, we also penalize pairs by multiplying ECD with the Leacock and Chodorow, (1998) distance penalization measure used to find similarity between concepts:

$$ECD(a,b) = ECD(a,b) \times -\log \frac{dis(a,b)}{2 * maxDistance} \quad (1)$$

Here dis(**a**,**b**) is the average (median) distance between events **a** and **b** in a particular scenario. If the two events appear in same sentence, then the distance is 1.0; if they appear in consecutive sentences, then distance is 2.0, and so on. MaxDistance is the maximum median distance between any two events in a scenario.

We rank contingent pairs according to their scores and evaluate contingency relationships ranked by both PMI and ECD using the interpolated precision-recall curve (explained in Section 4).

2) Cause and Effect Roles Assignment: We introduce a new metric, ECR (Effect-Control-Ratio), to identify the Cause and Effect roles.

$$ECR(a,b) = \frac{\frac{P(a,b)}{P(b) - P(a,b)} * \frac{P(a,b)}{\max_t P(a,b_t)}}{\frac{P(a,b)}{P(a) - P(a,b)} * \frac{P(a,b)}{\max_t P(a_t,b)}} \quad (2)$$

Similar to ECD, the first fractions of the numerator and the denominator capture the dependency of the Effect with respect to the Causing event. Since we need to capture only the dependency of the Effect event on the Causing event and not their importance (tackled by ECD), we have removed γ from both fractions and have tried to reduce the scale of the second fraction such that the decision is made only by the first fraction. The first fractions of the numerator and denomenator have values within the range (0,∞], while the second fractions of the numerator and denominator have values in (0,1]. When the first fraction can not identify the Cause and the Effect (i.e., its value for the numerator and denominator are the same or very close), the decision is made by the second fraction which considers how strongly the Causing Event is related to the Effect as compared with other events. The decision about the causal roles is as follows:

a) Predict **a** $\rightarrow$ **b**, if ECR(a,b) > 1.0

b) Predict **b** $\rightarrow$ **a**, if ECR(a,b) < 1.0

In this research we do not deal with the case when ECR is 1.0, since we need a deeper temporal or semantic analysis for such pairs to decide their causal roles. However, this does not affect our predictions much since only 1% of the event pairs in each scenario on average had a ratio of 1.0.

IV. Experiments and Evaluation

In this section we present the experiments and their evaluation. The parameters set manually in our system for each dataset are:

(1) PAM Dirichlet parameters same as used by Li and McCallum, (2006) including super-topics=2, and sub-topics=3

for DAG tree (2) minimum support of 5 for the FP-Growth algorithm (3) $\gamma = 0.01$ for ECD.

During the experiments we observed that a 4-level DAG with 3 subtopics performs well on the scenario learning task on both data sets. This number of subtopics is reasonable since we noticed that any domain can have between 3 to 10 major scenarios on average. A larger number of subtopics generates noisy scenarios for our domains of the Iraq war and Hurricane Katrina.

A. Evaluating the Scenario Generation Task

We evaluated each of the three topic-specific scenario clusters obtained on each text collection through blind judgments of cluster quality. Two human annotators were presented with each of the scenarios' top-50 ranked words to label them for the "relatedness" task. The relatedness task for a scenario's top words list requires annotators to label a word as "YES" if it is semantically similar to other words in that scenario cluster, otherwise "NO". In this annotation task, the annotators were asked to judge the semantic coherence of each scenario cluster as a whole (i.e., are the words in the clusters semantically related, identifying a particular scenario?). The evaluation using the relatedness task shows that for hurricane Katrina, clusters 1 (66% related words) and 3 (65% related words) are less noisy as compared with cluster 2 (57% related words) (Table III). Similarly, clusters 1 and 2 in the Iraq war collection were good. At the end of this process, the annotators discussed and agreed on appropriate scenario labels: Katrina – (C1) *"Hurricane Katrina disaster and damage*", (C2) "*Global Warming and climate change issues*", (C3) "*Rescue efforts and criticism of government rescue plans*"; and the Iraq war – (C1) "*War effects-economic progress in Iraq and side effects on the world's economy*", (C2) "*US accusations and the UN inspection*, (C3) "*Pre-war: War strategies and planning*".

Test	**Data**	**C1**	**C2**	**C3**
Relatedness	Katrina	66%	57%	65%
	Iraq	90%	83%	39.5%
Annotator-Agreement	Katrina	86%	94%	92%
	Iraq	80%	96%	86%

Table III
EVALUATION OF WORD RELATEDNESS AND INTER-ANNOTATOR AGREEMENT FOR ALL THREE SCENARIO CLUSTERS IN BOTH COLLECTIONS.

B. Evaluating the Contingency Relations Detection Task

Due to time constraints, we considered here only the ranked list of contingency relationships for the best cluster of each collection according to the relatedness test (Table III). Before evaluation, we performed an interesting cognitive study of how people identify contingency discourse relations. This is presented in detail next.

1) Human Annotation: Since contingency (causality, in a broader sense) is to some extent a matter of perception, we performed a study which looks at how people perceive causal and non-causal relationships between two events at different distances. For this, we selected the top-two clusters for hurricane Katrina (i.e., C1 and C3) and the Iraq war (i.e., C1 and C2) as given by the relatedness task. Then we randomly selected 80 test examples of event pairs (**a**,**b**) from each cluster. Each of the four test-sets were selected in such a way that the events **a** and **b** in a pair can be at any distance from 1 to 4 (i.e., 1 to 4 other events can separate them) and the event pairs were evenly distributed over these distances (i.e., 20 examples for each distance: 20 x 4 = 80 test-examples).

We presented these random test examples to two judges who annotated them as contingent or not-contingent according to the manipulation theory adopted from [1] (i.e., keeping constant as many other states of affairs of the world in the given text context as possible, modifying event **a** entails predictably modifying event **b**). Table IV shows that for all test-sets more contingency examples were observed at distance 1 than non-contingency and that for some test-sets (HK:C1 and IW:C2) the percentage of contingency examples decreases with the increase in distance on the human-agreed annotated test-sets. For IW:C2, the annotators agreed on only 5% contingency examples at distances 3 and 4. For three test-sets (HK:C3, IW:C1 and IW:C2) the annotators identified at least 50% non-contingency and at most 25% contingency examples for distances greater than 1 on human-agreed data. This leads to the conclusion that the smaller the distance between two events, the more likely it is for people to identify them as contingent and vice versa. This means that events appearing in the same sentences (i.e., distance 1) are more likely to be perceived in a causal relationship.

Test-set	$Dist_1$		$Dist_2$		$Dist_3$		$Dist_4$		Causal relations
	Y	N	Y	N	Y	N	Y	N	
HK:C1	55	40	40	30	35	35	25	30	53.4
HK:C3	30	20	15	60	25	55	25	50	33.9
IW:C1	40	25	10	60	10	60	15	50	27.7
IW:C2	35	25	10	65	5	50	5	55	22.0

Table IV
THE PERCENTAGE OF CONTINGENT AND NON-CONTINGENT EXAMPLES IN TEST-SETS OF HURRICANE KATRINA (HK:C1 AND HK:C3) AND IRAQ WAR (IW:C1 AND IW:C2) WITH RESPECT TO DISTANCES $Dist_i$ BETWEEN EVENTS (i RANGES FROM 1 TO 4). "Y" AND "N" REFER TO CONTINGENT AND NOT-CONTINGENT EXAMPLES, RESPECTIVELY AFTER HUMAN AGREEMENT (E.G. HK:C1 HAS 55% "Y" AND 40% "N" EXAMPLES WITH REMAINING 5% DISAGREEMENT FOR $Dist_1$) . THE LAST COLUMN SHOWS THE PERCENTAGE OF CONTINGENCY RELATIONS IN EACH TEST-SET AFTER ANNOTATORS AGREEMENT (E.G. HK:C1 HAS 53.4% CONTINGENCY EXAMPLES AFTER EXCLUDING EXAMPLES ON WHICH ANNOTATORS DISAGREED).

We can also relate the scenario clusters' quality to the

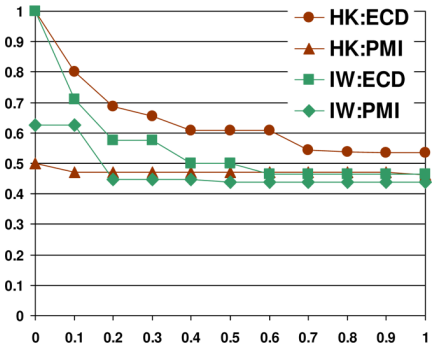

Figure 2. Interpolated Precision-Recall Curve.

percentages of contingency and non-contingency pairs identified on the human-agreed test sets. It can be easily noticed that the occurrence of contingency pairs is higher in the scenarios HK:C1 and IW:C1 ranked best by the relatedness task than non-contingency pairs (the last column of Table IV shows the percentage of contingency relations). Moreover, for the Iraq war test-sets, the annotators labeled more examples as non-contingency (27.7% and 22.0% causal relations in IW:C1 and IW:C2, respectively). A quick analysis of the annotators' comments on the disagreed examples shows that they found the Iraq war instances much more difficult to annotate. In addition, there were cases where the annotators had to have some advanced domain knowledge in order to annotate the examples. For example, for the Iraq war dataset, the annotators needed knowledge about the main participants and the government policies in each country involved. Moreover, for those examples where the contingency relation was implicit (i.e., no discourse marker), the annotators had to read the entire document before deciding on the relationship between the events.

2) Evaluation of Contingency Measures: Based on the observations obtained from the annotation experiments, we decided to evaluate the system only on distance 1 and 2 examples. Thus, the two judges also annotated the following collected testsets: (1) for Hurricane Katrina: the top 100 ranked relationships for cluster 1 using ECD (HK:ECD) and PMI (HK:PMI); and (2) for the Iraq war: the top 100 ranked relationships for cluster 1 using ECD (IW:ECD) and PMI (IW:PMI). The inter-annotator agreement on the four test-sets are shown in Table V. In order to judge the performance of the contingency detection task, we computed the interpolated precision at 11 recall levels (Figure 2), method of evaluation used frequently in Information Retrieval. The interpolated precision at recall level r (Interpolated-Precision(r)) is given by $\max_{i \geq r} Precision(i)$. The precision-recall curve shows the system's performance at different recall levels for the ECD and PMI measures. The idea is to choose the measure which has good precision at all recall levels. The ECD outperforms PMI by achieving maximum interpolated-precisions from 1 to 0.52 at 11 recall levels for the Katrina test-set (Figure 2). However, the PMI performance on the Katrina test-set is rather constant. It remains lower for almost all recall levels. At 0.60 recall level, ECD balances the precision and recall.

For the Iraq war test-sets ECD is also better than PMI up to recall 0.60 after which the measures converge. The Figure also shows that the performance on hurricane Katrina is higher than the one on the Iraq war, due to its complexity. PMI also remains quite stable and its precison does not go below 40%.

Considering that we do not perform any context analysis, these measures work quite reasonably on the contingency detection task. ECR shows reasonable performance for the contingency roles assignment task with the best accuracy of 72.5% achieved on IW:ECD test-set (Table V shows the Roles accuracy and the inter-annotator agreement) without using any temporal classifier. A higher inter-annotator agreement on the contingency annotation task (Table V) is achieved on all four test-sets because now annotators label distance 1 and 2 examples only.

Task	HK:ECD	HK:PMI	IW:ECD	IW:PMI
RA	63.4%	71.0%	72.5%	66.6%
CA-agreement	98%	93%	90%	85%
RA-agreement	100%	97%	95%	94%

Table V
ROLES ACCURACY (RA) FOR ALL TEST SETS. ROLES ACCURACY = # OF CORRECT ROLES PREDICTED/# OF CONTINGENCY EXAMPLES. CA-AGREEMENT AND RA-AGREEMENT SHOW THE INTER-ANNOTATOR AGREEMENT ON THE CONTINGENCY ANNOTATION AND CONTINGENCY ROLES ANNOTATION TASKS.

V. DISCUSSION AND CONCLUSIONS

In this paper we presented a system which makes contingency predictions based on two novel knowledge- and context-poor statistical measures. Even though our system obtains good results, it has also a number of limitations. For example, a small number of the identified contingency dependencies indicate other discourse relationships (e.g., elaboration or similar events). Consider the following example:

"*A decade of tourism development in Mississippi was wiped out in a few hours as the full extent of <Hurricane Katrina's destructive force emerged>. A casino barge sits among residential homes north of highway 90, bottom, in Biloxi, Miss., Tuesday, Aug. 30, 2005 after <hurricane Katrina passed> through the area.*"

Here, the events *<Hurricane Katrina's destructive force {emerged}>* and *<hurricane Katrina {passed}>* co-occur very often and are thus, strongly dependent. However this these events are similar rather than causally related.

Our approach based on scenario-specific events allows us to analyze and identify contingency relationships between strongly related events. We noticed that scenario-specific events tend to be strongly related and our approach captures the information flow through sequences of scenario-specific events within a scenario. This approach generates suitable event pair candidates for the contingency detection task which reduces the chance for noisy relationships.

Unlike previous works [16], [17], our approach was to build a knowledge- and context-poor, yet fast and easily domain portable model which constitutes an informative baseline for this task. The rationale was to narrow down the search space of causal information and to see how much performance such a system can achieve without any syntatic, semantic, or discourse analysis. Moreover, the datasets thus obtained would allow researchers to get new insights into the causality problem. Identifying causal relationships at the discourse level is a very challenging task since it requires a deep discourse and temporal analysis of the text, as well as extra-linguistic information such as world knowledge. We believe that our baseline model will provide a better framework for any context- and knowledge-rich system which employes such features. Moreover, comprehending discourse relations is also a matter of perception. Thus, such datasets would be very important for a deeper study of inter-annotator agreement.

REFERENCES

[1] B. Beamer and R. Girju, *Using a Bigram Event Model to Predict Causal Potential* Computational Linguistics and intelligent Text Processing (CICLING), 2009.

[2] D. M. Blei, A. N. Ng, and M. I. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research, 2003, 3:993-1022.

[3] N. Chambers and D. Jurafsky, *Jointly combining implicit constraints improves temporal ordering*, Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Morristown, NJ, 2008.

[4] D. Chang and K. Choi, *Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities*, Information Processing and Management, 2006, 24(3):662-678.

[5] I. Dagan and O. Glickman, *Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability*, Learning Methods for Text Understanding and Mining, Grenoble, France, 2004.

[6] R. Girju, *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.

[7] R. Girju and K. Woods, *Exploring Contingency Discourse Relations*, The 6th International Workshop on Computational Semantics(IWCS-6), Harry Bunt (ed.), The Netherlands, 2005.

[8] A. C. Graesser, K. K. Millis, and R. A. Zwaan, *Discourse Comprehension* Annual Review of Psychology, 1997, 48: 163-189.

[9] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometricia, 1969, 37:424-438.

[10] J. Han, J. Pei, Y. Yin, and R. Mao *Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach*, Data Mining and Knowledge Discovery, 2004, 8(1):53-87.

[11] C. Leacock and M. Chodorow, *Combining Local Context and WordNet Sense Similarity for Word Sense Identification*, WordNet: An Electronic Lexical Database (Language, Speech, and Communication), 1998.

[12] W. Lei and A. McCallum, *Pachinko allocation: DAG-structured mixture models of topic correlations*, International Conference on Machine Learning, ICML, Pittsburgh, Pennsylvania, 2006.

[13] J. Magliano and B. Pillow, *Learning: Causal Reasoning*, The Gale Group http://www.education.com/reference/article/learning-causal-reasoning/, 2006-2009.

[14] W. C. Mann and S. A. Thompson, *Rhetorical structure theory: Toward a functional theory of text organization*, Text, 1988, 8(3):243281.

[15] P. Menzies, *Counterfactual Theories of Causation*, Online Encyclopedia of Philosophy, 2008.

[16] E. Pitler, A. Louis, and A. Nenkova, *Automatic Sense Prediction for Implicit Discourse Relations in Text*, ACL-IJCNLP, 2009.

[17] E. Pitler and A. Nenkova, *Using Syntax to Disambiguate Explicit Discourse Connectives in Text*, ACL-IJCNLP, 2009.

[18] T. J. M. Sanders, W. P. M. S. Spooren, L. G. M. Noordman, *Toward a taxonomy of coherence relations. Discourse Processes*, 1992, 15(1): 1-35.

[19] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, *Scalable Techniques for Mining Causal Structures*, Data Mining and Knowledge Discovery, 2000, 4(2-3):163-192.

[20] C. Sporleder and A. Lascarides, *Using automatically labeled examples to classify rhetorical relations: An assessment*, Natural Language Engineering, 2008, 14(3): 369-416.

[21] Y. Sun, K. Xie, N. Liu, S. Yan, B. Zhang, and Z. Chen, *Causal relation of queries from temporal logs*, International Conference on World Wide Web WWW, Banff, Alberta, Canada, 2007.

[22] P. Suppes, *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing Company, 1970.

[23] M. Surdeanu and J. Turmo, *Semantic Role Labeling Using Complete Syntactic Analysis*, CoNLL, Shared Task, 2005.

[24] J. Woodward, *Causation and Manipulation*, Online Encyclopedia of Philosophy, 2008.