AN UNSUPERVISED APPROACH TO IDENTIFYING CAUSAL RELATIONS FROM RELEVANT SCENARIOS

BY

MEHWISH RIAZ

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Adviser:

Professor Corina Roxana Girju

ABSTRACT

Semantic relations between various text units play an important role in natural language understanding, as key elements of text coherence. The automatic identification of these semantic relationships is very important for many language processing applications. One of the most pervasive yet very challenging semantic relations is cause-effect. In this thesis, an unsupervised approach to learning both direct and indirect cause-effect relationships between inter- and intra-sentential events in web news articles is proposed. Causal relationships are leaned and tested on two large text datasets collected by crawling the web: one on the *Hurricane Katrina*, and one on *Iraq War*. The text collections thus obtained are further automatically split into clusters of connected events using advanced topic models. Our hypothesis is that events contributing to one particular scenario tend to be strongly correlated, and thus make good candidates for the causal information identification task. Such relationships are identified by generating appropriate candidate event pairs. Moreover, this system identifies both the Cause and Effect roles in a relationship using a novel metric, the Effect-Control-ratio. In order to evaluate the system, we relied on the manipulation theory of causality.

To my Parents and Husband

ACKNOWLEDGMENTS

I would like to convey my sincere appreciation to my advisor Corina Roxana Girju for her valuable guidance and insightful suggestions throughout my research work. Her commendable supervision helped me in working towards the solutions for challenging research issues. This work would not have been accomplished without her constant support and encouragement in developing and polishing research ideas.

I want to express thanks to all my Semantic Frontiers Research group fellows, working in supervision of Roxana Girju. Our group's research discussions have been excellent source of inspiration for me to come up with creative thoughts during this work.

Lastly, my cordial appreciation is for my parents, brother, sister and husband for their encouragement and support throughout my graduate studies.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

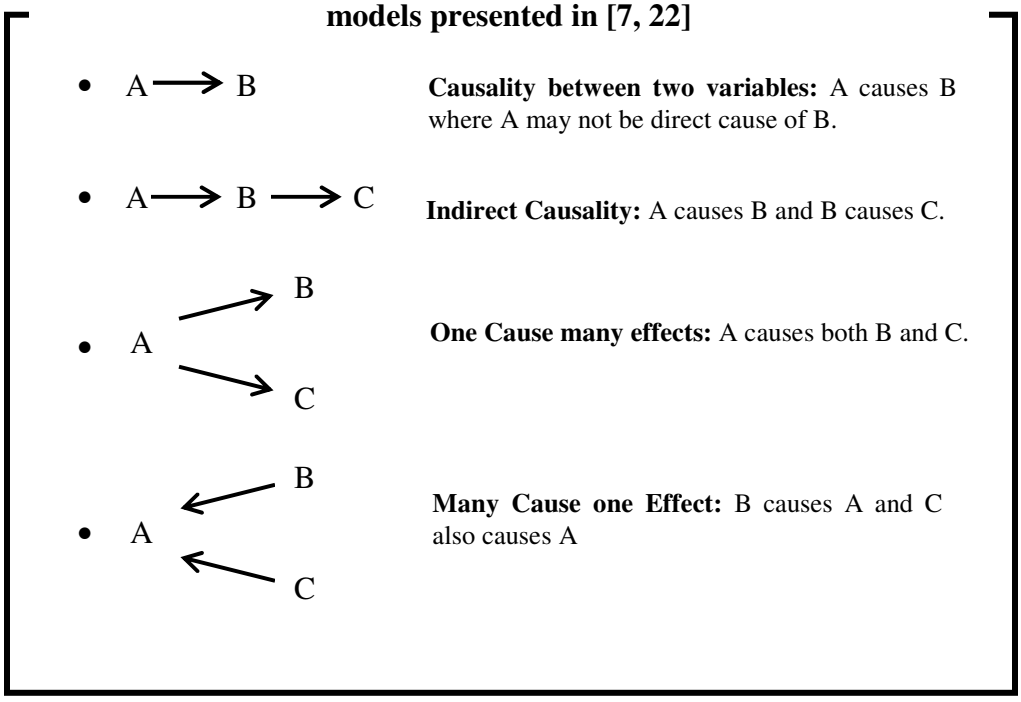
1.1 Problem Definition

Semantic relations between various text units play an important role in natural language understanding, as key elements of text coherence. Examples of such relationships are elaboration, explanation, contrast, attribution, etc. The automatic identification of these semantic relationships is very important for many language processing applications such as question answering, and summarization**.** One of the most pervasive yet very challenging semantic relations is *cause-effect*. The reason is that, most of the time, the automatic identification of causal relations requires a deep semantic analysis of the relevant causal contexts.

Various natural language processing (NLP) researchers [1, 6, 9, 10, 24] have focused their efforts in devising approaches for causal information identification from natural language text. However, most of these approaches focus on predefined linguistic patterns employed in supervised learning models. In this thesis, an unsupervised approach to automatically identifying causal information between inter- and intra-sentential events in web news articles without relying on deep processing of contextual information is proposed. 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

causal information from text. The basic causal context consists mainly of two events, the *cause* (**a**) and the *effect* (**b**), such that **a** → **b** (where '→' means 'cause'). This binary relationship leads to possible arrangements of events which can involve more than two events as shown in figure 1.1. In this thesis an events are defined as <[sub_e] $verb_e$ [obj_e]> instances, where the subject or the object can be missing. Our approach focuses on binary causal relationships where an event **a** is the direct or indirect cause of an event **b**.

Figure 1.1. Causal models. These models are similar to various models presented in [7, 22]



Binary causal relationships are learned and tested on two domain specific data sets (one on the Iraq war and one on hurricane Katrina) collected from the web. For each data set the model identifies event pairs that are potentially causal. We rely here on the hypothesis

that natural language events that explain a particular scenario tend to be strongly correlated and are thus good candidates for cause-effect information. Examples of such causal event pairs are given in Figure 1.2.

Figure 1.2. Binary causal examples for each data set

1. **Data set:** Hurricane Katrina

Example: <Six people were {killed}>, over a million customers were without electricity after <Hurricane Katrina {struck} south Florida> as a Category 1 storm.
Type: Intra-sentential causal relationship
Causal Relation: "<Hurricane Katrina {struck} south Florida>"→ "<Six people were {killed}>"

2. **Data set:** Iraq War

Example: <Pentagon {fears} last-ditch Iraqi chemical attack>. Iraqi leaders could wait for US and British troops to reach Baghdad to < {launch} a chemical weapons attack>, a US official said.
Type: Inter-sentence causal relationship
Causal Relation: "<{launch} a chemical weapons attack>" → "<Pentagon {fears} last-ditch Iraqi chemical attack >"

1.2 Overview of Approach

Identifying automatically the text snippets that follow one of the causal models shown in Figure 1.1 is an overambitious task and requires automatic discovery of cause-effect event pairs (which we call *bigram causality* or *binary causal relationship*) which in turn help us to learn all the remaining trigram causal chains. In this respect the proposed approach focuses on the discovery of binary causal relationships, i.e. **a → b** where an event (**a**) is the cause of event (**b**). This causal relationship can be direct or indirect due to

the high variability of natural language. Two domain specific data sets on the Iraq War and Hurricane Katrina have been used to test the proposed model and to generate potential candidate event pairs which are filtered later on in order to distinguish cause-effect pairs. Appropriate candidate pairs encoding causality are determined based on the hypothesis that natural language events that explain a particular scenario tend to be strongly correlated. Therefore it is appropriate to consider highly correlated events that explain a scenario as encoding a potential causal relationship. Figure 1.2 shows bigram causality examples for each data set mentioned above.

Many NLP semantic processing tasks are solved through pipelining where each layer provides results for the next in order to solve some major task in the end. The proposed approach is also tackled in three layers of processing. Main processing objectives of each of these layers are given below:

1. *Identifying Topic-Specific Scenarios and their Events* – For each dataset, first topic-specific scenarios are discovered using a hierarchical topic model which identifies fine grained topics by capturing relationships between words [15]. Then events associated with each topic-specific scenario are identified.
2. *Generating Event Pair Candidates* – Using as input the scenarios and their events identified in layer 1, similar events are grouped and then candidate event pairs are generated by mining frequent pairs of events.
3. *Identifying Causal Event Pairs and their Roles* – Using the event pair candidates, statistical measures of independence and strong dependence [22] are applied to identify causal dependencies. Once a pair is identified as causal, the system

determines its Cause and Effect roles based on a novel metric, the Effect-Control-ratio.

1.3 Challenges

Some of the important research challenges in the identification of causal relationships in natural language text are presented in this section. These research challenges are classified into "system design and development" and "evaluation challenges".

1.3.1 System Design and Development Challenges

Following are some of system design and development challenges raised while devising this approach.

1. How can we automatically identify topic-specific scenarios existent in a particular text collection with no prior domain knowledge?
2. What is the relationship between these scenarios? How can we exploit this relationship for our task of causality detection?
3. How are the cause-effect relationships expressed in natural language text?
4. What text snippets (contributing to a scenario) encode causality and how should we identify these text snippets?
5. How to tackle implicit causality expressed in text without relying on predefined linguistic cue phrases?
6. How to distinguish cause-effect relationships from other semantic relationships?
7. There must be a concrete way to explain causality. Multiple perspectives of causality have been developed and this raises the question of what would be the best tests of causality we need to identify before devising our approach?

8. Is this general notion of causality applicable to natural language text?

1.3.2 Evaluation Challenges

Following are some of challenges need to be addressed for evaluation of such systems.

1. Keeping in mind the complexity of this task, what should be the mechanism to evaluate such system?
2. In theory, the two events encoding a cause-effect relationship can appear anywhere in a natural language text. Causality can exist between events in the same sentence or at a distance of multiple sentences and this makes it difficult for humans to evaluate. Moreover, can we empirically identify the distance between two events encoding a cause-effect relationship? What would be the largest distance for which people can easily and reliably identify the events are being causally related?
3. Some data sets are difficult to evaluate e.g. Iraq war data set causal pairs are difficult to evaluate as compared with Hurricane Katrina because annotators are required to have some background knowledge about the particular domain (e.g. Iraq War) before they annotate test data for causality task evaluation. This raises the challenge of educating annotators about domain specific knowledge to have accurate evaluation possible for such systems.

Our approach has tried to address all the above challenges while devising approach for causal relationships automatic detection.

This thesis is structured as follows. The next chapter presents relevant previous work.

Chapter 3 describes the model and introduces the modules used at each processing layer.

The system evaluation is presented in Chapter 4, followed by discussions on future work in chapter 5.

CHAPTER 2

BACKGROUND

2.1 Introduction

Causal relationships between text units are an important feature of natural language that makes understanding and reasoning possible for humans. Each text analysis requires the exploration of semantic relationships and causality can be distinguished here as one of the most important relationships. Causality can be expressed both explicitly and implicitly in natural language in various ways [4, 8, 10, 11]. However, most of the computational approaches for cause-effect relationship detection focus mainly on the use of some predefined lexico-syntatic patterns of language [6, 9, 10]. These approaches however, have their limitations due to the challenges imposed by the highly ambiguous nature of these language patterns. Moreover, these patterns are not well suited for a detailed temporal analysis. The absence of causal markers makes the task even more challenging for automatic causal systems [11, 18, 19, 20, 23].

This chapter provides a review of the English language features of the explicit and implicit causality expressions in English, along with appropriate definitions of causality and computational approaches for the discovery of the cause-effect text components.

2.2 Causality in Natural Language

A detailed analysis of various English expressions of cause-effect relationships is mandatory before devising any approach the automatic identification of causal relations

from text. Various researchers have identified and analyzed relevant causal language features [4, 8, 10, 11]. Cause-effect relationships between two text units can be expressed either explicitly by using cue phrases or implicitly. Girju and Moldovan 2002 [10] reviewed four major classes of explicit causal relations which are presented below:

1. **Causal Connectives**

   These are cue phrases which connect two text units in causal relationship. A few such examples of adverbial phrases are "The meaning of a word can vary a great deal depending on the context. *For this reason*, pocket dictionaries have a very limited use.", "A local man was kept off a recent flight *because of* a book he was carrying." [10].

2. **Causative Verbs**

   These are causal verbs which connect two elements e.g. '*lead to*', *kill*', '*poison'* etc. Causative verbs combine two causal roles "cause" and "effect" mainly in the format "NP-Cause Verb NP-Effect". However, the semantics of these causative verbs is ambiguous since they do not refer to causality in all contexts.

3. **Conditionals**

   Conditional statements express causality by relying on "if .. then.." statements. One such example is "*If* this bridge falls *then* people will die".

4. **Causative Adverbs and Adjectives**

   Adverbs and adjectives can encode cause-effect information. A few such examples are "Brutus *fatally* wounded Caesar" and "Caesar's wound was *fatal*" [8, 10].

Causality detection approaches using such causative expressions must have a mechanism which helps in disambiguating the linguistic patterns. Causality can also be expressed

implicitly as well and this brings more challenges to this task. Implicit causal contexts do not link text elements based on cue phrases. Instead, they require deep semantic inference in order to judge causal relationships. Some implicit causative expressions presented previously by Girju and Moldovan 2002 [10] are presented next:

1. Complex Nominals and Other Complex Expressions

Complex nominals are expressions of the type "N N" (where N is a noun) -- e.g. "*cold tremble*" [10]. There are also complex expressions with ambiguous explicit cues of form "NP_1-producing NP_2" (e.g., "*malaria-producing mosquitoes*"). These expressions are ambiguous and they can appear in non-causal contexts as well (e.g., "*leather-producing factory*").

There are also other complex expressions with unambiguous explicit cues (e.g., "NP_1 caused NP_2" where "caused" is unambiguous, specifically when used in causal contexts.

2. Implicit causality of verbs

Implicit causality of verbs dictates the reader to determine that pronoun in text is pointing to which referent [4, 10]. For example Caramazza et al [4] observed examples like "The actor *admired* policeman because he was brave" to explain implicit causality of verb *admire* is the reason of determining that pronoun "he" is referring to policeman rather than actor.

3. Discourse Structure

Causal relations can also occur and thus be analyzed at the discourse level. Discourse passages are characterized by characters, goals, motives, events, plans etc [11]. This set of factors contributes towards generating inferences about causal events and their

effects. Researches who focus on the detection of intra-sentential causal connections need to understand and make inferences by relying on these discourse factors. Our approach captures intra-sentential causal relationships as well. Although such contexts require a deep and accurate analysis, our proposed approach has tried to approximate it by looking at the scenarios in which events would be contributing towards same kind of goal with strong semantic relationships between them. This is a necessary approach since the state-of-the-art in discourse processing does not allow for an in-depth and accurate analysis of text [18, 19, 20].

2.3 Definition of Causality

Causality has been studied for a long time from different perspectives by philosophers, logicians, linguists, data-mining researchers, bio-statisticians, and economists [7, 12, 16, 22, 26].

In philosophy and logic, two of the most influential theories are the counterfactual theory of causality [16] and the manipulation theory of causality [26]. These theories identify as causal conditions (1) the temporal precedence of the cause (**a**) and the effect (**b**) events, and (2) the causal dependency between them (**a→b**). Since the manipulation theory of causality was proven to provide an easy and objective notion of causality on some language tasks [1], it was used for annotation and evaluation purposes for our approach. For example, Beamer and Girju 2009 [1] identify an important condition for causality: 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** (details are given in [1]). This annotation test is both simple to execute mentally and is relatively objective.

Consider for instance the example "*Mary shot the thief. He died in an hour*." In this context the shooting event caused the thief's death. Had Mary not shot him, one could necessarily infer that the thief would not have died is true.

The automatic identification of causal relations has been addressed by the data mining community as well. Various approaches [7, 22] have been proposed to learn chains of events in structured datasets, such as census data and transaction databases. These approaches identify causality by employing the constraint based model for Bayesian network inference. A causal network using the Markov condition can be learned using the following rules (cf. [22]):

Rule1) If variables a, b, c are pair wise dependent and if a and c become independent conditioned on b, then three causal models are possible:

a → b → c OR a ← b → c OR c → b → a

In order to choose the correct model from these three choices we need to rely on prior information. For example, Silverstein et al [22] analyze an example from the census data where the variables "*voting*", "*drive a car*", and "*18 years old*" are pair-wise dependent (e.g., since "*voting*" entails "*one can drive a car*" , they are thus dependent). If the system predicts that conditioned on the variable "*18 years old*", the other two variables "*voting*" and "*drive a car*" become independent. Thus, three causal arrangements are possible:

"*voting*" → "*18 years old*" → "*drive a car*", or

"*drive a car*" → "*18 years old*" → "*voting*", or

"*voting*" ← "*18 years old*" → "*drive a car*"

The correct arrangement can be inferred based on the prior information about which of these variables has no cause. In this example, "*18 years old*" has no prior cause (nothing

causes somebody to be "*18 years old*") and thus, the last arrangement can be inferred (i.e., "*voting*" ← "*18 years old*" → "*drive a car*").

Rule2) If variables a, b and a, c are pair wise dependent, but b and c are independent then if b and c become dependent conditioned on a, then it can be inferred that b and c cause a (i.e., (b and c) → a).

Silverstein et al. [22] failed to successfully apply Rule1 to textual data since they did not have any prior information available to choose the correct model out of the three choices specified by Rule1 above. In order to apply on text data the causal inference method proposed by Silverstein et al [22], one needs to take into consideration the following issues:

- Textual data is unstructured and, unlike for structured data (e.g. census data), its discrete variables are not readily available. Causality can exist between various segment of text (e.g., noun phrases, verb events, etc). Therefore the text has to be processed first and the meaningful text units that encode causality need to be identified.
- In textual data there is no prior information available for the identification of the Cause and the Effect roles as required by Rule1 above.

In this thesis, we also employ the statistical measures introduced by Silverstein et al. [22]. However, these measures are adjusted in such a way that they are suited for unstructured linguistic data.

2.4 Learning Causality

In linguistics, many researchers have focused on the analysis of English expressions which can encode causal relations. These expressions are lexico-syntatic patterns (e.g. "*mosquitoes cause malaria"* is of the type "NP-Cause verb NP-Effect") which are employed most of the time in supervised learning models [6, 9, 10]. However, in these approaches, the patterns are identified either manually or semi-automatically, which makes the systems difficult to port to different domains. Moreover, most of these patterns are ambiguous and thus, they need to be disambiguated in context. One such approach to identifying causal relations [9] which relies on a pattern disambiguation procedure (for patterns of type "NP Verb NP") achieves a precision of 73.91%. The approach tests first if the component noun phrases in such patterns are present in WordNet [27], and then queries the Internet or some large text collection using the pattern "* verb/expression *". Then semantic constraints are generated from WordNet to ensure that the pattern identifies causal relation instances. The extracted instances are annotated and provided as input to the C4.5 decision tree learning model [21] which is then tested on unseen examples.

Causality can also be expressed implicitly with no causal markers (e.g., *as a result of, due to, because of, cause*), especially at the discourse level. One such example is "*I just missed the bus. I will be late for the meeting*". Discourse passages are characterized by protagonists, goals, motives, events, plans, etc [11], which are presented in a cohesive and coherent way in text. Such causal discourse relations are inferred from the context,

rather than being explicitly stated. Thus, the absence of causal markers makes the task even more challenging for automatic causal systems [11, 18, 19, 20, 23].

In this thesis, proposed research focus is on both explicit and implicit causal relationships at intra- and inter-sentential levels. Our approach tries to approximate the context by identifying first topic-specific scenarios containing events that describe them and which are connected by strong causal relationships. This is a necessary approach since the state-of-the-art in discourse processing is still far from providing a deep analysis of textual context [18, 19, 20].

Other researchers [1] have made use of statistical methods to learn approximate solutions for this hard problem without relying on cue phrases. These approaches employ special data sets (e.g., those where the events are temporally ordered), which makes the causal learning task easier. Since the causal task requires a semantic as well as a temporal analysis of the events to ensure accurate results, using such data sets reduces the complexity of the task. One such recent approach [1] relies on a statistical measure, *Causal Potential*, which is applied on a text corpus of screen plays where the verb events are already temporally ordered. The obtained degree of correlation between the causal potential prediction and the human judgments was 0.497. The authors used the Spearman's rank correlation coefficient and verified that the human ranking and the ranking predicted by their measure were positively correlated. Event pairs which score very highly had very high observed causal frequencies and vice versa.

Other researchers [24] have proposed another measure, the *Event Causality Test* (ECT), which was employed in the discovery of causal relationships between events in search queries extracted from temporal query logs. Their model determines if two queries are

causally related if the change in one query causes the other query to change during a time period. They assumed that if the frequencies of two queries increase over a time period, then they might be correlated. Following this assumption queries were extracted from temporal query logs and then checked for causality based on the Event Causality Test (ECT) and then re-ranked using the *Granger Causality Test* (GCT) [12]. For the top-100 examples predicted, the model achieves an accumulated precision from 32% to 21% for instances ranked 1 to 99. Instead of relying on datasets with temporally-ordered events or on temporal classifiers which are difficult to build accurately [5], in this thesis an approach is proposed which finds first which events are strongly correlated and then, based on a novel metric (called *Effect-Control-ratio*), identifies the Cause and the Effect roles of a pair instance once it is identified as causal.

CHAPTER 3

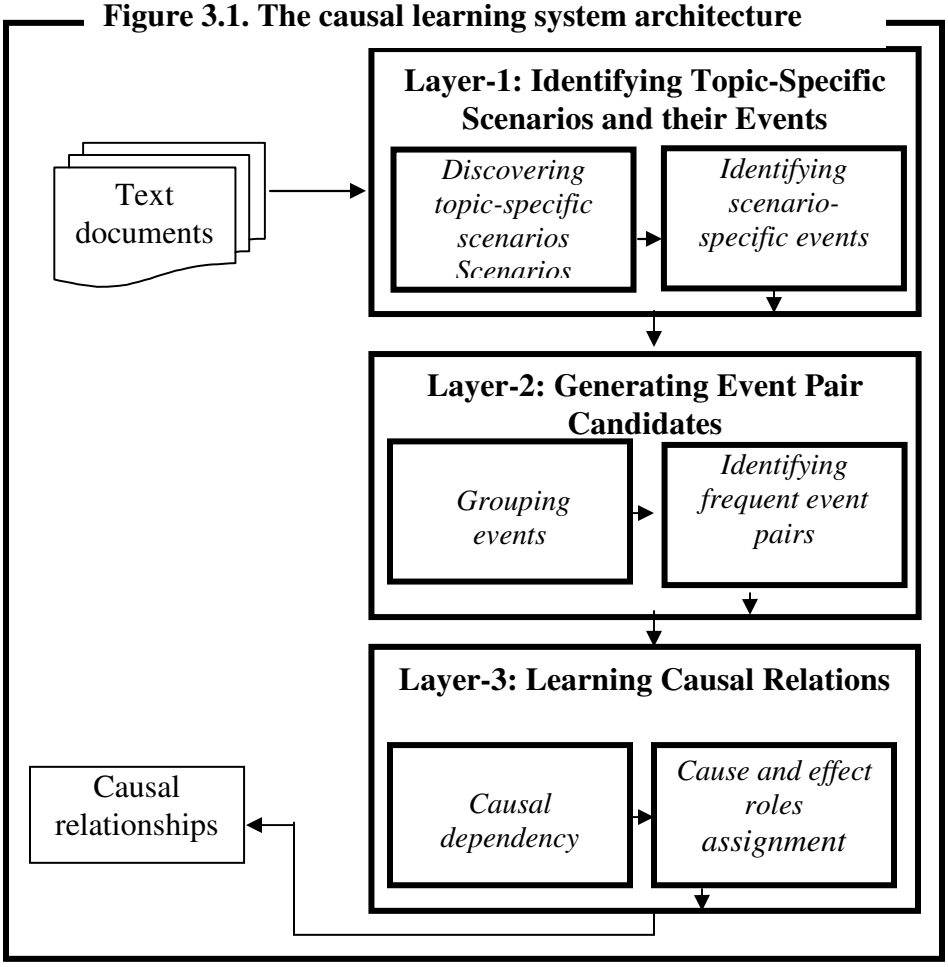
METHODOLOGY

3.1 Introduction

This chapter explains three-layer unsupervised statistical system (Figure 3.1) proposed to automatically identify causal information in domain specific text collections. The system discovers inter- and intra-sentential causal information without relying on a deep context analysis. This will give better insights into how far we can go with such a knowledge-poor approach to causality detection. Moreover, our approach is totally unsupervised which saves us the trouble of getting manually annotated data which is very expensive.

The proposed approach focuses on the detection of inter- as well as intra-sentential causal relations. Although such contexts require a deep and accurate analysis which is impossible today without heavy supervision, we have tried to approximate it by looking at topic-specific scenarios containing events that describe them and which are connected by strong dependency relationships. We hypothesize that events contributing to one particular scenario tend to be strongly correlated, and thus make good candidates for the causal task. Identifying causal relations and assigning the Cause and the Effect event roles are done based on a set of statistical measures. Moreover, the approach is applied and tested on two text corpora. Three layers of processing are briefly reviewed here along with details on implementation in section 3.3.

Figure 3.1. The causal learning system architecture



1. *Identifying Topic-Specific Scenarios and their Events* – For each dataset, first topic-specific scenarios are discovered using a hierarchical topic model which identifies fine grained topics by capturing relationships between words [15]. Then events associated with each topic-specific scenario are identified.
2. *Generating Event Pair Candidates* – Using as input the scenarios and their events identified in layer 1, similar events are grouped and then generate candidate event pairs by mining frequent pairs of events.
3. *Identifying Causal Event Pairs and their Roles* – Using the event pair candidates, statistical measures of independence and strong dependence [22] are applied to

identify causal dependencies. Once a pair is identified as causal, the system determines its Cause and Effect roles based on a novel metric, the Effect-Control-ratio.

3.2 Text Corpus

Our proposed approach requires a domain specific text collection to determine causal relationships between events. In order to acquire such domain specific collections, we crawled the web and collected two datasets: one on the Hurricane Katrina[1] and one on Iraq War[2]. The downloaded text archives have been post-processed such that the various formatting tags were removed (e.g., html etc.). This way we have collected data (excluding stop words) with following statistics:

Table 3.1. Text Corpus Statistics

Text Corpus	News Articles	Word-Tokens	Word-Types
Hurricane Katrina	447	189,840	14,996
Iraq War	556	304,481	20,629

Here word-tokens refer to words separated by space in text corpus and word types refer to unique words (correspond to text corpus vocabulary).

News articles for each of the two domains were used to run the proposed unsupervised causal learning tasks. The models were evaluated on a subset of these text collections as explained in chapter 4.

[1] http://websearch.archive.org/katrina/list.html, http://www.news.google.com/archivesearch
[2] http://www.comw.org/warreport/

3.3 Unsupervised Learning for Causal Relationships

This section explains three layered approaches proposed for the causality detection task (see Figure 3.1).

3.3.1 Layer-1: Identifying Topic-Specific Scenarios and their Events

The first layer module identifies scenarios and their events. This module clusters text units according to their probability distributions to build 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 allegations and the inspection process in Iraq" and "the come back of American forces and post-war developments", etc). Thus, our intuition is that the events describing a particular topic-specific scenario are strongly correlated. Examples are given in Table 3.2.

Table 3.2. Examples of topic-specific scenarios and their events

Scenario1: **"**War effects - economic progress in Iraq and side effects on the world's economy**"**	**Scenario 2:** **"**US accusations and the UN inspection**"**
"rehabilitate Iraqi people"; "writing new constitution"; "destroying chemical weapons plant"	"Iraq might have developed chemical weapons"; "UN teams asking scientists various questions"; "Suspecting the existence of a chemical weapon plant in Iraq"

For example, sentences such as "*Iraq might have developed chemical weapons*" and "*the UN team asking scientists various questions*" denote events which identify the scenario **"**US accusations and the UN inspection**"** and are strongly correlated.

These scenarios can be identified by clustering semantically similar text units into different clusters. For this we rely on topic modeling. Basic topic models like PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) [2, 14] cluster the words in a given text collection into topics. Topic models are generative models where each topic is a multinomial distribution over words and each document is generated by the mixture of topics. Therefore a text document can be generated by multiple topics. More advanced topic models, such as the Pachinko Allocation Topic Model (PAM) [15], capture not only correlations between words to determine topics but also identify relationships between topics. In this research we employ PAM to discover topic-specific scenarios only, and leave the correlation between topics for future research. This can be done in two ways: 1) applying PAM to words or 2) applying PAM to events. Since statistical models such as PAM require large data sets as input, it is not feasible to run it on events since they are less frequent than simple words. Therefore we run PAM on words and then extract the events corresponding to each identified scenario. This procedure is explained next.

3.3.1.1 Discovering Topic-Specific Scenarios

Using as input a large text collection, PAM generates an *n*-level Directed Acyclic Graph (DAG) structure (see Figure 3.3 for a 4-level DAG). It starts with the root node at level 0 which is connected to all other nodes at level 1 (called *super topics*). In turn, each super node is fully connected with other nodes (their children) at level 2. Nodes at level 3 are

the leaves which contain words. Each internal node is a super topic which has a multinomial distribution over all its subtopics (children at the next level). Words are assigned to sub-topics based on their probabilities of occurrence conditioned on subtopic Z_t, i.e $P(w| Z_t)$. Figure 3.3 also shows correlations between topics along the edges between super topics and subtopics. However, this research does not consider correlations between topic clusters.

In order to learn the DAG structure, PAM uses a generative model in which each super topic is represented as a multinomial distribution over subtopics and each subtopic has a multinomial distribution over words. Each super topic is associated with a Dirichlet distribution parameterized by α_i where the i dimension is equal to the number of subtopics and each subtopic at level L-1 is associated with a single Dirichlet distribution parameterized by β. For the sake of simplicity, parameter β remains fixed and is not re-estimated for a particular data set. For each document, the PAM generative model samples a multinomial distribution of super topics and then samples the path for each word w (path from root to level L-1) in document d from the multinomial distributions, and finally samples words from the level L-1 topic distribution $\theta^{(d)}_{tL-1}$ (see Figure 3.2).

PAM uses Gibbs sampling for the parameters' inference procedure. With a four level DAG, for each word w in a document d the joint probability of super topics and sub-topics is as follows:

$$P(z_{w2} = t_k, z_{w3} = t_p | D, z_{-w}, \alpha, \beta) \propto \frac{n^{(d)}_{1k} + \alpha_{1k}}{n^{(d)}_{1} + \sum_{k'} \alpha_{1k'}} \times \frac{n^{(d)}_{kp} + \alpha_{kp}}{n^{(d)}_{k} + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m} \quad (3.1) \text{ (cf. [15])}$$

Here Z_{ts} is the super-topic and Z_{tr} is the sub-topic. Z_{-w} means that it does not consider the current assignment for word w, $n_{kp}^{(d)}$ is the number of words in document d labeled with the super topic k and the sub topic p, n_{pw} is the number of times word w labeled with topic p, W is total number of words, α_{xy} and β are the Dirichlet parameters. α_{xy} captures the correlation between the super-topic x and the sub-topic y. In order to learn the DAG structure, the above inference procedure is used for a number of iterations to label word w in document d as belonging to various topics. During each iteration α_{xy} the parameters are updated using an approximate method of moment matching. For the learning topic-specific scenarios and their events we use the same PAM procedure on words (excluding stop words) for the given data sets (see Table 3.1).

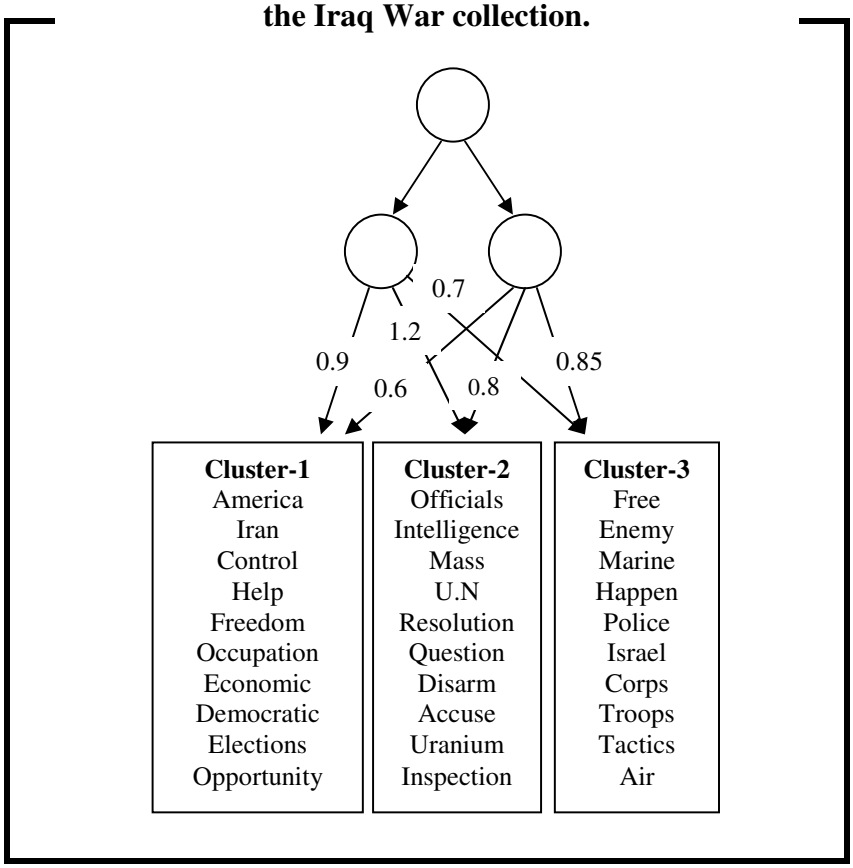
Figure 3.2. The PAM generative model (cf. [15]).

1. Sample $\theta_{t_1}^{(d)}$, $\theta_{t_2}^{(d)}$, ..., $\theta_{t_s}^{(d)}$ from $g_1(\alpha_1)$, $g_2(\alpha_2)$, ..., $g_s(\alpha_s)$, where $\theta_{t_i}^{(d)}$ is a multinomial distribution of topic t_i over its children.
2. For each word w in the document,
   - Sample a topic path $\mathbf{z}_w$ of length L_w: $< z_{w1}, z_{w2}, ..., z_{wL_w} >$. z_{w1} is always the root and z_{w2} through z_{wL_w} are topic nodes in T. z_{wi} is a child of $z_{w(i-1)}$ and it is sampled according to the multinomial distribution $\theta_{z_{w(i-1)}}^{(d)}$.
   - Sample word w from $\theta_{z_{wL_w}}^{(d)}$.

The scenarios identified by PAM for data set are then analyzed for event identification. Figure 3.3 shows the top 10 representative words for each of the three scenarios identified for the Iraq War collection. Two human annotators labeled the discovered scenarios as cluster-1 ("War effects - economic progress in Iraq and side effects on the world's economy"), cluster-2 ("US accusations and the UN inspection"), and cluster-3 ("Pre-war: War strategies and planning").

We use the words in each topic-specific scenario for each dataset to extract the events belonging to these scenarios. The event identification process is explained in the next subsection.

Figure 3.3. The three topic-specific scenarios for the Iraq War collection.



3.3.1.2 Identifying Scenario-specific Events

At this point, we need to identify events contributing to a particular scenario. We can do this by recovering the sentences which correspond to the scenario clusters discovered by PAM and then by identifying the events in these sentences. We represent each cluster as a vector *v* whose elements are words. Each word has a weight *w* which is the number of times the word was assigned to a scenario cluster. Since we want to ignore unimportant words, we keep only those elements in *v* with the frequency greater than a threshold *t* (set up here to t=10). Thus, each sentence *s* (the weight for each word in sentence *s* is 1.0) is assigned to a cluster *v* based on the normalized cosine similarity measure between *s* and *v* (N is the vocabulary size):

$$\text{cosine-similarity}(\overrightarrow{s}, \overrightarrow{v}) = \frac{\overrightarrow{s} \bullet \overrightarrow{v}}{\sqrt{\sum_{i=1}^{N} s_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}} \quad (3.2)$$

The procedure to assign sentence *s* to scenario cluster *v* is as follows:

1. Calculate cosine-sim(s,v_i) for all scenario clusters. Here *i* is the number of clusters.

2. Assign sentence *s* to cluster v_i with which it has the highest cosine measure. In case of a tie between clusters assign the sentence to all clusters with the same cosine measure.

From these scenario-specific sentences we identify their events (i.e., <sub_e $verb_e$ obj_e> instances). Thus, we rely here on a semantic role labeler, SWiRL [25], to identify the

subject and the object. Figure 3.4 shows a snapshot of scenario-specific sentences and their events.

Figure 3.4. Sentences along with their events (shown in italic) assigned to the scenarios identified for the Iraq war collection.

Scenario 1 – "War effects - economic progress in Iraq and side effects on the world's economy**"**

Event in context: <*Financial markets {wobble}*> as Iraq war unfolds.

Scenario 2 -- "US accusations and the UN inspections**"**

Event in context: <*Pentagon {fears} last-ditch Iraqi chemical attack*>.

Scenario 3 -- "Pre-war: War strategies and planning**"**

Event in context: If the <*Kurds join the Shiites*> in a general offensive against the Sunnis, <*the Sunnis will probably lose*>.

3.3.2 Layer-2: Generating Event Pair Candidates

This layer module generates event pair candidates. First, all the events generated at layer 1 are grouped together based on their similarity. Then frequent event pairs are identified. These procedures are explained in this section.

3.3.2.1 Grouping Events

Similar events identified for each scenario in the previous layer need to be grouped together. For example, the instances "*UN teams suspect 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 based on the naïve lexical similarity between events. The procedure is presented next.

Procedure: Grouping events

Input: Events $e_1, e_2, \ldots, e_n$

Output: Event Groups $G_1, G_1, \ldots, G_m$ where $m \leq n$

1. Initially place every event $e_i = <[\text{sub}_{ei}]\ \text{verb}_{ei}\ [\text{obj}_{ei}]>$ into its own group ($G_i = \{e_i\}$)
2. For each event $e_i \in G_i$

   For each group $G_{k \neq i}$ 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)

   Identify that G_k for which the average_cosine_similarity(e_i, G_k) is maximum. Then add event e_i to G_k and discard G_i. In case of tie put event e_i randomly in any of the tie groups.
3. Return the resulting event groups $G_1, G_1, \ldots, G_m$.

3.3.2.2 Identifying Frequent Event Pairs

Once events are grouped as shown above, candidate event pairs need to be generated for the causal detection step. For this task, we rely on the FP-Growth algorithm [13] to mine frequent event pairs with minimum support of 5. These are pairs (**a**, **b**) which appear in at least 5 documents (i.e., news articles). The FP-Growth algorithm is used frequently in the data mining community [3, 22] to generate patterns (combinations of items) and learn associations between various pattern items. Transaction database records containing information about what items people are purchasing together provide interesting patterns,

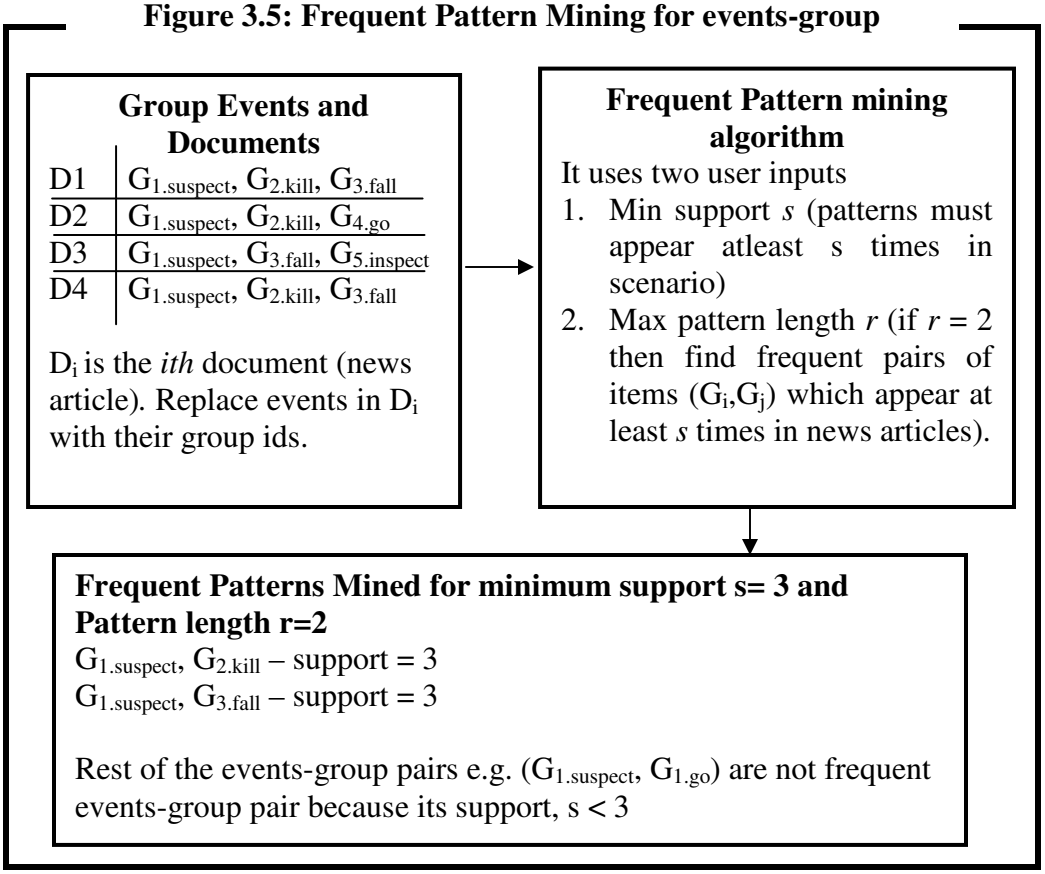
such as (laptops, hard drive) when people tend to buy hard drives when purchasing laptops. Another example of frequent pattern is (bread, butter) – indicating that people tend to purchase bread and butter together with a frequency of at least *n* (called *minimum support*). Silverstein et al [22] also used the frequent pattern mining algorithm to mine patterns before identifying causality between census variables and text words.

FP-Growth algorithm was applied on our text data sets considering documents as transactions. 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 explained by the Grouping events procedure (see Figure 3.5 for an example of application of FP-Growth algorithm). Next, we apply FP-Growth which generates event group pairs (G_i, G_j) with minimum support 5. Here group G_i can contain an event instance like "*US suspects Iraq*", while a group G_j contains an event instance such as "*Iraq develops weapons of mass destruction*". If the event pair (G_i, G_j) appears in at least 5 documents, we say that it is a frequent pair of minimum support 5 (see Figure 3.5 for example of FP-Growth application with minimum support 3).

Causality learning statistics are applied on these frequent events-group pairs (explained in next section) where each group G_k, of a frequent pair (G_i, G_j) represents one event.

Therefore we are using words, event(s) and events pair instead of group(s) and events-group in next section.

Figure 3.5: Frequent Pattern Mining for events-group



3.3.3 Layer-3: Learning Causal Relations

This section focuses on the identification of causal relations between two events. Various novel statistical measures are introduced. This task is based on two procedures which are explained below:

1. Determine if two events identified as frequent in the previous layer are strongly correlated. This condition is similar to the condition introduced by the manipulation theory of causality which states that modifying the causing event entails predictably modifying the effect event [26].
2. Assign the Cause and the Effect roles once the two events are identified as causal.

3.3.3.1 Causal Dependency

For this task we used the Chi-square test of independence and the dependence test proposed in [3, 22]. The Chi-square test of independence is used to test the hypothesis according to which two events (**a**, **b**) are independent. A two-tailed test of independence proposed in [22] is defined as follows:

1. If $\chi^2 > \chi_\alpha^2$ at the level of significance α, then the two events are correlated. For example, α = 5% means that 5% of the uncorrelated pairs are incorrectly judged as correlated.
2. If $\chi^2 < \chi_\alpha^2$ at the level of significance α, then the two events are uncorrelated or independent. For example, if α = 95% then 5% of the uncorrelated pairs are incorrectly judged as correlated.

Silverstein et al [22] claimed that it is easier to predict independence (uncorrelation) using the Chi-square measure. However, there are pairs of variables (**a**, **b**) which are neither correlated nor uncorrelated. Therefore, they have defined an interesting equivalence between the Chi-square test and the correlation measure (i.e., $\chi^2 = np^2$ where *n* is data set size) and assumed that the variables are dependent or strongly correlated only if "the correlation coefficient is greater than a cut off value", i.e. the confidence level is below $np^{2.}$:

$$\chi^2 = np^2 = \frac{n(cn-ab)^2}{ab(n-a)(n-b)} \qquad (3.3)$$

Here *n* is the data set size (the number of documents) and *c* is the support ***(a, b)***, i.e. the number of documents in which **a** and **b** appear together (*a* is the number of documents in which event **a** appears and *b* is number of documents in which event **b** appears).

For this research, we used a confidence level of 95% and a significance level of 5% to identify the dependence between events. Since we consider event pairs (**a**, **b**), only two variables are being used. Thus, the degree of freedom for the Chi-square test is 1. For the current task, we exclude the pairs with negative correlations. However a negative correlation may lead to interesting causal relationships (e.g., the occurrence of event **a** causes event **b** not to occur). Silverstein et al [22] also used a support threshold requirement for the pairs (i.e., the events in a pair (**a**, **b**) need to appear together at least *s* times). They used this support threshold requirement to avoid infrequent pairs which may not be interesting. Moreover, the effectiveness of Chi-square increases when it is applied to frequent pairs. Similarly we have used the FP-Growth algorithm for frequent event pair mining with support of at least s (we consider s=5).

3.3.3.2 Cause and Effect Roles Assignment

Once we identify two events as dependent, we have to determine which of them is the Cause and which is the Effect. For example, after the system has identified the events "*Iraq used chemical weapons*" and "*US accused Iraq*" as dependent (i.e., the occurrence of one event is dependent on the occurrence of the other event) we still have to find which event occurrence causes the other event to occur. Therefore, a mechanism is required which can assign roles to the events in a causal context. Thus, we have devised a novel metric called Effect-Control-ratio to label events with their corresponding roles. The Effect-Control-ratio is defined and explained below:

Effect-Control-ratio *(a,b)* =

$$\frac{\frac{sup(a,b)}{sup(b)-sup(a,b)} \cdot \frac{sup(a,b)}{\max\limits_{t}(sup(a,b_t))}}{\frac{sup(a,b)}{sup(a)-sup(a,b)} \cdot \frac{sup(a,b)}{\max\limits_{s}(sup(a_s,b))}} \qquad (3.4)$$

Here *sup(a,b)* is the support of the event pair (**a, b**) (i.e., in how many documents events **a** and **b** appear together). We consider only the event pair for which the minimum support *(a,b)* is 5. If event **a** appears with *m* other events in the text documents then, *max(sup(a,b_t))* chooses the event b_t with which event **a** has the highest support and uses that support here. Similarly *max(sup(a_s,b))* is defined for event **b**, (i.e., it chooses the event a_s with which event **b** has the highest support). The Effect-Control-ratio makes the decision as follows:

1. Predict **a** → **b,** if Effect-Control-ratio *(a,b)* > 1.0
2. Predict **b** → **a,** if Effect-Control-ratio *(a,b)* < 1.0

In this research we do not deal with the case when the Effect-Control-ratio is 1.0, since we need a deeper temporal or semantic analysis for such pairs in order to decide which event is the Cause and which is the Effect. However, this does not affect our predictions much since only 2% of the event pairs in each scenario had a ratio of 1.0.

The intuition behind this ratio is simple. First consider for example, the numerator:

$$\frac{sup(a,b)}{sup(b)-sup(a,b)} \cdot \frac{sup(a,b)}{\underset{t}{max}(sup(a,b_t))} \qquad (3.5)$$

The numerator term indicates that event **a** is the Cause and event **b** is the Effect. The two fractions of the numerator indicate that:

1. If the event **a** causes **b**, then **a** can appear independently while **b**'s occurrence is controlled by **a**. In this case, the fraction *sup(a,b)/(sup(b)-sup(a,b))* will be greater than 1.0 if **b** appears more frequently with event **a** than alone (i.e., *sup(a,b)* > *sup(b) - sup(a,b)*).

2. The second fraction assumes that event **a** can be the cause of many other events, so if it appears with event $\mathbf{b}_t$ more often than with any of *m-1* other events (i.e. $sup(a,b_t) > sup(a,b_{l \neq t})$), then it indicates what fraction of this maximum support it appears with event **b.** Clearly, if $b_t = \mathbf{b}$ then this fraction would be 1.0, otherwise it will be less than 1.0.

Similarly the denominator indicates that **b** is the Cause event and **a** is the Effect event. Thus the prediction would be as follows:

1. Predict **a → b,** if the numerator > denominator (i.e., the ratio > 1.0),

2. Predict **b → a,** if the denominator > numerator (i.e., the ratio < 1.0)

Figure 3.6 shows examples of causal examples predicted by system for different scenarios from Hurricane Katrina and Iraq War.

Figure 3.6. Binary causal examples for each data set

1. Data set: Hurricane Katrina

Scenario: Hurricane Katrina disaster and damage
Example: <Six people were {killed}>, over a million customers were without electricity after <Hurricane Katrina {struck} south Florida> as a Category 1 storm.
Type: Intra-sentential causal relationship
Causal Relation: "<Hurricane Katrina {struck} south Florida>"→ "<Six people were {killed}>"

2. Data set: Iraq War

Scenario: US accusations and the UN inspection
Example: <Pentagon {fears} last-ditch Iraqi chemical attack>. Iraqi leaders could wait for US and British troops to reach Baghdad to < {launch} a chemical weapons attack>, a US official said.
Type: Inter-sentence causal relationship
Causal Relation: "<{launch} a chemical weapons attack>" → "<Pentagon {fears} last-ditch Iraqi chemical attack >"

CHAPTER 4

SYSTEM EVALUATION

4.1 Introduction

This chapter describes experiments and evaluation task for three layer unsupervised approach to identifying causality relations from relevant scenarios. We performed an evaluation at two important levels: Identifying Topic-Specific Scenarios and their Events (layer 1) and Learning Causal Relations (layer 3).

4.2 Experiment Set Up

This section presents the experiment set up on the two data sets: one on Hurricane Katrina and another on Iraq War.

Here is the list of parameters that needed to be specified for various processing layers:

1. For the PAM topic model [15], we considered two super topics and three subtopics, and we ran the model with 3,000 iterations using an initial α_{xy}=0.01 (for all super topics(x) and subtopics(y)), and β=0.01. α_{xy} and β are the parameters for the Dirichlet distributions. α_{xy} parameters are estimated to capture the relationships between a super topic and a subtopic (i.e., each super topic is associated with a subtopic with a Dirichlet distribution parameterized by α), while β which captures relationships between words with respect to topics remains constant (i.e., each subtopic is associated with a single Dirichlet distribution parameterized by β).

2. For frequent event pair mining, we used a minimum support of 5.
3. For causal dependency task we run our system for 95% and 99% confidence intervals.

The same parameter setup was used for each of the two datasets. During the experiments we observed that a DAG with 3 subtopics performs well on the scenario learning task on both data sets. A very large number of subtopics generates noisy scenarios. Therefore we preferred a small number of subtopics for both data sets.

4.3 Evaluation

We performed an evaluation at two important levels: Identifying Topic-Specific Scenarios and their Events (layer 1) and Learning Causal Relations (layer 3).

4.3.1 Evaluating the Scenario Generation Task

We evaluated each of the three clusters obtained on each text collection through blind judgments of cluster quality.

Table 4.1. Labels assigned to topic-specific scenario clusters

Text Corpus	Cluster 1-label	Cluster 2-label	Cluster 3-label
Hurricane Katrina	Hurricane Katrina disaster and damage	Global Warming and climate change issues	Rescue efforts and criticism of government rescue plans
Iraq War	War effects - economic progress in Iraq and side effects on the world's economy	US accusations and the UN inspection	Pre-war: War strategies and planning

We evaluated the PAM clusters against human judgments for each scenario. Our human evaluation procedure is similar to that used by Li and McCallum [15] who compared the results of two topic models, LDA and PAM. Here instead, we compare the scenarios identified by PAM with those labeled by human observers.

For each scenario, we provided the top-50 ranked words to two human annotators and asked them to label them as "YES", if they are semantically similar, representative for a particular scenario (topic), or "NO" otherwise. In particular, the annotators were asked to judge the semantic coherence of each cluster as a whole (e.g., are the words in the clusters related, identifying a particular topic? - we call this test *evaluating words' relatedness*). The results show that for hurricane Katrina, clusters 1 and 3 are less noisy as compared with cluster 2, and for the Iraq war the clusters 1 and 2 were also good. Tables 4.2 and 4.3 show a good inter-annotator agreement for each topic-specific cluster for both data sets. The annotators found that cluster-3 in the Iraq War dataset is noisier and difficult to judge as compared with other scenario clusters. There results are shown in Tables 4.2 and 4.3.

After the cluster annotation and evaluation which were performed independently by each annotator, the annotators discussed and agreed on the appropriate labels for these clusters (see Table 4.1).

Table 4.2. Evaluation for word relatedness for each of Hurricane Katrina scenario clusters along with the inter-annotation agreement

Test	Cluster 1	Cluster 2	Cluster 3
Related words	66%	57%	65%
Inter-Annotator Agreement	86%	94%	92%

Table 4.3. Evaluation for word relatedness for each of Iraq War scenario clusters along with the inter-annotation agreement

Test	Cluster 1	Cluster 2	Cluster 3
Related words	90%	83%	39.5%
Inter-Annotator Agreement	80%	96%	86%

We also calculated the Jensen-Shannon divergence (Formula (4.1) below) between each two scenario distributions to find how different they are. Jensen-Shannon divergence is a symmetrized version of KL-divergence and indicates the difference between two distributions. The results show that the three scenarios for both data sets are very different (Table 4.4, where "JS (i-j)" refers to the Jensen-Shannon divergence for distributions of clusters *i* and *j*).

$$JS(P||Q) = D(P||M)+D(Q||M) \qquad (4.1)$$

Where $D(P||M)$ is KL-divergence measure and $M=(P+Q)/2$

Table 4.4. The Jensen-Shannon divergence for scenario distributions. "JS (i-j)" refers to the Jensen-Shannon measure for distributions of clusters *i* and *j*

Test	JS (1-2)	JS (1-3)	JS (2-3)
Hurricane Katrina	0.75	0.69	0.68
Iraq War	0.67	0.72	0.66

4.3.2 Evaluating the Causality Detection Task

Our system discovered strong semantic relationships between events from Clusters 1 and 3 for hurricane Katrina, and from Clusters 1 and 2 for the Iraq War (see Tables 4.5 and

4.6). We evaluated the system's performance against human judgments on 100 randomly selected events pair examples for each text corpora (after the annotator agreement) as follows:

1. Hurricane Katrina: 100 events pair examples for each Cluster 1 and 3. The annotators labeled a total of 200 examples for this data set.
2. Iraq War: 100 events pair examples for each Cluster 1 and 2. Thus annotators labeled a total of 200 examples for this data set.

Our system detects inter- and intra-sentence causal relationships, irrespective of the distance between the two events. However, in order to make the evaluation task easier we considered only those events which were separated by at most 3 sentences in the articles in which they occur. In this research we focused only on those pairs of events which belong to the same article (Figure 3.6 shows examples of event pairs occurring at different distances in the same document).

Thus, the two events are provided with their surrounding context to make the annotation task easier. The annotations were done according to following guidelines:

1. Label example "YES" or "NO" if the events are *causally* dependent in context. For this task follow the causal condition adopted from [1] which says that, 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** [1].
2. Assign the Cause and the Effect roles to each causal events pair from previous step.

Table 4.5. The evaluation of the causality detection procedure for hurricane Katrina and the Iraq War data sets with a 95% confidence interval

Data set	Precision (Roles Accuracy*)	Recall	F1 measure	Annotator Agreement
HK: Cluster-1	58.3% (67.3%)	83.0%	68.4%	87%
HK: Cluster-3	50.9% (40.7%)	81.0%	62.4%	89%
IW: Cluster-1	39.1% (62.06%)	74.3%	51.2%	85.7%
IW: Cluster-2	32. % (85.7%)	80.7%	46.1%	81%

HK: Hurricane Katrina
IW: Iraq War
*Roles Accuracy is the measure of accuracy of correct roles predicted for causal examples.
Roles Accuracy = *# of correct roles predicted/# of correct causal examples.*
The annotator agreement is for given here for the causal dependency labels. For roles' annotation there was a 98% to 99% agreement between the two annotators for all clusters.

System was also tested with a 99% confidence interval (with only 1% chance of making an error). The results are shown in Table 4.6.

Table 4.6. The evaluation of causality detection for hurricane Katrina and the Iraq War data sets with a 99% confidence interval

Data set	Precision (Roles Accuracy*)	Recall	F1 measure
HK: Cluster-1	55% (66%)	55.6%	55%
HK: Cluster-3	50% (40.9%)	66%	56.8%
IW: Cluster-1	36.9% (66%)	61.5%	46.2%
IW: Cluster-2	33%(85.7%)	80.7%	47.0%

The results obtained for the causality task show that for a 95% confidence interval (i.e., there is a 5% chance of making errors) the system is likely to predict causal dependencies between events (**a**, **b**) which explains the high recall for each cluster test data (Table 4.5).

Meanwhile, the results for a 99% confidence interval show that the recall drops because the system chooses only those causal dependencies with a much higher confidence factor. Compared to recall, the precision is not as high. In order to explain this, we did some error analysis of the test data and observed that scenario-specific events are likely to be strongly dependent. However, these dependencies can happen for many reasons, not just causal (e.g., elaboration, and other semantic relationships between events). Since there can be various kinds of semantic dependencies between events, the annotators were more likely to label the examples as "not-causal" rather than "causal". Therefore, in order to improve precision, the strongly dependent events pairs thus predicted need to be filtered further to eliminate those which are due to other semantic relationship. A further filter of dependent pairs requires a deep semantic analysis which currently is a quite challenging task for inter-sentential causal events. Moreover, our system obtained a better precision for the 95% than for the 99% confidence interval. This reduction can be explained by the fact that that system is more strict in predicting the causal dependencies. This can lead to missed interesting causal events (**a, b**) which do not appear together frequently in the corpus (i.e., the events **a** and **b** are independent in the corpus), but encode a causal relationship.

The performance obtained by the system on the Cause and the Effect role assignment task is good. The Effect-Control-ratio achieved a good accuracy of 85.7% on the Iraq War cluster 2. In this research we are not using any temporal classifier nor any deep semantic analysis tool to decide which event is the cause and which is the effect. Thus, the high performance obtained by our system for the role assignment task shows that this

measure can be applied effectively when the temporal information about such events is missing.

4.4 Discussions

This section analyzes out system's results along with some issues raised in this research.

As mentioned in the previous section, our system's precision is low compared with the recall. This is due to the strong dependencies identified between the natural language passages representing the events which are not only of causal nature. Thus, the system has to further filter the predicted causal event pairs such that it identifies only the true causal instances. 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 {emerge}*> and <*hurricane Katrina {passed}*> appear together very often, and thus the system predicted them as causally dependent. The annotators labeled this example as non-causal, since these events are similar rather than causally related.

Another problem is generated by the usage of the Chi-square measure and the correlation for identifying causal dependencies. This is because Chi-square measures do not perform well for very low frequencies. Thus, better dependency measures are required.

Moreover, the event grouping process which is required to combine similar events uses a shallow lexical similarity metric. Relying only on lexical similarity is not sufficient for this task. For example, it would be better to focus on verbs and other words which are

semantically similar in the context of each event (e.g., "*the hurricane hit Florida*", "*the hurricane struck Florida*"). Such events might not be grouped and thus, might not be identified as encoding a causal relationship, for example when the data has more frequent events pair V = ("hit Florida", "kill people"), but less frequent pairs U = ("struck Florida", "kill people"). On one hand the system will detect V as a strongly dependent pair and will incorrectly predict U as independent when in reality, the pairs are similar.

In this thesis we introduced a novel metric for the assignment of Cause and Effect, the Effect-Control-ratio. This metric is very easy to apply since it does not rely on a deep semantic and temporal analysis. The good performance obtained by the Effect-Control-ratio shows that it is a good measure for the identification of these roles when a deep context analysis is difficult, or when the context is missing.

One observation interesting to make here is the fact that the annotators found it more difficult to label causal event pairs when the events are further apart (e.g., for a distance of 3 or higher). This however, can lower the inter-annotator agreement. We also noticed that in case of event pairs with events appearing in different documents, the annotators found it more difficult to label the instances due to the time needed to identify and search the relevant documents. Thus, although in this research we focused only on causal events occurring in the same document, this issue is worth exploring in future developments.

We hope that all these issues are considered as food-for-though by researchers interested in this problem.

CHAPTER 5

FUTURE WORK

5.1 Introduction

This chapter discusses possibilities of future research issues regarding causality learning problem. Causality learning problem is very challenging and it requires lots of research and NLP resources to overcome complexities in detecting causal relationships. This thesis provides an unsupervised approach to handle this problem but there are lots of research questions raised during experiments and evaluation of this system which makes this problem even more challenging.

5.2 Future Research

The presented results and observations show that it is possible to obtain a good performance on this task without a deep context analysis, although causal contexts also play an important role and their consideration may lead to better performance. Specifically, the problem of identifying inter-sentential causal information is very hard because it requires a deep discourse and temporal analysis. Our approach based on scenario-specific events allows us to analyze and identify causal relationships between strongly related events. We noticed that scenario-specific events tend to be strongly related and our approach can capture the information flow through sequences of scenario-specific events within a scenario. Such set of events are good candidates for causal

semantic relationships. Our approach identifies scenarios and effectively generates suitable events pair candidates for the causality detection task. Thus, this approach based on scenario-related events generation reduces the chance for noisy relationships.

One important feature of our approach is the use of the Effect-Control-ratio metric which effectively assigns Cause and Effect roles with high accuracy (cf. Table 4.5 and 4.6). Various causality detection statistical approaches are bound to work on temporal data since in temporal data it is easier to identify as causal those events which are temporally ordered [1, 24]. However, events in raw text data are not always temporally ordered, and thus systems need to rely on temporal classifiers or some deep semantic analysis. Instead, our statistical measure of Effect-Control-ratio helps identify roles without the need for deep semantic and temporal analysis. This approach also provides insights about how far we can go in identifying causal information without relying on a deep causal context analysis. We believe that this research has the potential to open up new avenues of research which, although challenging they are important for text analysis:

- How important is context in detecting causal relationships?
- How complex is this text understanding task and how difficult it is for human annotators to annotate causal relationships with high agreement specifically when the events are further apart (e.g., at distance 4 or more)?
- What is the best way to capture and evaluate relationships between scenarios such that we avoid the inclusion of noisy relationships as well as missing interesting causal relationships?
- Statistical measures are good at finding strong dependencies between events. However, in natural language text strong dependencies can indicate a number of

semantic relationships (e.g., elaboration, etc). What kind of semantic analysis is required to differentiate among these relationships?

- How can we improve the statistical measures such that we can take context into account?
- We observed that some domains (e.g. Iraq War) are harder to annotate as compared with simple domains (e.g. Hurricane Katrina). This requires us to educate annotators to have background knowledge about complex domains to achieve better evaluation possible.

REFERENCES

[1] Beamer, B. and Girju, R. 2009. Using a Bigram Event Model to Predict Causal Potential. In *Proceedings of the 10th international Conference on Computational Linguistics and intelligent Text Processing* (Mexico City, Mexico, March 01 - 07, 2009).

[2] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (Mar. 2003), 993-1022.

[3] Brin, S., Motwani, R., and Silverstein, C. 1997. Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international Conference on Management of Data* (Tucson, Arizona, United States, May 11 - 15, 1997).

[4] Caramazza, A., Grober, E., Garvey, C., and Yates, J. 1977. Comprehension of anaphoric pronouns. In *Journal of Verbal Learning and Verbal Behavior*, 1977.

[5] Chambers, N. and Jurafsky, D. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii, October 25 - 27, 2008). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 698-706.

[6] Chang, D. and Choi, K. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Inf. Process. Manage.* 42, 3 (May. 2006), 662-678.

[7] Cooper, G. F. 1997. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Min. Knowl. Discov.* 1, 2 (Jan. 1997), 203-224.

[8] Cresswell, M. 1981. Adverbs of Causation. In *Words, Worlds, and Contexts: New Approaches in Word Semantics, Eikmeyer, Rieser (eds),* 1981.

[9] Girju, R. 2003. Automatic detection of causal relations for Question Answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12* (Sapporo, Japan). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 76-83.

[10] Girju, R. and Moldovan, Dan. 2002. Mining Answers for Causation Questions. In *AAAI symposium on mining answers from texts and knowledge bases,* 2002.

[11] Graesser, A.,Millis, K., and Zwaan, R. 1997. Discourse Comprehension. In Annual Review of Psychology, 48, 1997, 163-189.

[12] Granger, C. W. 2001. Investigating causal relations by econometric models and cross-spectral methods. In *Essays in Econometrics: Collected Papers of Clive W. J. Granger*, E. Ghysels, N. R. Swanson, and M. W. Watson, Eds. Causality, Integration And Cointegration, And Long Memory, vol. II. Harvard University Press, Cambridge, MA, 31-47.

[13] Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international Conference on Management of Data* (Dallas, Texas, United States, May 15 - 18, 2000). SIGMOD '00. ACM, New York, NY, 1-12.

[14] Hofmann, T. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval Berkeley, California, pages: 50-57, ACM Press,* 1999.

[15] Li, W. and McCallum, A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international Conference on Machine Learning* (Pittsburgh, Pennsylvania, June 25 - 29, 2006). ICML '06, vol. 148. ACM, New York, NY, 577-584.

[16] Menzies, P. 2008. Counterfactual Theories of Causation. In *The Online Encyclopedia of Philosophy.*

[17] Mimno, D., Li, W., and McCallum, A. 2007. Mixtures of hierarchical topics with Pachinko allocation. In *Proceedings of the 24th international Conference on Machine Learning* (Corvalis, Oregon, June 20 - 24, 2007). Z. Ghahramani, Ed. ICML '07, vol. 227. ACM, New York, NY, 633-640.

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

[19] Pitler, E., and Nenkova, A. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009*.

[20] Pitler, E., Raghupathy, W., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. 2008. Easily Identifiable Discourse Relations. In *Proceedings of the COLING 2008*. Poster paper.

[21] Quinlan, J. R. 1993 *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[22] Silverstein, C., Brin, S., Motwani, R., and Ullman, J. 2000. Scalable Techniques for Mining Causal Structures. *Data Min. Knowl. Discov.* 4, 2-3 (Jul. 2000), 163-192.

[23] Sporleder, C. and Lascarides, A. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment.*Nat. Lang. Eng.* 14, 3 (Jul. 2008), 369-416.

[24] Sun, Y., Xie, K., Liu, N., Yan, S., Zhang, B., and Chen, Z. 2007. Causal relation of queries from temporal logs. In*Proceedings of the 16th international Conference on World Wide Web*, pages 1141-1142.

[25] Surdeanu, M. and Turmo, J. 2005. S*emantic Role Labeling Using Complete Syntactic Analysis*. Proceedings of CoNLL 2005 Shared Task, June 2005.

[26] Woodward, J. 2008. Causation and Manipulation. In *The Online Encyclopedia of Philosophy*

[27] Word Net
http://wordnet.princeton.edu