MINING NOVEL SOURCES OF KNOWLEDGE TO IDENTIFY CAUSAL INFORMATION IN TEXT

BY

MEHWISH RIAZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Roxana Girju, Chair
Professor ChengXiang Zhai
Assistant Professor Julia Hockenmaier
Associate Professor Barbara Di Eugenio, University of Illinois at Chicago

Abstract

The abundance of information on the internet has impacted the lives of people to a great extent. People take advantage of the internet to acquire information for several day to day social and political activities. Though the plenty of information on the internet is of great use, it takes lot of time to go through a number of text articles to understand events and the causal relations between events that build a particular social or political news story. In this thesis, we focus on the problem of automated extraction of causal information in text. This can be of great assistance to the people who strive to acquire the flow of events in text to make various decisions and predict consequences of their decisions.

In natural language, causal relations can be encoded using various linguistic constructions. Each construction with its own semantics can pose various challenges for the problem of identifying causality. In this thesis, we address the tasks of identifying causality between two verbs and a verb and a noun by deeply analyzing semantics of these constructions. After the successful use of linguistic features for various Natural Language Processing (NLP) tasks, several approaches have been proposed to identify causality using such features in the framework of supervised learning. However, it is not practical to depend merely on these features because there are many factors involved in identifying causality such as background knowledge, semantic and pragmatic features of events, world knowledge, etc. In addition to the above, the supervised learning approaches are sensitive to the size of training corpus and the type of contexts of training instances. For example, the unambiguous training instances do not provide a better supervision for the ambiguous and implicit instances of semantic relations including causality [Sporleder and Lascarides 2008]. Therefore, in this work instead of merely relying on the linguistic features extracted from the contexts of training instances, we propose an approach to derive novel sources of knowledge for identifying causal information in text. In the first part of this thesis, we introduce methods to acquire background knowledge and the knowledge of causal semantics of verbs for the task of identifying causality between the two state of affairs represented by verbs. After the knowledge acquisition step, we integrate the above types of knowledge with a supervised classifier employing linguistic features to obtain optimal predictions for the current task. Similarly, in the second part of this thesis, we propose methods to acquire and employ the knowledge of causal semantics

of nouns, verbs and verb frames to identify causality between the two state of affairs represented by verbs and nouns. With the addition of novel sources of knowledge, our models for the current tasks gain lots of progress in performance over the baseline of supervised classifiers relying merely on linguistics features. Moreover, in comparison with these supervised classifiers, performance of our models is more robust on all types of context –i.e., unambiguous, ambiguous and implicit contexts.

To my parents and husband.

Acknowledgments

I strongly believe that my PhD dissertation would not have been completed without the guidance of my adviser Roxana Girju. I want to start with expressing my deepest gratitude to my adviser for her support and encouragement throughout my graduate studies. During initial days of my graduate studies, she was the one who guided me to develop ideas to deal with various challenging research problems. In order enhance my research skills, she advised me on which courses should I take to acquire in-depth knowledge of Natural Language Processing (NLP). She always encouraged me to go deep into the natural language semantics to address the topic of my PhD thesis research and other research areas in NLP. Moreover, Roxana helped me a lot in developing well written research papers.

I owe special thanks to my PhD committee members Prof. ChengXiang Zhai, Prof. Julia Hockenmaier and Prof. Barbara Di Eugenio for their valuable suggestions on my research work. I got opportunity to discuss my research with them during and after my prelim exam. I found those discussions very effective in improving my thesis research.

I also want to convey thanks to the Semantic Frontiers research group members Brandon Beamer, Michael Paul, and Rania Al-Sabbagh supervised by my adviser. My discussion with Brandon Beamer on the topic of causality was quite helpful for me in understanding the notion of causality and developing an approach for identifying causality in text. In general our research group's discussions on various topics of NLP helped me a lot in developing research skills.

Some part of my PhD studies was financially supported in terms of full and partial Sohaib and Sara Abbasi Computer Science Fellowships. I am highly obliged to Sohaib Abbasi, CEO of Informatica, and his wife Sara Abbasi for their generous financial support for me.

Finally, I am truly grateful to my parents, husband, sister and brother for their love and support required to achieve various milestones of graduate studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The current era with abundance of information on the internet has impacted the lives of people to a great extent. People take advantage of the internet to acquire information for their several day to day activities ranging from finding the location of a shopping center to getting news on the probe for the missing Malaysian airplane. Though the plenty of information available on the internet is of great use, it takes lot of time and energy to manually process this unstructured information to reach to a point of interest. For example, a person may need to go through a number of news articles to understand the flow of events that build a particular social or political story. Let us consider Figure 1.1 where the stories of Egyptian Revolution of 2011 and Hurricane Katrina extracted from the news articles are characterized by a number of events in causal relations. For example, for these two stories we observe the following causal relations taken from Figure 1.1:

- **Egyptian Revolution of 2011:** Hosni Mubarak's son was expected to succeed his father as the next president of Egypt. Due to this show of inheritance of power, the political groups expressed opposition and millions of people demanded the overthrow of the Mubarak's government. And as a result of protests, the Egypt's parliament was dissolved on 13 February 2011.
- **Hurricane Katrina:** The levee system catastrophically failed in New Orleans which led to the flooding of New Orleans. As a result of the flooding event, a significant number of people died in New Orleans.

A real challenge for current natural language understanding systems is to process text effectively to identify events and the pairs or chains of events encoding causality. In this thesis, we focus on the problem of identifying causal information in text. The automated extraction and presentation of events encoding causality can be of great assistance to the people who strive to acquire the flow of events in text to make various decisions and predict consequences of their decisions. Moreover, in natural language processing, the recognition of semantic relations including causality has always been considered an important topic of research because success in this area is critical for various language processing applications such question answering, document summarization, generation of coherent ordering of events and prediction of future, etc.

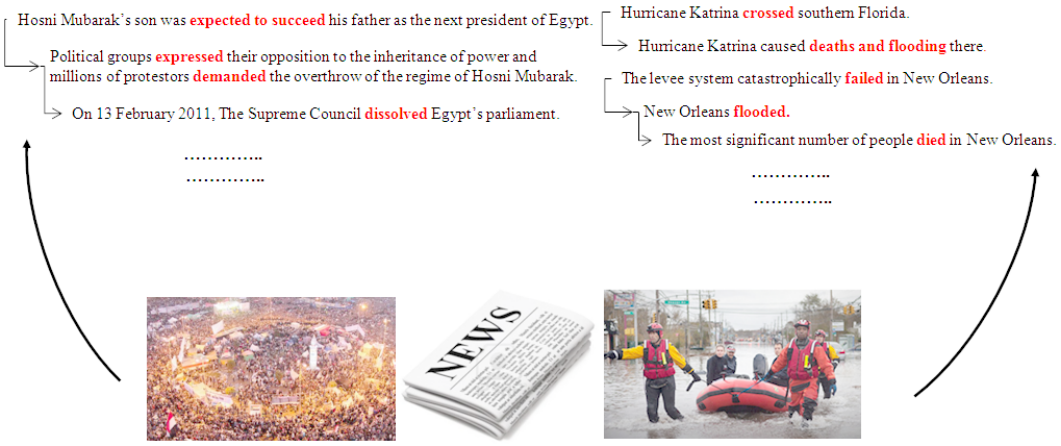

Figure 1.1: A number of events encoding causality in the Egyptian Revolution of 2011 and Hurricane Katrina.

[Girju 2003, Chklovski and Pantel 2004, Barzilay et al. 2002, Radinsky and Horvitz 2013] For example, a question answering system should rely on a model for identifying causality to answer the following question:

Why did American Airlines cancel the last flight from New York to Florida on Tuesday?

Similarly, in order to generate a coherent summary of document(s), it is important to know the order of events i.e., which event happens before another event or is caused by another event [Barzilay et al. 2002, Chklovski and Pantel 2004].

Before discussing the task of identifying causality in text, we need to formally define the cause-effect (or causal) relations. Given various perspectives involved in describing a cause-effect relation, there is a number of notions of causality available to us from the areas of logic, philosophy, statistics, economy, computer science, etc [Menzies 2008, Woodward 2008, Suppes 1970, Granger 1969, Pearl 2000]. Just to give a flavor, three notable notions of causality are as follows:

- **Counterfactual Theory of Causality:** Menzies's counterfactual theory of causality determines truth of the following three conditions to determine if a cause-effect relation is encoded between the two events **a** and **b** or not: (1) event **a** must temporally precede or overlap event **b** in time, (2) the effect event **b** must wholly depend on the cause event **a** and (3) if event **a** has not taken place then **b** must also have not taken place [Menzies 2008, Beamer and Girju 2009]. This notion has a shortcoming that the condition (2) above does not consider the case when an event has more than one cause [Beamer and Girju 2009].

- **Manipulation Theory of Causality:** Woodward's manipulation theory of causality determines truth of the following two conditions to determine if a cause-effect relation is encoded between the two

events **a** and **b** or not: (1) event **a** must temporally precede or overlap event **b** in time and (2) while keeping as many state of current affairs constant as possible, modifying event **a** must entail predictably modifying event **b** [Woodward 2008, Beamer and Girju 2009].

- **Probabilistic Theory of Causality:** Suppes (1970) adopted the probabilistic theory of causation to determine if a cause-effect relation is encoded between the two events **a** and **b** or not. According to this theory if **b** is the effect of event **a** then the probability of occurrence of event **b** when **a** has occurred should be greater than the probability of occurrence of event **b** by itself.

In previous work [Beamer and Girju 2009, Riaz and Girju 2010], NLP researchers have relied on manipulation theory of causality to annotate events in text with cause and non-cause relations. Since, this notion of causality has proven to be an objective notion of causality for the NLP annotations task [Beamer and Girju 2009], in this work we rely on this notion to define a cause-effect relation.

1.1 Automated Identification of Causal Information

In natural language, causal relations can be encoded using various constructions e.g., between two verbs, two nouns, a verb and a noun, two discourse segments and even between two documents. Each construction with its own semantics can pose various challenges for the problem of identifying causality [Girju 2003, Girju et al. 2009, Bethard and Martin 2008, Sporleder and Lascarides 2008]. In order to effectively tackle this broad problem it is useful to split this problem into the smaller tasks. Therefore, in this thesis we focus on the tasks of identifying causality between two verbs and a verb and a noun by deeply analyzing the semantics of these constructions. Consider the following examples:

1. Yoga **builds** stamina because you **maintain** your poses for a certain period of time.
2. At least 1,833 people **died** in the **hurricane**.

Here, in example (1) two events represented by the verbs "build" and "maintain" and in example (2) two events represented by the verb "die" and the noun "hurricane" are encoding causal relations. In example (1), two verbs appearing in the same sentence are causally connected by an explicit and unambiguous discourse marker (i.e., "because"). However, in English, not all the discourse markers unambiguously identify causality [Prasad et al. 2008]. For example, Bethard and Martin (2008) proposed a corpus of 1000 causal and non-causal verb-verb pairs where the two verbs are conjoined by the ambiguous marker "and". Consider the following examples taken from this corpus:

3. The man who had brought it in for an estimate had **returned** to collect it and was **waiting** in the hall[1].

4. Indeed , the "art of doctoring" does **contribute** to better health results and **discourages** unwarranted malpractice litigation .

In example (3) "and" can be replaced by the expression "and as a result". However, in example (4) the verb-verb pair conjoined by the marker "and" does not encode a cause relation. The causal relations can also be encoded in implicit contexts - i.e., those where no discourse marker is present. In example (5), two verbs "rage" and "collapse" are appearing in different sentences and there is no causal marker to signal a cause relation between them. According to a rough estimate by Sporleder and Lascarides (2008), around half of the sentences in the British National Corpus (BNC) lack a discourse marker.

5. The monster storm Katrina **raged** ashore along the Gulf Coast Monday morning. There were early reports of buildings **collapsing** along the coast.

In this thesis, we propose a model for identifying causality between two verbs appearing in the same and different sentences of text. Also, two verbs can appear in any context i.e, explicit and unambiguous, ambiguous or implicit context. Therefore, the ambiguity of natural language and the implicit encoding of causality are two of the serious challenges to be tackled for the current task.

Besides the verb-verb constructions, we also focus on the task of identifying causality in verb-noun pairs i.e., identify causality between the state of affairs represented by a verb and a noun. Girju (2003) pointed out that the association of figurative (non-literal) readings with natural language expressions (e.g., nouns) can complicate the task of identifying causality. Consider example (6) where the noun phrase "The United States" refers to an event of "raid in Abbottabad on May 2, 2011" rather than merely referring to the country. Here the association of the figurative reading with "the United States" results in encoding of causality between "the United States" and an event represented by the verb "kill". Note that a country cannot physically kill someone. It is the event of "raid in Abbottabad on May 2, 2011 by the United States" that led to "killing of Osama bin Laden". Similarly in example (7) the noun "Vietnam" (a country) refers to an event of "war in Vietnam" and thus encodes a cause relation with an event represented by the verb "haunt". However, in example (8) the pair "hit-Cuba" encodes a non-cause relation with no figurative reading associated with the noun "Cuba". Therefore a model for identifying causality should know the real reading of a natural language expression to identify causality.

6. **The United States** has **killed** Osama bin Laden and has custody of his body.

[1]The example is taken from Bethard and Martin (2008)

7. Sex, drugs, and **Vietnam** have **haunted** Bill Clinton's campaign[2].

8. Sandy **hit Cuba** as a Category 3 hurricane.

In the next section, we present some research questions which we address in this thesis to devise a better approach for the current tasks.

1.2 Research Questions

Despite the importance of the problem of identifying causal information in text, NLP researchers have not yet got successful in solving this problem with a very high performance. After the successful use of linguistic features (e.g., lexical items, part-of-speech tags, syntactic structures, senses of words, etc.) for various NLP tasks, several approaches have been proposed to identify causality using such features in the framework of supervised learning [Girju 2003, Bethard and Martin 2008, Sporleder and Lascarides 2008, Pitler and Nenkova 2009, Pitler et al. 2009]. On the task of disambiguating the lexico-syntactic patterns of form <Noun Phrase_1 verb Noun Phrase_2 > (e.g., Earthquakes generate tidal waves[3]) for the cause and non-cause relations, Girju (2003) has reported around 80% F-score using a supervised classifier employing linguistic features. The performance of the Girju (2003)'s supervised classifier on the instances of above mentioned lexico-syntactic patterns is very encouraging. However, performance of the supervised classifiers employing linguistic features drops significantly on other types of constructions e.g., for verb-verb pairs or pairs of discourse segments [Bethard and Martin 2008, Sporleder and Lascarides 2008]. On the task of identifying causality in verb-verb pairs where two verbs represent events, Bethard and Martin (2008) achieved 37.1% F-score using a supervised classifier. Bethard and Martin (2008) attributed low performance of their supervised classifier to the lack of enough training data. Since performance of a supervised classifier is sensitive to the size of training corpus, researchers have previously utilized unambiguous discourse markers to automatically acquire a large training corpus of discourse relations such as explanation, result, contrast, etc. [Sporleder and Lascarides 2008, Marcu and Echihabi 2002]. For example, the unambiguous marker "because" ("although") represents explanation (contrast) relation, respectively. Using a large training corpus, Sporleder and Lascarides (2008) performed a detailed analysis of the supervised classifiers for the task of identifying discourse relations including causality. Although a large number of training instances of discourse relations can be acquired effortlessly using explicit and unambiguous discourse markers, these instances are representative of only one type of context –i.e., explicit and unambiguous

[2]The example is taken from Markert and Nissim (2009).
[3]The example is taken from Girju (2003).

context. Sporleder and Lascarides (2008) showed that even with supervision from the millions of training instances of unambiguous contexts, a supervised classifier brings a very low performance for identifying discourse relations in the ambiguous and implicit contexts [Sporleder and Lascarides 2008]. Later, Pitler et al. (2009) took advantage of the manual annotations of implicit discourse relations in Penn Discourse TreeBank (PDTB) [Prasad et al. 2008] to build a supervised classification model for implicit discourse relations. They compared their model with the baseline of a supervised classifier trained using automatically generated training corpus with the help of unambiguous discourse markers. With a best combination of features, Pitler et al. (2009)'s supervised classifier achieved 16% improvement in F-score for identifying implicit contingency (causal) discourse relations over the baseline. Unfortunately, to the best of our knowledge there is no big manually generated training corpus available for the task of identifying causality in verb-verb pairs. Our current research is motivated by the above stated observations of previous researchers [Bethard and Martin 2008, Sporleder and Lascarides 2008, Pitler et al. 2009] regarding the supervised classifiers. In this work, we study the research question of how to build a model with the capability to identify causality in all types of contexts (i.e., explicit and unambiguous, ambiguous and implicit contexts) with a better performance. Since there is no large training corpus of verb-verb pairs available to us, we also take advantage of explicit and unambiguous discourse markers to automatically acquire a training corpus for the current task. Using this training corpus, we build a supervised classifier for verb-verb pairs employing the linguistic features introduced in Bethard and Martin (2008) and some features we introduce in this work. Moreover, in order to perform well on all types of context, we propose to incorporate additional sources of knowledge such as background knowledge on top of the supervised classifier for verb-verb pairs. In this thesis, we show that our model brings lot of improvement in performance over the supervised classifier for the current task. In addition to this, performance of our model remains robust on all types of contexts in comparison with the above stated supervised classifier.

The supervised classifiers [Girju 2003, Bethard and Martin 2008, Sporleder and Lascarides 2008, Pitler et al. 2009] introduced earlier in this section mainly depend on linguistic features to identify semantic relations including causality. Although these features provide useful knowledge about the context of the sentence(s), humans also make use of other information such as background knowledge to comprehend causality. The complexity of the task of detecting causality stems from the fact that there are many factors involved, such as linguistic features of an instance, background knowledge, semantic and pragmatic features of events, world knowledge, etc. Therefore, a model employing the knowledge sources other than linguistic features is critically needed to achieve progress on the current tasks [Girju 2003, Riaz and Girju 2013]. In this work, for the task of identifying causality in verb-verb pairs we acquire background knowledge in the form of causal

associations of these pairs. We learn these causal associations using a large collection of unlabeled text. In previous work [Riaz and Girju 2010, Do et al. 2011], researchers have proposed methods to learn causal associations between two events in an unsupervised fashion. These events can be represented by either verbs or nouns. Do (2012) used causal associations between events as a source of background knowledge to identify causality. Riaz and Girju (2010) and Do et al. (2011) learned these causal associations by relying on the probabilities of co-occurrences of events acquired through a large collection of unlabeled text. The metric Cause-Effect Association proposed by Do et al. (2011) identifies causal associations mainly using PointWise Mutual Information (PMI) scores for the event predicates and the arguments of these predicates. In this work, through the empirical analysis of this metric we observe that the PMI scores employing the probabilities of co-occurrences of events do not distinguish well causality from any other type of correlation. Considering this observation, we study the question of how to devise the methods capable to distinguish causal associations from any other type of correlation to provide a better source of background knowledge. In this work, we propose a novel method in which the prior information of causal associations is acquired by mainly relying on the probabilities of co-occurrences of verb-verb pairs and then this information is improved by considering supervision from a large training corpus of cause and non-cause relations. In addition to this, our method takes care of unambiguous as well as ambiguous and implicit contexts of verb-verb pairs to derive a better source of background knowledge as compared to the previously proposed approach of Do et al. (2011).

Do et al. (2011) have previously studied the task of identifying causality in verb-noun pairs but they focused only on a small list of predefined nouns representing events. In this work, we introduce a supervised classifier by employing a set of linguistic features for the current task. This supervised classifier depends merely on linguistic features and requires more sources of knowledge to identify causality with a better performance. For example, Girju and Moldovan (2002) observed that noun phrases must represent events, conditions, states, phenomena, processes or facts in order to encode causation in the lexico-syntatic patterns of form <Noun Phrase$_1$ verb Noun Phrase$_2$ > (or <NP$_1$ verb NP$_2$ >). The supervised classifier lacking this information about nouns can lead to lot of noise in predictions. Girju and Moldovan (2002) employed the following 5 senses *human action, phenomenon, state, psychological feature* and *event* of WordNet to identify the semantics of nouns. They argued the pattern <NP$_1$ verb NP$_2$ > has the highest tendency to encode causation if all senses of the head noun of each noun phrase lie in the semantic hierarchies originated by the above 5 senses. Notice that this approach works for the relatively unambiguous nouns with all of their senses lying in the above mentioned 5 senses of WordNet. Later Girju (2003) used 9 noun hierarchies of WordNet i.e., *entity, psychological feature, abstraction, state, event, act, group, possession* and *phenomenon*

as features of a supervised classifier for identifying causality in the lexico-syntatic patterns of form <NP_1 verb NP_2 >. For each of two noun phrases, the above 9 noun hierarchies are used as features. Each of these features is set to 1 if any sense of the head noun of a noun phrase lies in that hierarchy, otherwise set to 0. On the limited context of patterns <NP_1 verb NP_2 >, Girju (2003) reported a very high F-score (i.e., 80% F-score). However, in this work we observe that these features are not very effective for the verb-noun pairs where a verb and a noun do not need to appear in any specific lexico-syntactic pattern. Therefore for the current task we study the question of how to identify semantics of both ambiguous and unambiguous nouns and use the semantics of nouns to identify causality in verb-noun pairs. In this work, we propose a method which employs the annotations of FrameNet corpus [Baker et al. 1998] for noun phrases to identify tendencies of nouns to encode a cause or non-cause relation. This information is then used in our model to reduce noise in the predictions of a supervised classifier for the verb-noun pairs. In addition to the above, we propose an approach to identify the association of figurative readings with nouns because such readings can alter the meaning of an instance as demonstrated by examples (6) and (7) in the last section.

1.3 Thesis Contributions

As discussed in section 1.1, we study the tasks of identifying causality in verb-verb and verb-noun pairs. In order to perform well on these tasks, we deeply exploit the semantics of verbs, nouns, verb-verb and verb-noun pairs to extract the knowledge useful for the current tasks. Specifically, we propose an approach to acquire the following four types of knowledge for the current tasks: (1) Background Knowledge, (2) Causal Semantics of Verbs, (3) Causal Semantics of Nouns and (4) Causal Semantics of Verb Frames. We integrate the above types of knowledge with the supervised classifiers for the tasks of identifying causality in verb-verb and verb-noun pairs. These supervised classifiers identify causality using linguistic features extracted from the instances of verb-verb and verb-noun pairs. After the acquisition of the above types of knowledge, we allow our models to automatically analyze instances of verb-verb and verb-noun pairs from multiple dimensions – i.e., the dimensions of linguistics features acquired from the contexts of such instances, background knowledge, causal semantics of verbs, nouns and verb frames. We take advantage of the learning and inference framework of Integer Linear Programming (ILP) for NLP [Roth and Yih 2004] to integrate all of the above types of knowledge for the current tasks. Do et al. (2011) have previously used the ILP framework to acquire minimal supervision for the tasks of identifying causality in verb-verb and verb-noun pairs. For example, using the ILP framework, they forced their model to assign a label of non-cause relation to all the verb-verb and verb-noun pairs from the two discourse segments connected by a non-causal marker.

However, in this work we employ the ILP framework to incorporate the above four types of knowledge in our models for identifying causality.

In the rest of this section, we briefly introduce the background knowledge, the knowledge of causal semantics of verbs, nouns and verb frames we acquire and employ for the current tasks.

1.3.1 Background Knowledge

Humans often rely on background knowledge to perform causal reasoning on events. For example (5), our background knowledge allows us to recognize the causal relation between **rage** and **collapse** even when no causal discourse marker is available. In this work we propose a method to learn causal associations in verb-verb pairs by distinguishing well the cause-effect relations from any other type of correlation. In addition to this, our method also takes care of unambiguous as well as ambiguous and implicit contexts of verb-verb pairs to derive a better source of background knowledge. We exploit the information available from a large number of unlabeled instances of verb-verb pairs to mine causal associations in these pairs. We also introduce two novel forms in which the background knowledge can be provided to our model.

We store the information regarding causal associations of verb-verb pairs in a resource named the knowledge base of causal associations (KB_c). In this resource, we keep the scores of likelihood of each verb-verb pair to encode causation. Depending on the likelihood of causality in verb-verb pairs, we create three categories of these pairs: Strongly Causal, Ambiguous and Strongly Non-causal. The category strongly causal (strongly non-causal) contains the verb-verb pairs with the highest (least) likelihood to encode causal relations, respectively. However, the category ambiguous contains the verb-verb pairs which have the tendency to encode both cause and non-cause relations. For example, the pair (kill, arrest) has a high tendency to encode causation irrespective of the context of an instance in which it is used, thereby a good indicator of causality. Unlike this pair, the pair (build, maintain) seems ambiguous because it can encode both cause and non-cause relations depending on the context, as shown by examples 1 (cause) and 9 (non-cause). Thus, a model for identifying causality should have knowledge about which of the verb-verb pairs are strongly causal (non-causal) in nature and for which pairs the context of an instance plays an important role to signal causality. The resource KB_c introduced above provides a rich source of background knowledge for identifying causality in verb-verb pairs [Riaz and Girju 2013, Riaz and Girju 2014b].

9. Republicans had not cut the funds for **maintaining** the levee and **building** up the ecological protections.

1.3.2 Causal Semantics of Verbs

Philospher Jaegwon Kim [Kim 1993] (as cited by Girju and Moldovan (2002)) pointed that the entities representing either causes or effects are often events, but also conditions, states, phenomena, processes, and facts. Considering this observation, Girju and Moldovan (2002) stressed that the noun phrases must represent events, conditions, states, phenomena, processes or facts in order to encode causation in the lexico-syntactic patterns of form <NP_1 verb NP_2 >. Following the previous work of Girju and Moldovan (2002), we assume that in order to identify causality in verb-verb pairs, it is important for a model to know if in an instance of verb-verb pair both verbs represent events or not. In linguistics and computational linguistics, researchers mainly categorize words (or phrases) into two aspectual classes: STATE and EVENT [Dowty 1979, Saeed 1997, Verkuyl 1972, Vendler 1957, Pustejovsky et al. 2003]. STATE describes an unchanging situation over a period of time (e.g., know, love) and an EVENT describes a situation which involves internal structure (e.g., run has an internal structure of raising a foot in air, moving it forward and putting it down on floor[4]). In NLP, several approaches have been proposed to categorize verbs into the classes of EVENT and STATES [Bethard and Martin 2006, Sauri et al. 2005]. Benefiting from these approaches we identify if both verbs of an instance of verb-verb pair represent events or not and provide this information to our model for identifying causality.

In addition to the above, we also incorporate the semantic classes of events to learn the causal semantics of verbs. Verbs play a key role in language to represent events and these events can be of various semantic types. For example, for the TimeBank's corpus, Pustejovsky et al., (2003) have categorized the instances of verbal events into seven semantic classes – i.e., OCCURRENCE, PERCEPTION, ASPECTUAL, STATE, I_STATE, I_ACTION and REPORTING. Based on the definitions of these classes [Pustejovsky et al. 2003], we argue that each verb can have its own causal semantics depending on the class of event it represents. For example, the reporting verbal events (i.e., events represented by the verbs such as say, tell, etc.) merely explain the action of a person or discuss another event instead of encoding causality with it. In order to support our argument we propose a data intensive method to identify tendencies of each of the above stated semantic classes and provide this information to our model for identifying causality. We acquire and employ the knowledge of causal semantics of verbs for identifying causality in verb-verb and verb-noun pairs [Riaz and Girju 2014a, Riaz and Girju 2014b, Riaz and Girju 2014c].The empirical evaluation of our models in this thesis reveals that this type of knowledge is quite useful for identifying causality in both verb-verb and verb-noun pairs.

[4]The examples of STATES and EVENTS are taken from Bethard and Martin (2006).

1.3.3 Causal Semantics of Nouns

As discussed in section 1.2, in order to identify causality in verb-noun pairs it is important for our model to have the knowledge of semantics of nouns. In this work, we propose a method which automatically acquires the semantic classes of nouns with a high and low tendency to encode causal relations. In our model, the information about these semantic classes of nouns is referred to as the knowledge of causal semantics of nouns. We use this type of knowledge to filter false positives. For example, a named entity such as LOCATION or any noun expression representing a location may not encode a causal relation unless a figurative reading (or metonymy) is associated with it. Therefore, in addition to acquiring the above stated semantic classes of nouns, we identify the association of metonymies with the nouns (or noun phrases) by using an approach proposed in this thesis. The empirical analysis of our model reveals that the information about the semantic classes of nouns helps reduce lots of false positives. Moreover, the information of metonymies boosts performance of our model by knowing which nouns are being used in literal or non-literal (figurative) sense in the instances of verb-noun pairs. We leverage the FrameNet annotations [Baker et al. 1998] to identify the semantic classes of nouns and the association of metonymies with nouns [Riaz and Girju 2014a, Riaz and Girju 2014c].

1.3.4 Causal Semantics of Verb Frames

We identify the causal semantics of verb frames and use this knowledge to identify causality in verb-noun pairs [Riaz and Girju 2014c]. In order to understand this type of knowledge, consider the following two examples[5]:

10. **The Great Storm of October 1987** almost totally **destroyed** the eighty year old pinetum at Nymans Garden in Sussex.

11. **The explosion occurred** in the city's main business area.

The above two examples show that the verbs "destroy" and "occur" have their own tendencies to encode causation with their subjects. Particularly, in above examples the verb frames of form {destroy, subject} and {occur, subject} encode a cause and non-cause relation, respectively. In this work, we leverage the annotations of verbs in FrameNet [Baker et al. 1998] to identify tendencies of verb frames to encode causation. In addition to above, we also determine the likelihood of a subject (or any grammatical relation) of any verb to encode causation with its verb.

[5]The examples are taken from the FrameNet corpus [Baker et al. 1998]

1.4 Thesis Organization:

The rest of this thesis is organized as follow:

- Chapter 2 provides an overview of the previous work to identify causal information in text. In this chapter, we provide details of the models proposed earlier to identify causal information in text. We also discuss approaches proposed in NLP for the tasks of identifying event mentions, semantic classes of events and temporal relations. These tasks are related to our current research for identifying causality.

- Chapter 3 presents our approach for acquiring knowledge for the task of identifying causality in verb-verb pairs. We propose methods to derive the background knowledge and the knowledge of causal semantics of verbs. In this chapter, we first introduce a novel method to acquire a training corpus of verb-verb pairs. This training corpus is employed to build a supervised classifier. On top of this classifier, the additional sources of knowledge are added to achieve progress on the current task.

- Chapter 4 provides the details of our model for identifying causality in verb-verb pairs. In this model, we integrate various types of knowledge necessary for the current task. This chapter also provides a detailed experimental study for assessing performance of our model with addition of each type of knowledge acquired for the current task.

- Chapter 5 discusses the process of knowledge acquisition for the task of identifying causality in verb-noun pairs. In this chapter, we first introduce a supervised classifier for the current task and then propose methods to derive the knowledge of causal semantics of nouns, verbs and verb frames. We also introduce our methods to identify the association of metonymies with the noun phrases. In addition to above, we determine if a verb and a noun represents the same or distinct state of affairs to help making better predictions for the current task.

- Chapter 6 provides the details of our model for identifying causality in verb-noun pairs. In this model, we incorporate various types of knowledge to identify causality. The experimental study provided in this chapter presents the contribution of each type of knowledge towards solving the current task.

- Chapter 7 concludes the current research by summarizing the current work and identifying the future research directions to achieve more on the problem of identifying causality.

Chapter 2

Relevant Work

Causation has long been studied from various perspectives by philosophers, logicians, statisticians, linguistics, bio scientists, data-mining researchers and computer scientists [Menzies 2008, Woodward 2008, Suppes 1970, Sanders et al. 1992, Cooper 1997, Silverstein et al. 2000, Pearl 2000]. In this chapter, we present previous work by focusing on the research done in natural language processing to identify causality in text. In addition to this, we discuss research for the related tasks of identifying event mentions, semantic classes of events and the temporal ordering between events.

2.1 Identification of Causality in Natural Language Text

The natural language provides a rich set of linguistic constructions to express causality. Girju and Moldovan (2002) have provided a comprehensive overview of explicit causative constructions (e.g., causal connectives, causative verbs, conditionals, causative adverbs and adjectives) and implicit causative constructions (e.g., complex nominals, implicit causality of verbs and discourse relations) for the English Language. In NLP, various approaches have been proposed to identify causal information in text by considering specific types of constructions e.g., lexico-syntactic patterns [Girju 2003, Chklovski and Pantel 2004, Khoo et al. 2000], verb-verb pairs [Bethard and Martin 2008, Beamer and Girju 2009, Riaz and Girju 2010, Do et al. 2011], noun-noun pairs [Girju 2003, Chang and Choi 2006, Girju et al. 2009], verb-noun pairs [Do et al. 2011] and pairs of discourse segments [Marcu and Echihabi 2002, Sporleder and Lascarides 2008, Pitler and Nenkova 2009, Pitler et al. 2009].

In NLP, several approaches have been proposed earlier for identifying causality using the supervised, unsupervised and minimally supervised learning frameworks. The bulk of research using the supervised learning framework depends on the linguistic features extracted from the contexts of training instances [Girju 2003, Chang and Choi 2006, Bethard and Martin 2008]. For example, Girju (2003) proposed to disambiguate the lexico-syntatic patterns of the form <NP_1 verb NP_2 > (e.g., Earthquakes generate tidal

waves[1]) for cause and non-cause relations using linguistic features. The features she employed for her model are head nouns of the noun phrases, the verb (or words of the verbal phrase) and the WordNet's senses for head nouns of the noun phrases. She used 9 noun hierarchies from WordNet (i.e., entity, psychological feature, abstraction, state, event, act, group, possession, phenomenon) as features to obtain the semantics of each of two noun phrases of the pattern <NP_1 verb NP_2>. For example, each of these 9 features is set to true for a noun phrase only if any sense of its head noun lies in that semantic hierarchy. Using these features, her model achieved 73.91% precision and 88.69% recall (80.60% F-score) for the disambiguation of the above stated patterns – i.e., <NP verb NP> where the verb belongs to a class of 60 verbs. These 60 verbs are semantically similar to the causative verb "cause". Though her approach achieves a quite better performance, it is yet to be determined how this approach scales up for the patterns with verbs not belonging to the above mentioned class of 60 verbs.

Another example of supervised learning models is the discriminative classification model of Bethard and Martin (2008) for identifying causality in verb-verb pairs. For their model, they used a set of linguistic features (e.g., verbs, words of the verb phrases, part-of-speech tags of verbs, etc.) to set up a supervised classifier using SVM classification algorithm. In order to generate the training and evaluation data for the verb-verb pairs, they focused only on a simple linguistic structure in which the two verbs of each instance should be conjoined by the marker "and". They generated a set of 1000 instances annotated with causality and temporal relations (BEFORE, AFTER and OVERLAP). They employed 697 (303) instances for training (evaluation) purpose, respectively. Using the above linguistic features, they reported 27% precision and 59.4% recall (37.1% F-score) on this problem. Though their model did not achieve a very high performance, their analysis of results brought important insights into this problem. For example, Bethard and Martin (2008) linked the low precision of their model with the lack of training data by showing the direct relationship in the steady improvement of precision and the percentage of verbs seen during training. Considering the issue of small size of a training corpus, we introduce a novel method in this work to automatically generate a large training corpus of verb-verb pairs encoding cause and non-cause relations. This saves us from the trouble of annotating verb-verb pairs to acquire a training corpus. Moreover, it allows us to focus more on the task of acquiring additional sources of knowledge other than linguistic features to achieve progress in performance for identifying causality. Another interesting observation from the work of Bethard and Martin (2008) is the improvement of 15% in F-score with the addition of features of the gold-standard labels for temporal relations between two verbs. Although the gold-standard labels for temporal relations boost the performance of their model, it is not always possible to acquire such features for any real data set. Moreover, the current classifiers

[1]The example is taken from Girju (2003).

for temporal relations are quite far from achieving a performance close to the humans [Mani et al. 2006, Chambers et al. 2007, Bethard and Martin 2008, Bethard et al. 2007, Do and Roth 2012].

In recent years, researchers have shifted their attention from the supervised identification of causality and have employed the unsupervised metrics and the minimal supervision for this problem [Beamer and Girju 2009, Riaz and Girju 2010, Do et al. 2011]. For example, Riaz and Girju (2010) proposed an unsupervised metric, Effect-Control Dependency (ECD), to identify cause-effect relations between two events of various news scenarios. Do et al. (2011) later introduced an improved version of the metric ECD. Their metric known as Cause-Effect Association (CEA) depends on Pointwise Mutual Information (PMI) and a component of the metric ECD to predict causal relations in verb-verb, verb-noun and noun-noun pairs. They studied the noun-noun and verb-noun pairs by considering a small list of predefined nouns representing events. However, in this work, we identify causality in verb-noun pairs where nouns can be of any type. Do et al. (2011) showed that their metric achieves a very high performance in comparison to the metric ECD. Therefore, in this work, we compare the performance of our model with the state-of-the-art CEA metric for the unsupervised identification of causality. Do et al. (2011) also acquired minimal supervised for the current problem by exploiting discourse markers. For example, using the ILP framework, they force their model to assign a label of non-cause relation to all the event-event pairs from the two discourse segments connected by a non-causal marker. Do et al. (2011) have used the ILP framework to acquire minimal supervision. In this work, we also take advantage of this framework to combine the entirely novel types of knowledge as discussed in the section 1.3. Do et al. (2011) evaluated their model on a set of 20 documents and achieved 38.6% F-score using the metric CEA. With the addition of minimal supervision they acquired 3.1% improvement in F-score. On verb-verb pairs, they reported 38.3% F-score and 1-2% improvement in F-score with addition of minimal supervision. Do et al. (2011) discussed that it is very difficult to achieve a higher human inter-annotator agreement on the annotations of cause-effect relations for a highly skewed data set. On their evaluation data set they achieved 58% human inter-annotator agreement on causal relations. This results in around 2-3% causal instances in their evaluation set, respectively.

In NLP, researchers have also proposed models to automatically identify causal relations between two discourse segments. Various theories have been proposed in linguistics to interpret discourse relations e.g., Rhetorical Structure Theory (RST), Discourse Representation Theory (DRT) and Segmented Discourse Representation Theory (SDRT) [Mann and Thompson 1987, Kamp and Reyle 1993, Asher and Lascarides 2003]. In these theories, the term of contingency discourse relations is used to refer to the relations of cause, purpose, explanation and reason, etc. Since the release of the RST corpus [Carlson et al. 2002] and Penn Discourse Treebank (PDTB) [Prasad et al. 2008] with instances of the discourse relations, a number of su-

pervised classification approaches have been proposed to identify the discourse relations in both explicit and implicit contexts [Soricut and Marcu 2003, Pitler and Nenkova 2009, Pitler et al. 2009]. These models depend on the linguistic features e.g., word pairs selected from the two discourse segments, polarities of words, Levin verb classes [Levin 1993], etc. Instead of relying on the manually labeled training corpus, some researchers exploited the unambiguous discourse markers to automatically label a massive number of instances of discourse segments with the relations of Contrast, Explanation, Result, Summary and Continuation, etc. [Marcu and Echihabi 2002, Sporleder and Lascarides 2008]. Using this method, Sporleder and Lascarides (2008) employed a massive training corpus to identify discourse relations in unambiguous, ambiguous and implicit contexts by omitting the discourse markers from the training instances. They observed that their training instances are representative of only explicit and unambiguous context and thus do not provide accurate supervision for the ambiguous and implicit contexts. Due to this reason, the performance of their supervised classifier drops significant on both ambiguous and implicit instances of discourse relations. Inspired by the work of Sporleder and Lascarides (2008), we automatically generate a training corpus of verb-verb pairs to identify causality to avoid the trouble of manual generation of a massive training corpus. In addition to this, we incorporate additional sources of knowledge discussed in section 1.3 to identify causality with a better performance on explicit and unambiguous, ambiguous and implicit contexts.

As discussed in the section 1.3, our objective is to employ the novel sources of knowledge to identify causation. For this purpose, we introduce a method to acquire a resource of background knowledge. We derive the background knowledge in terms of the causal associations of verb-verb pairs. We cannot depend on state-of-the-art resources on verbs semantics, such as WordNet, VerbNet, Levin Verb Classes, etc. [Miller 1990, Levin 1993, Kipper et al. 2000] to acquire the causal associations in verb-verb pairs because these resources mainly provide information about the semantic classes, thematic roles and selectional restrictions of verbs. Among these, WordNet is the only resource which provides information about the cause-effect relations between verbs, but it has a very limited coverage. VerbOcean, a semi-automatically generated resource, with fine-grained relations on verb-verb pairs is relevant to our task but it also has a limited coverage. To generate this resource, Chklovski and Pantel (2004) used explicit lexico-syntactic patterns as means of mining enablement (cause-effect) relations between verbs. For example, they consider the pattern "verb * by verb" to extract enablement relations where the unambiguous marker "by" signals this relation. Such approaches help detecting causality with a high precision but suffer from the limited coverage due to the consideration of only explicit and unambiguous contexts of causality of certain forms. Moreover, the current resources do not provide any information about the likelihood of causal relations encoded by verb-verb pairs. Therefore, we derive our own resource for the causal associations of verb-verb pairs. However, in order to

derive the other types of knowledge i.e., the causal semantics of verbs, nouns and verb frames, we leverage the annotations of the TimeBank and FrameNet corpus [Pustejovsky et al. 2003, Baker et al. 1998]. Using these annotations, we acquire the semantic classes of verbs and nouns with a high and low tendency to encode causation and the tendencies of the verb frames to encode cause and non-cause relations.

2.2 Identification of Events, Semantic Classes and Temporal Relations

In a related line of research to the tasks being considered for this thesis, several models have been proposed for identifying event mentions, their semantic classes and temporal order of events. As described in section 1.3.2, in linguistics and computational linguistics researchers mainly categorize words (or phrases) into two aspectual classes: STATE and EVENT where STATE describes an unchanging situation over a period of time (e.g., know, love) and an EVENT describes a situation involving internal structure (e.g., run). Pustejovsky et al. (2003) closely followed the above definition of events to annotate 8312 instances of TimeBank with labels of event and non-event. However, there are some exceptions in their annotations. For example, they considered some states as events depending on the context. In the example "They **lived** in U.N.-run refugee camps for 2 1/2 years.", the verb "lived" is a state that persists for a long time but Pustejovsky et al. (2003) considered this as an event. Pustejovsky et al. (2003) organized the instances of events into the following seven semantic classes[2]:

- **OCCURRENCE:** These events describe something that happens or occurs in the world. Some examples are launched, exploded, landed.
- **PERCEPTION:** These events involve the physical perception of another event. Some examples are see, watch, view.
- **ASPECTUAL:** These events focus on different facets of an event's history. Some examples are begin, start, commence.
- **STATE:** These events refer to the circumstances in which something holds true for a certain period of time. In the example "They **lived** in U.N.-run refugee camps for 2 1/2 years." the event "lived" is the state that persists for 2 1/2 years.
- **I_STATE:** These events are intentional states that refer to an alternative or possible world (or state of affairs). In the example "Russia now **feels** *the US must hold off at least until UN secretary general*

[2]The definitions of the semantic classes of events are taken from Pustejovsky et al. (2003).

Kofi Annan visits Baghdad.", the event "feels" is the I_STATE event referring to the state of affairs shown in italic.

- **I_ACTION:** These events are intentional actions which introduce an event argument describing an action from which we can infer something given its relation with the I_ACTION. In the example "The Organization of African Unity will **investigate** the Hutu-organized *genocide* of more than 500,000 minority Tutsis.", the event "investigate" is the I_ACTION event which introduces the event "genocide".

- **REPORTING:** These events describe the action of a person or an organization, declare something, narrate another event, etc. Some examples are say, tell, report.

In order to identify events and their semantic classes, researchers [Bethard and Martin 2006, Sauri et al. 2005] have proposed supervised machine learning and rules based models build on the TimeBank corpus [Pustejovsky et al. 2003]. Bethard and Martin (2006) addressed this problem in the framework of word-chunking by assigning labels B-I-O (i.e., B (Beginning of event), I (Inside event) and O (Outside of event)) to words and building a classification model on top of these these labels. Their model also predicts the semantic class of an event by enhancing the semantic class labels with B-I-O tags (e.g., OCCURRENCE label is learned and predicted as B_OCCURRENCE, I_OCCURRENCE). Using linguistic features (e.g., lexical items, morphological features, part-of-speech tags and the syntatic chunk labels), they have reported 88.3% and 70.7% F-scores for the tasks of identifying verbal events and their semantic classes, respectively. In this thesis, we employ the annotations of TimeBank corpus for events and semantic classes of events to learn the knowledge of causal semantics of verbs.

The TimeBank corpus also has annotations for the temporal relations (e.g., IBEFORE, BEGINS, ENDS, SIMULTANEOUS, INCLUDES, BEFORE). Benefiting from these annotations of the temporal relations, researchers have also proposed supervised learning models for identifying such relations [Mani et al. 2006, Chambers et al. 2007]. These models rely either on the gold-standard features or the features automatically learned for this task. Another interesting research in this direction is the automated induction of the narrative event chains sharing a common protagonist. For example, "_accused X $\rightarrow$ X claimed_ $\rightarrow$ X argued $\rightarrow$ dismissed X" is a narrative chain of events for the scenario of "Firing of Employee" [Chambers and Jurafsky 2008, Chambers and Jurafsky 2009]. Chambers and Jurafsky (2008) are perhaps the first who studied the automated induction of such chains. They used distributional similarity metric – Pointwise Mutual Information (PMI) – to build the chain of events by focusing only on those verbs which share arguments with each other. To interpret the order of the events contained in a chain, they applied a classifier to arrange events using BEFORE relation. We believe that in the future these chains of events can

be further refined by considering the causal relations.

Chapter 3

Knowledge Acquisition for Verb-Verb Pairs

To build a model for recognizing causality in verb-verb pairs, it is essential to acquire the knowledge necessary for this task. Unlike the traditional approach relying merely on linguistic features to identify causality, we employ novel sources of knowledge for this task. Specifically, we propose our methods to derive the background knowledge and the knowledge of causal semantics of verbs to achieve progress over the traditional approach. For the current task of identifying causality in verb-verb pairs, we consider only main verbs which normally represent events in text. Before introducing our methods for knowledge acquisition, we need to formally define the notions of a verbal event (i.e., an event represented by a verb) and a causal relation between two events. We follow our previous work [Riaz and Girju 2010] to define a verbal event. A verbal event (denoted by e_{v_i}) is defined as a 3-tuple ($[subject_{v_i}]$, v_i, $[object_{v_i}]$), where v_i is the main verb and the rest of the elements of this 3-tuple are core arguments of the verb v_i i.e., subject and object. These arguments are not always explicitly available in an instance. We assign a NULL value for a missing argument of the 3-tuple. In order to define the causal relations between two events, we rely on the broad notion of causality as adopted by Riaz and Girju (2010). According to this notion, the causal relations between events are broadly seen as contingency semantic relations (cause-consequence, argument-claim, instrument-goal, purpose and reason/explanation). These are different from the additive relations (list, continuation, comparison, opposition, exception, enumeration, temporal, and concession). Employing the above stated notions, our current task is to identify causality on both intra- and inter-sentential instances of the verbal event-event pairs. These two types of instances are defined as follows. In an intra-sentential instance of event-event pair, both events belong to the same sentence (e.g., the pair e_{build}-$e_{maintain}$ from example (1)). In an inter-sentential instance of event-event pair, both events belong to the different sentences of text (e.g., the pair e_{rage}-$e_{collapse}$ from example (2)).

1. Yoga **builds** stamina because you **maintain** your poses for a certain period of time ($e_{maintain} \rightarrow e_{build}$).

2. The monster storm Katrina **raged** ashore along the Gulf Coast Monday morning. There were early

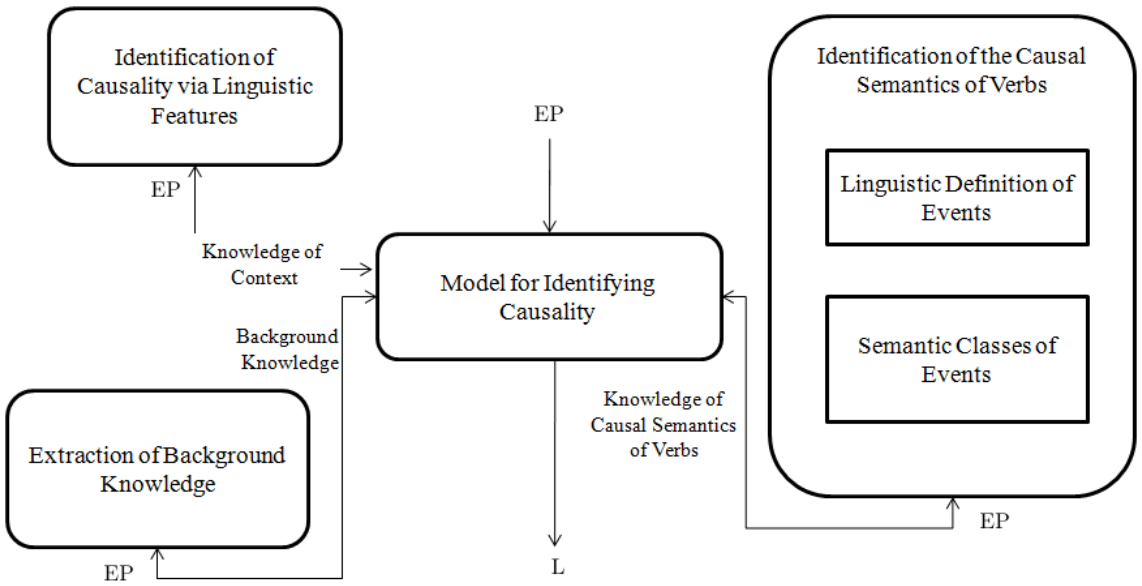

Figure 3.1: A model for identifying causality for a set EP of instances of e_{v_i}-e_{v_j} pairs – i.e., EP = {e_{v_i}-e_{v_j} | e_{v_i} and e_{v_j} are the events represented by the verbs v_i and v_j}. The output of this model is the set L of instances of e_{v_i}-e_{v_j} pairs with the assignments of labels C or $\neg C$ – i.e., L= {(e_{v_i}-e_{v_j}, l) | e_{v_i} and e_{v_j} are the events represented by the verbs v_i and v_j and $l \in \{C, \neg C\}$}.

reports of buildings **collapsing** along the coast ($e_{rage} \rightarrow e_{collapse}$).

In above examples the $\rightarrow$ shows the direction of causality – i.e., in example (1) $e_{maintain}$ causes e_{build} and in example (2) e_{rage} causes $e_{collapse}$.

Figure 3.1 shows the structure of our model for the current task. As shown in Figure 3.1, our model takes as input a set EP of instances of e_{v_i}-e_{v_j} pairs and produces the label Cause (C) or Non-Cause ($\neg C$) on all instances of this set. In our model, the component "Identification of Causality via Linguistic Features" is a supervised classifier which identifies causality by exploiting linguistic features. These features are extracted from the contexts of instances of e_{v_i}-e_{v_j} pairs. This component provides the labels C or $\neg C$ on the instances of e_{v_i}-e_{v_j} pairs and the probabilities of assignments of labels. We use the term "knowledge of context" for the probabilities of assignments of labels. The above supervised classifier serves as a baseline for our model. On top of the knowledge of context, we plug in background knowledge from the component "Extraction of Background Knowledge" and the knowledge of causal semantics of verbs from the component "Identification of the Causal Semantics of Verbs". In order to integrate the above stated types of knowledge, we employ the learning and inference framework of Integer Linear Programming (ILP) for NLP [Roth and Yih 2004]. In this chapter, we discuss our methods for knowledge acquisition which is a prerequisite for setting up an integer linear program introduced in chapter 4 of this thesis.

3.1 Identification of Causality via Linguistic Features

In this section, we present a supervised classifier for the current task. As discussed in section 2.1, Bethard and Martin (2008) proposed a supervised classification model by generating a small scale data set annotated with C and $\neg C$ labels. This data set consists of only 1000 instances of e_{v_i}-e_{v_j} pairs. They employed 697 (303) instances for the training (evaluation) purpose, respectively. Given the various notions of causality, the manual generation of a large enough training corpus for the current task is laborious and time consuming [Bethard 2007]. Bethard and Martin (2008) attributed low performance of their supervised classifier for e_{v_i}-e_{v_j} pairs to the lack of enough training data. Therefore, in this section we begin with proposing a novel procedure to automatically generate a large training corpus of e_{v_i}-e_{v_j} pairs. Using this corpus, we extract a list of linguistic features from the contexts of the training instances to identify causality in e_{v_i}-e_{v_j} pairs.

3.1.1 Acquisition of Training Corpus

Following previous approaches [Marcu and Echihabi 2002, Sporleder and Lascarides 2008], we propose a method to leverage unambiguous discourse markers to acquire a training corpus for building a supervised classifier. For example, the discourse marker "because" in example (1) encodes a causal relation between the verbal events e_{build} and $e_{maintain}$. Previously, Sporleder and Lascarides, (2008) have utilized a number of unambiguous discourse markers to acquire the training instances of semantic relations between the discourse segments. However, the process is not simple for the current task since it is not always clear how to create a causal instance of a verbal event-event pair. Consider the following meta instance I with the discourse marker "because".

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

It is composed of main verbs (v_1, v_2, ..., v_r), discourse markers (m_1, m_2), and sentence boundaries (<s>, </s>). Here, we assume that the discourse markers or the sentence boundaries whichever appear first in I represent the boundaries of the discourse segments for the marker "because"[1]. In the instance I, there are k and $r - k$ main verbs appearing before and after the marker "because", respectively. The problem here is to determine the verbal event-event pair encoding causality out of $k \times (r - k)$ choices. For the following examples, the pair e_{lose}-e_{place} and e_{turn}-e_{focus} encode causal relations out of 2 and 3 available choices, respectively.

[1]We assume that only those markers which have discourse usage in the instance I define the boundaries of the discourse segments. We use the list of 100 explicit discourse markers provided by Prasad et al. (2008) and the supervised learning approach of Pitler and Nenkova (2009) to detect the markers and the discourse versus non-discourse usage of these markers.

3. A Michigan woman **lost** custody of her young daughter *because* she **placed** the child in day care while **attending** college classes.

4. Some of these groups had been **turned** down *because* they were **told** they **focused** only on women's issues, have now been **admitted** on appeal.

Considering the above problem, we need to rely on some approximate solution to prepare the training instances of causal event-event pairs. We assume that the most dependent pair among $k \times (r-k)$ choices in the instance I is the best candidate to encode causality. We propose the following function $f(I)$ to pick the most dependent pair:

$$f(I) = \underset{(\mathrm{v_i} \prec m_c, \mathrm{v_j} \succ m_c)}{\arg\max} CD(\mathrm{v_i}\text{-}\mathrm{v_j}) \times PS_I(\mathrm{v_i}\text{-}\mathrm{v_j}) \tag{3.1}$$

Here, $i(j)$ refers to all verbs that appear before (after) the causal discourse marker (i.e., m_c) (e.g., "because" in the instance I). CD (equation 3.2) is a component of the predicate-predicate association of the metric CEA [Do et al. 2011] to determine causal dependency of a pair $\mathrm{v_i}$-$\mathrm{v_j}$. Do et al. (2011) used CD to determine causality in an unsupervised fashion but here we employ this to build a training corpus of event-event pairs.

$$CD(\mathrm{v_i}\text{-}\mathrm{v_j}) = PMI(\mathrm{v_i}\text{-}\mathrm{v_j}) \times max(\mathrm{v_i}\text{-}\mathrm{v_j}) \times IDF(\mathrm{v_i}\text{-}\mathrm{v_j}) \tag{3.2}$$

The functions PMI (i.e., Pointwise Mutual Information), max and IDF (i.e., Inverse Document Frequency) depend on the probabilities of co-occurrences and idf scores to determine causal dependency [Riaz and Girju 2010, Do et al. 2011]. These functions are defined as follows. PMI (equation 3.3) assumes that two verbs are causally dependent on each other if the probability of co-occurrences of $\mathrm{v_i}$ and $\mathrm{v_j}$ is greater that probability of occurrence of these verbs by themselves i.e., $P(\mathrm{v_i}\text{-}\mathrm{v_j}) > P(\mathrm{v_i})P(\mathrm{v_j})$.

$$PMI(\mathrm{v_i}\text{-}\mathrm{v_j}) = log(\frac{P(\mathrm{v_i}\text{-}\mathrm{v_j})}{P(\mathrm{v_i})P(\mathrm{v_j})}) \tag{3.3}$$

In the above mentioned function $P(\mathrm{v_i}\text{-}\mathrm{v_j})$ is computed by counting the number of instances in which $\mathrm{v_i}$ and $\mathrm{v_j}$ appear together divided by the total number of instances. Notice that $\mathrm{v_i}$ and $\mathrm{v_j}$ can appear in any order in an instance i.e., $\mathrm{v_i}$ can either appear before or after $\mathrm{v_j}$ in text.

The function max (3.4) is the component of ECD metric [Riaz and Girju 2010] which identifies causal dependency in verbs $\mathrm{v_i}$ and $\mathrm{v_j}$ by determining how frequently these verbs appear with each other than with any other verb.

$$max(\text{v}_\text{i}\text{-v}_\text{j}) = \max\{\frac{P(\text{v}_\text{i}\text{-v}_\text{j})}{\max_k[P(\text{v}_\text{i}\text{-v}_\text{k})] - P(\text{v}_\text{i}\text{-v}_\text{j}) + \epsilon}, \frac{P(\text{v}_\text{i}\text{-v}_\text{j})}{\max_k[P(\text{v}_\text{k}\text{-v}_\text{j})] - P(\text{v}_\text{i}\text{-v}_\text{j}) + \epsilon}\} \tag{3.4}$$

Here, a small value $\epsilon = 0.01$ is added to avoid 0 value in the denominator. The first fraction determines how frequently v_i co-occurs with v_j as compared with any other verb v_k with which it occurs most of the times (i.e., $\text{v}_k = max_k[P(\text{v}_i\text{-v}_k)]$). This fraction has maximum value if $\text{v}_j = \text{v}_k$. Similarly, the second fraction is computed for the other direction i.e., how frequently v_j co-occurs with v_i as compared with any other verb v_k with s.t., $\text{v}_k = max_k[P(\text{v}_k\text{-v}_j)]$.

The IDF function (3.5) assumes that the verbs appearing in a large number of documents are less important and less discriminative, thereby frequently encode non-causal relations [Do et al. 2011].

$$\begin{aligned} IDF(\text{v}_\text{i}\text{-v}_\text{j}) &= idf(\text{v}_\text{i}) \times idf(\text{v}_\text{j}) \times idf(\text{v}_\text{i}\text{-v}_\text{j}) \\ idf(p) &= \frac{D}{1+N} \end{aligned} \tag{3.5}$$

Here D is the total number of documents and N is the number of documents in which p occurs. We use a large number of documents and extract a set of unlabeled intra- and inter-sentential instances of verb-verb pairs to compute the functions PMI, max and IDF. This set of intra- and inter-sentential instances is referred to as the development set for our model. In section 3.2, we provide details of this development set. If $P(\text{v}_\text{i}\text{-v}_\text{j}) = 0$ then we do not provide the score CD for the function $f(I)$ (3.1). In this case the function $f(I)$ depends only on the score PS_I to identify the most dependency pair.

Above, we have explained the causal dependency score (CD) to select the most dependent pair using the function $f(I)$. Next, we define a novel penalization factor PS_I (3.6) for the verbs of a pair appearing at greater distance from the causal marker. For example, this assumes that for the instance I, the verbs of the pair $\text{v}_2\text{-v}_{k+2}$ are less likely to be in a cause relation as compared with $\text{v}_k\text{-v}_{k+1}$. We use the penalization factor PS_I to select the most dependent pair because in our previous work [Riaz and Girju 2010] we observed reduction in the likelihood of causality with respect to increase in distance between two events.

$$PS_I(\text{v}_\text{i}\text{-v}_\text{j}) = -\log \frac{pos(\text{v}_\text{i}) + pos(\text{v}_\text{j})}{2.0 \times (C(v_p) + C(v_q))} \tag{3.6}$$

Here, $\text{C}(\text{v}_p)$ ($\text{C}(\text{v}_q)$) is the count of the main verbs appearing before (after) the causal marker (e.g., "because" in the instance I), respectively. The distance of the verb is measured in terms of its position (i.e., $pos(\text{v}_\text{i})$) with respect to the causal marker. The position is 1 for the verb closest to the causal marker

Relation	Discourse Marker
Causal	because, for this (that) reason, consequently, as a consequence of, as a result of
Non-causal	but, in short, in other words, whereas, on the other hand, nevertheless, nonetheless, in spite of, in contrast, however, even, though, despite the fact, conversely, although.

Table 3.1: A list of unambiguous discourse markers employed for the acquisition of a training corpus of e_{v_i}-e_{v_j} pairs encoding causal and non-causal relations.

and 2 for the verb next to the closest verb. For example, for the instance I, $pos(v_k) = 1$, $pos(v_{k+1}) = 1$, $pos(v_2) = 2$ and $pos(v_{k+2}) = 2$ and so on. For the instance I, PS_I has maximum value for the pair v_k-v_{k+1} and it reduces for other pairs with verbs at a greater distance from the causal marker.

Above, we have explained our method to collect the causal training instances. The process for the collection of non-causal training instances is relatively simple. In order to extract the non-causal event-event pairs, we utilize the instances of two discourse segments conjoined by non-causal markers (e.g., "but" which represents comparison (non-causal) relation). Any event-event pair collected from the two discourse segments in non-causal relation encodes non-causality. Therefore, we select the closest verb-verb pair from the instances of the form I with a non-causal marker conjoining the two discourse segments. Table 3.1 shows the complete list of unambiguous discourse markers we employ for the purpose of acquiring a training corpus of e_{v_i}-e_{v_j} pairs. In this thesis, we present performance of a supervised classifier trained using a training corpus of $244,552$ instances (50% for each of C and $\neg C$ labels).

In this research, we also employ the manually annotated training corpus Penn Discourse Tree Bank (PDTB) for our purpose [Prasad et al. 2008]. This corpus provides labels for contingency (causal) and non-contingency (non-causal) relations on the pairs of discourse segments (also known as Elementary Discourse Units (EDU)). For our task, we apply the above stated method to automatically acquire the training instances of e_{v_i}-e_{v_j} pairs from the instances of EDU-EDU pairs of the PDTB corpus. The above method can only be applied on those EDU-EDU pairs where each EDU contains at least one main verb in it and it is a text segment of contiguous words. In order to select the most dependency verb-verb pair from a EDU-EDU pair of contingency relation, we need to calculate the scores CD and PS_I of function $f(I)$ (3.1). It is straightforward to obtain the score of CD. However, in order to calculate the score PS_I, we need to define positions of the main verbs contained in both EDUs. We use three possible structures of EDU-EDU pairs given in Figure 3.2 to define the positions of the main verbs. The first structure with a discourse marker between the two EDUs corresponds to the meta instance I. Therefore, for this structure we employ the scheme stated above to define the positions of the main verbs for the instances of form I. The second structure represents the case where two EDUs either encode a relation in an implicit context or the discourse marker connecting two

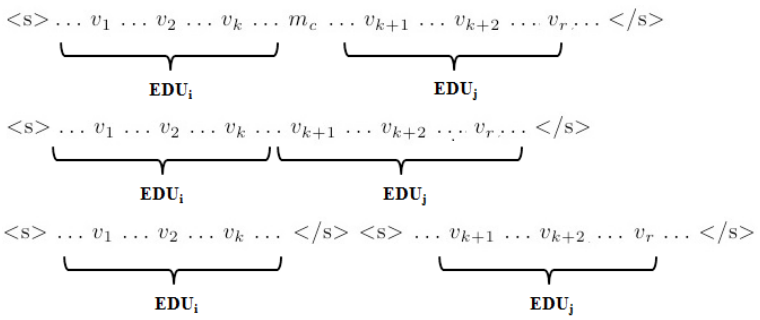

Figure 3.2: The three possible structures of EDU-EDU pairs acquired from the PDTB corpus.

EDUs does not appear between two text segments. The third structure represents the case where two EDUs appear in different sentences. For the last two structures, we assume that the text order has preference i.e., for the last two structures $pos(v_1) = 1$, $pos(v_{k+1}) = 1$, $pos(v_2) = 2$, $pos(v_{k+2}) = 2$ and so on. Using PDTB corpus, we have acquired $5,277$ C and $16,640$ $\neg C$ instances of e_{v_i}-e_{v_j} pairs.

In the rest of this thesis, we use the terms "Explicit$_{e_{v_i}\text{-}e_{v_j}}$" and "PDTB$_{e_{v_i}\text{-}e_{v_j}}$" to refer to the training corpora acquired using the explicit and unambiguous discourse markers and the manually annotated PDTB corpus, respectively. In our experiments, we also employed Bethard and Martin (2008)'s manually annotated training corpus of 1000 e_{v_i}-e_{v_j} pairs. The results with this small training corpus were quite lower as compared to the results achieved using the "Explicit$_{e_{v_i}\text{-}e_{v_j}}$" and "PDTB$_{e_{v_i}\text{-}e_{v_j}}$" corpora. Therefore, in this work we depend on the "Explicit$_{e_{v_i}\text{-}e_{v_j}}$" and "PDTB$_{e_{v_i}\text{-}e_{v_j}}$" corpora to learn and identify causality.

In the next section, we present a list of linguistic features we extract from the training corpus to build a supervised classifier.

3.1.2 Linguistic Features

In this section, we present a list of linguistic features for building a supervised classifier for e_{v_i}-e_{v_j} pairs. We use example (2) (inter-sentential instance) to elaborate the following list of features (see Table 3.2). Among the following list of features, the first two types of features (i.e., Verbs and Verb Phrases) were proposed by Bethard and Martin (2008) and the rest of the features are contributions of this research.

- **Verbs:** words, lemmas, part-of-speech tags [Toutanova et al. 2003] and all senses of both verbs from WordNet. Since we do not know the actual sense of a verb, we use all senses for this type of feature.

- **Verb Phrases:** words, lemmas, part-of-speech tags and all senses of the words of both verb phrases. We take senses from Word for only verbs and nouns. In order to collect the verb phrases, we use Stanford's syntactic parser [Klein and Manning 2003] to acquire the syntactic structure of the sentence(s)

Feature Type	Examples
Verbs	raged, rage, VBD, 3 senses of rage from WordNet (use 3 sense keys from the WordNet), collapsing, collapse, VBG, 7 senses of collapse from WordNet (use 7 sense keys from the WordNet)
Verb Phrases	raged, rage, VBD, 3 senses, ashore, ashore, RB, along, IN, the, DT, gulf, gulf, NNP, 3 senses, coast, coast, NNP, 4 senses, monday, monday, NNP, 1 sense, morning, morning, NN, 4 senses; collapsing, collapse, VBG, 7 senses, along, IN, the, DT, coast, coast, NN, 4 senses
Verb Arguments	subject$_{rage}$: katrina, katrina, NNP, no sense in WordNet, object$_{rage}$ No object; subject$_{collapse}$ No subject, object$_{collapse}$ No object
Verbs and Arguments Pairs	katrina-collapse, rage-collapse
Context Words	the, monster, storm, katrina, rage, ashore, along, the, gulf, coast, monday, morning, there, were, early, report, of, building, collapse, along, the, coast
Context Main Verbs	raged, rage, collapsing, collapse
Context Main Verb Pairs	rage-collapse
Example: The monster storm Katrina **raged** ashore along the Gulf Coast Monday morning. There were early reports of buildings **collapsing** along the coast.
$min_{context}$**:** <s>The monster storm Katrina raged ashore along the Gulf Coast Monday morning</s>. <s>There were early reports of buildings collapsing along the coast<s>.

Table 3.2: The instances of linguistics features employed by the supervised classifier for identifying causality in e_{v_i}-e_{v_j} pairs.

of a e_{v_i}-e_{v_j} pair.

- **Verb Arguments:** words, lemmas, part-of-speech tags and all senses of the subject and object of both verbs. In order to obtain the subject and object, we employ the Stanford dependency parser [Marneffe et al. 2006].

- **Verbs and Arguments Pairs:** For this feature, we take the cross product of both events of a pair e_{v_i}-e_{v_j} where e_{v_i} = [subject$_{v_i}$] v_i [object$_{v_j}$] and e_{v_j} = [subject$_{v_j}$] v_j [object$_{v_j}$]. Some examples of this feature are subject$_{v_i}$-subject$_{v_j}$, subject$_{v_i}$-v_j, subject$_{v_i}$-object$_{v_j}$, etc (see Table 3.2 for the features). In this work, we use unordered pairs as features – i.e., v_i-v_j is same as v_j-v_i because the temporal order of the events is unknown for the unlabeled instances. In the future, this feature can be improved by adding temporal information.

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

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

- **Context Words:** lemmas of all words from the $min_{context}$. This feature captures words other than two events.

- **Context Main Verbs:** all main verbs and their lemmas from the $min_{context}$. It collects information about all verbs that appear with the causal and non-causal event-event pair.

- **Context Main Verb Pairs:** the pairs of main verbs from the $min_{context}$. The lemmas are taken from the feature "Context Main Verbs" and then the pairs on these lemmas are used as this feature. For example, for the lemmas of the verbs $v_1, v_2, \dots, v_k$, the pairs of verbs (i.e., v_1-v_2, v_1-v_k, etc.) are used for this feature. This feature is used to get information about the interesting causal chains of verbs that may appear in the causal instances. For example (3), a chain of causality i.e., e_{attend} causes e_{place} and e_{place} causes e_{lose} can be captured through this feature.

We use both Naive Bayes and Maximum Entropy classifiers to obtain binary predictions for the labels C and $\neg C$ along with their probabilities. These classifiers provide the knowledge of context to our model

which is defined as the probabilities of assignments of labels C and $\neg C$ to the instances of e_{v_i}-e_{v_j} pairs. For Maximum Entropy classification we employ the MALLET toolkit [McCallum 2002] and for Naive Bayes we obtain the probabilities of assignments of labels as follows:

$$\begin{aligned} P(e_{v_i}\text{-}e_{v_j}, C) &= 1.0 - \frac{\sum_{k=1}^{n} logP(f_k \mid C)}{\sum_{k=1}^{n} \sum_{l \in \{C, \neg C\}} logP(f_k \mid l)} \\ P(e_{v_i}\text{-}e_{v_j}, \neg C) &= 1.0 - P(e_{v_i}\text{-}e_{v_j}, C) \end{aligned} \tag{3.7}$$

where f_k is a feature, n is total number of features and $P(f_k \mid l)$ is the smoothed probability of a feature f_k given the training instances of label l.

3.2 Extraction of Background Knowledge

The model for identifying causality can take advantage of background knowledge when the linguistic features do not provide enough information to comprehend causation. In this section, we introduce our method to acquire background knowledge. We derive this knowledge in terms of causal associations of verb-verb pairs. For example a pair "kill-arrest" has a high tendency to encode causation irrespective of the context. On the other hand a pair "produce-create" may have a high tendency to encode non-causation because the verbs "produce" and "create" are nearly synonyms. Such information about the likelihood of verb-verb pairs to encode cause or non-cause relation is used as the source of background knowledge in our model.

In order to acquire causal associations of verb-verb pairs, we extract a set of large number of unlabeled intra- and inter-sentential instances of these pairs. This set is referred to as the development set of our model. This development set is also used to collect the statistics for function $f(I)$ used to prepare a training corpus (see section 3.1.1). We are using around 12,000 documents to extract instances for the development set. We collect instances of verb-verb pairs from the same sentences (intra-sentential) and adjacent sentences (inter-sentential) of the documents as follows. We remove stopwords and retain only main verbs in the sentences. For each sentence, we collect all pairs of main verbs (i.e., (v_i-v_j)) to generate the intra-sentential instances of verb-verb pairs. For each of two adjacent sentences s_l and s_m of a document, we collect all pairs of main verbs i.e., v_i-v_j where v_i (v_j) appears in s_l (s_m), respectively. This results in generation of the inter-sentential instances of verb-verb pairs. We use these intra- and inter-sentential instances to derive the causal associations of verb-verb pairs. To determine causal associations with confidence, we retain only those verb-verb pairs which have at least 30 instances in the development set. Also, we consider only those intra-sentential instances of verb-verb pairs in which two verbs are separated by at least two words. The

verbs appearing close enough may represent the same event e.g., "He **failed** to **defeat** his competitor" where "failed to defeat" is one event with two verbs in it. In the above set of 12000 documents, we add the intra- and inter-sentential instances from the 3000 articles on the topics of Hurricane Katrina and Iraq War. The portion of these articles were collected and used previously by Riaz and Girju (2010) to identify causal relations in news scenarios. We use these collections because the natural disaster and war related news articles are rich in causal relations and chains of such relations. In our development set, there is a total of $10,774$ distinct verb-verb pairs. Using intra- and inter-sentential instances of these verb-verb pairs, We compute the likelihood of these pairs to encode causation. In the rest of this section, we introduce our metrics for this purpose.

3.2.1 Explicit Causal Association (ECA)

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

$$ECA(\mathrm{v_i\text{-}v_j}) = \frac{1}{\mid VP \mid} \sum_{I_{\mathrm{v_i\text{-}v_j}} \in VP} (CD(\mathrm{v_i\text{-}v_j}) \times C_I) \tag{3.8}$$

where VP is the set of intra- and inter-sentential instances of verb-verb pairs. An instance of $\mathrm{v_i}$-$\mathrm{v_j}$ pair is denoted by $I_{\mathrm{v_i\text{-}v_j}}$. CD determines the causal dependency of a verb-verb pair in an unsupervised fashion (equation 3.2), and C_I finds the tendency of instance I of $\mathrm{v_i}$-$\mathrm{v_j}$ pair to belong to the cause class as compared to the non-cause class using the training corpus of event-event pairs. The goal of ECA is to combine the unsupervised causal dependency score (i.e., CD) with the supervised score of instance I of belonging to the cause class than the non-cause one (i.e., C_I). Here, CD represents the prior knowledge about the causal association based on the co-occurrence probabilities and idf scores (equation 3.2). It can discover lots of false positives because the co-occurrence probabilities can fail to differentiate causality from any other type of correlation. We improve the prior knowledge obtained from CD with the help of supervision from the training corpus of both C and $\neg C$ relations. The global decision of causal association of a verb-verb pair is made by taking the average of scores on all instances of that pair. Notice that CD can also be moved out from the summation function in equation 3.8.

We define the score C_I as follows:

$$C_I = \frac{P(I, C)}{P(I, \neg C)} \tag{3.9}$$

Here, the probability $P(I, C)$ is the probability of assignment of the label C to the instance I. We can

obtain these probabilities using both Naive Bayes and Maximum Entropy classifiers introduced in section 3.1. However, in our model we do not employ the Maximum Entropy classifier for the calculation of C_I because it works very slow on the massive development set. Therefore, we employ the following function for the fast computation:

$$C_I = \sum_{k=1}^{n} log(\frac{P(f_k \mid C)}{P(f_k \mid \neg C)}) \tag{3.10}$$

The notation f_k represents a feature on an instance I. In section 3.1.2, we have introduced a set of linguistic features we employ to predict the labels C and $\neg C$. $P(f_k \mid C)$ and $P(f_k \mid \neg C)$ are the smoothed probabilities of a feature f_k given the cause and non-cause training instances. The value of C_I is positive only when the instance I has more tendency to encode a cause relation than a non-cause one. To avoid negative values, we map the scores of C_I to the range $[0, 1]$ using $\frac{C_I - C_{min}}{C_{max} - C_{min}}$ where C_{min} (C_{max}) is the minimum (maximum) value of C_I obtained on the development set, respectively. Also, we add a small value ϵ to C_I to avoid 0 value. Similarly, to avoid negative scores of PMI in equation 3.2 we can map it to the range [0,1].

We employed both training corpora $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ and $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ (see section 3.1.1) to calculate the scores of metric ECA. Our empirical evaluation revealed that the ECA scores acquired using $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus provides better source of background knowledge than the scores acquired using $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$. This makes sense because we acquire causal associations by considering the scores of ECA on the massive development set and the training corpus $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ is very small for this purpose.

We selected top 500 scored verb-verb pairs using the metric ECA. Following are some examples of causal verb-verb pairs from these top 500 pairs: destroy-rebuild, convict-arrest, receive-download, ask-reply, score-win, etc. We also observed some false positives in the top 500 pairs i.e., those pairs which do not seem to encode a cause-effect relation. Some examples of these pairs are jump-rise, hit-strike, drop-fall, climb-gain, meet-discuss, etc. Notice that in these examples some pairs contain nearly synonymous verbs (e.g., jump-rise, hit-strike) or the verbs in temporal only relation (e.g., drop-fall, climb-gain, meet-discuss). In the next chapter we empirically evaluate performance of the metric ECA by using the causal associations in verb-verb pairs derived from this metric in our model for identifying causality.

Natural language allows the expression of semantic relations in both ambiguous and implicit contexts. This fact increases the complexity of the current task to a large extent. Sporleder and Lascarides (2008) raised an important observation that people tend to avoid unnecessary redundancy while expressing semantic relations. For example, they prefer not to use a discourse marker when a semantic relation can be inferred from other elements of the context. Taking this observation forward, we assume that when a verb-verb

pair is strongly causal in nature (e.g., kill-arrest) then people may hardly use an explicit and unambiguous discourse marker to express the causation encoded by this pair. The strong causal link of this pair is obvious to the readers even when an ambiguous or no discourse marker is available to signal causality encoded by this pair. Therefore, the causality of such verb-verb pairs can remain undiscovered by the metric ECA because this metric relies on the supervision from Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus in which two events of each training instance appear in explicit and unambiguous context. We use the term training data sparseness for this problem where the strongly causal verb-verb pairs hardly appear in the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ training corpus. Due to this problem, we can mistakenly consider the strongly causal verb-verb pairs as non-causal. In the next section, we introduce a metric which addresses this problem to derive the better scores of causal associations in verb-verb pairs.

3.2.2 Implicit Causal Association (ICA)

In this section, we propose a metric ICA to handle the problem of training data sparseness discussed in the previous section. This metric makes use of functions for the identification of roles of events in a cause relation. After briefly describing the roles of events in a causal relation below, we continue with the description of ICA.

- **Roles of Events in a Causal Relation:** Each of the two events in a causal relation can be assigned either cause or effect role. For example (3) from section 3.1.1, the verb appearing after "because" represents a cause event and the verb before "because" represents an effect event. These roles of events are given below:

  5. A Michigan woman lost custody of her young daughter *because* she **placed** the child in day care while attending college classes. (e_{place}, R_C)
  6. A Michigan woman **lost** custody of her young daughter *because* she placed the child in day care while attending college classes. (e_{lose}, R_E)

  The notations R_C and R_E represent the cause and effect roles, respectively. Table 3.3 shows the assignment of roles to the events connected by the unambiguous discourse markers. We used these discourse markers to generate the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ training corpus.

  We use core features of events to determine the likelihood of their roles in causation. These features include lemmas, part-of-speech tags, all senses from WordNet of both verbs and their arguments (i.e., subject and object). Next, we use these features to handle training data sparseness.

Discourse Marker	Roles Information
Because	$(e_{v_{before}}, r_E)$, $(e_{v_{after}}, r_C)$
For this (that) reason	$(e_{v_{before}}, r_C)$, $(e_{v_{after}}, r_E)$
Consequently	$(e_{v_{before}}, r_C)$, $(e_{v_{after}}, r_E)$
As a consequence of	$(e_{v_{before}}, r_E)$, $(e_{v_{after}}, r_C)$
As a result of	$(e_{v_{before}}, r_E)$, $(e_{v_{after}}, r_C)$

Table 3.3: A list of causal discourse markers and the assignment of roles to the events of causal relations signaled by these markers. The event $e_{v_{before}}$ ($e_{v_{after}}$) is represented by the verb appearing before (after) the causal discourse marker in text, respectively.

- **Handling of Training Data Sparsity:** To deal with the problem of training data sparsity, we define the metric ICA as follows:

$$ICA(\mathrm{v_i}\text{-}\mathrm{v_j}) = \frac{1}{\mid VP \mid} \sum_{I_{\mathrm{v_i}\text{-}\mathrm{v_j}} \in VP} (CD(\mathrm{v_i}\text{-}\mathrm{v_j}) \times C_I \times ERM_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}}) \tag{3.11}$$

where CD and C_I are defined earlier and ERM determines the likelihood of the roles of events in a cause relation. We remind the reader that CD is the unsupervised causal dependency of verb-verb pair and C_I is the tendency of instance I of a verb-verb pair to belong to the cause class than the non-cause one using the full set of features from section 3.1.2.

Events Roles Matching ($\mathrm{ERM}_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}}$) (equations 3.12 and 3.13) is the negative log-likelihood of events $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ appearing as cause or effect role determined using the causal training instances of $\mathrm{Explicit}_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}}$ corpus and the core features of events discussed above.

$$ERM_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}} = -1.0 \times \max(S(e_{\mathrm{v_i}}, R_C) + S(e_{\mathrm{v_j}}, R_E), S(e_{\mathrm{v_i}}, R_E) + S(e_{\mathrm{v_j}}, R_C)) \tag{3.12}$$

$$S(e_{\mathrm{v_i}}, R_C) = \sum_{k=1}^{n} log(P(f_k \mid R_C)) \tag{3.13}$$
$$S(e_{\mathrm{v_j}}, R_E) = \sum_{k=1}^{n} log(P(f_k \mid R_E))$$

Here, $\mathrm{S}(e_{\mathrm{v_i}}, R_C)$ is the score of $e_{\mathrm{v_i}}$ being a cause event and $\mathrm{S}(e_{\mathrm{v_j}}, R_E)$ is the score of $e_{\mathrm{v_j}}$ being an effect event. These scores are computed using smoothed probabilities – i.e., $\mathrm{P}(f_k \mid R_C)$ and $\mathrm{P}(f_k \mid R_E)$.

Similarly, S($e_{\mathrm{v_i}}, R_E$) and S($e_{\mathrm{v_j}}, R_C$) are calculated and max is taken. A high score of ERM represents low matching of an event-event pair (verbs and their arguments) with the explicit contexts of causal training instances of Explicit$_{\mathrm{e_{v_i}\text{-}e_{v_j}}}$ corpus. The high score of ERM of an event-event pair can have one of the following two interpretations: (A) it is a non-causal pair, or (B) it is a causal pair but this pair and the pairs which are semantically closer to it hardly appear in explicit and unambiguous causal contexts. In the metric ICA, CD($\mathrm{v_i}$-$\mathrm{v_j}$) $\times$ C_I is used as a guiding score to interpret the scores of ERM as follows:

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

ICA is a boosting factor to determine the causal verb-verb pairs that remain undiscovered due to the problem of training data sparseness. We selected top 500 scored verb-verb pairs using the metric ICA. Following are some examples of causal verb-verb pairs from these 500 pairs: shoot-hold, fall-break, develop-provide, hit-hold, break-make, etc. These examples of pairs are not included in the top 500 list of pairs by the metric ECA due to the problem of training data sparseness. We also observed some false positives in the top 500 scored pairs. Some examples of these pairs are cut-raise, carry-leave, fall-boost, give-take, raise-lower, etc. Notice that in these examples some pairs contain nearly antonymous verbs (e.g., cut-raise, carry-leave, fall-boost, raise-lower) or the verbs in temporal only relation (e.g., give-take). In the next chapter we empirically evaluate performance of the metric ICA by using the causal associations in verb-verb pairs derived from this metric in our model for identifying causality.

We also define a Boosted Causal Association (BCA) metric by adding ICA to the original ECA metric as follows:

$$BCA(\mathrm{v_i}\text{-}\mathrm{v_j}) \quad = \quad \frac{1}{\mid VP \mid} \sum_{I_{\mathrm{v_i}\text{-}\mathrm{v_j}} \in VP} (CD(\mathrm{v_i}\text{-}\mathrm{v_j}) \times C_I) + (CD(\mathrm{v_i}\text{-}\mathrm{v_j}) \times C_I \times ERM_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}}) \quad (3.14)$$

We acquire the likelihood of each verb-verb pair to encode causation via above metrics and store this information in a resource called the knowledge base of causal associations of verb-verb pairs (i.e., KB_c).

3.2.3 Forms of Background Knowledge

In this section, we propose two novel forms in which we provide background knowledge to our model. We derive these two forms using the scores of likelihood of verb-verb pairs to encode causation. These scores of likelihood of causality in verb-verb pairs are available from the above stated resource KB_c. The two forms of background knowledge are given below.

Ranking Scores of Verb-Verb Pairs

For this form of background knowledge, we assign a ranking score to each v_i-v_j pair based on its likelihood to encode causation as compared with other verb-verb pairs of the language. We use the notation $RS(\text{v}_\text{i}\text{-v}_\text{j})$ for the ranking score of v_i-v_j pair. We define this score as follows:

$$RS(\text{v}_\text{i}\text{-v}_\text{j}) = \frac{|KB_c| - rank_{\text{v}_\text{i}\text{-v}_\text{j}}}{|KB_c|} \tag{3.15}$$

where $|KB_c|$ is the total size of the knowledge base KB_c – i.e., how many distinct verb-verb pairs are contained in this knowledge base. In KB_c, there is a total of $10,774$ verb-verb pairs. In order to calculate the ranking score RS, we employ the ranked list of verb-verb pairs of the resource KB_c. These pairs are ranked in descending order with respect to their likelihood to encode causation. We determine the value of the function $rank_{\text{v}_\text{i}\text{-v}_\text{j}}$ for a v_i-v_j pair based on its position in the ranked list. For example, if the ranked list of pairs is generated based on the scores of ECA then the pair with the highest score has $rank = 0$ and the pair with the next highest score has $rank = 1$ and so on. We assign the same rank to two pairs with the same likelihood to encode causation. The RS function value lies in range (0,1] and it is maximum for the verb-verb pair with the highest likelihood of encoding causality. The score (1.0 - RS) is the likelihood of a verb-verb pair to encode non-causality. We employ the value of $RS(\text{v}_\text{i}\text{-v}_\text{j})$ to provide background knowledge to our model. We use the notation KB_1 for this form of background knowledge.

Categories of Verb-Verb Pairs

For the second form of background knowledge, we divide verb-verb pairs into three categories: (1) Strongly Causal (S_c), (2) Ambiguous (A_c) and (3) Strongly Non-causal ($S_{\neg c}$) to provide background knowledge. To generate these categories we employ the ranked list of all verb-verb pairs of KB_c with respect to their likelihood to encode causation. Then, we assume that all pairs are uniformally distributed across three categories - i.e., top one-third and bottom one-third ranked pairs belong to the Strongly Causal (S_c) and Strongly Non-causal ($S_{\neg c}$) categories and the rest of the pairs are considered Ambiguous (A_c). In this work,

we employ this uniform categorization to provide background knowledge but in the future researchers can perform empirical study of how to automatically cluster verb-verb pairs into three or more categories with respect to causation. The category S_c ($S_{\neg c}$) contains the verb-verb pairs with the highest (least) tendency to encode a causal relation, respectively and A_c contains the verb-verb pairs with tendency to encode both types of relation depending on the context. We use the notation KB_2 for this form of background knowledge.

In our model, we add background knowledge using either the form KB_1 or KB_2. In the next chapter, we explain our approach to incorporate this knowledge to our model for identifying causality.

3.3 Identification of the Causal Semantics of Verbs

Verbs are the expressions of language for representing events of various semantic types. We argue that each verb can have its own semantics with respect to the relation of causality. In order to support our argument, we identify the causal semantics of verbs by focusing on the linguistic definition of events (see section 3.3.1) and the semantic classes of events represented by the verbs (see section 3.3.2).

3.3.1 Linguistic Definition of Events

As discussed in section 1.3.2, in linguistics and computational linguistics researchers mainly categorize each word (or phrase) into two aspectual classes: STATE and EVENT where a STATE describes an unchanging situation over a period of time (e.g., know, love) and an EVENT describes a situation involving the internal structure (e.g., run has an internal structure of raising a foot in air, moving it forward and putting it down on floor). In our model, we started with a naive assumption that each main verb (v_i) represents an event along with its arguments – i.e., a 3-tuple e_{v_i} =([subject$_{v_i}$] , v_i, [object$_{v_i}$]) is an event. Now, we propose to incorporate the above linguistic definition of events in our model. Using this definition, we want to determine if the pair e_{v_i}-e_{v_j} has a tendency to encode a causal relation or not. Particularly, for a pair v_i-v_j, we first automatically identify if any of the two 3-tuples (i.e., e_{v_i} or e_{v_j}) represents an event according to the linguistic definition or not. After acquiring this information, we stress on the fact that the pair e_{v_i}-e_{v_j} can encode causation only if at least one of e_{v_i} and e_{v_j} is an event according to the linguistic definition. In other words, we assume that there is a high tendency for the e_{v_i}-e_{v_j} pair to encode a non-cause relation if both e_{v_i} and e_{v_j} are not events according to the linguistic definition. The reason for this assumption is the previous work of Girju and Moldovan (2002) in which it was pointed out that the events have a high tendency to encode causation as compared with any other state of affairs.

In order to acquire information regarding the linguistic definition of events, we predict the labels of

Event (E) or Non-Event (¬E) on both e_{v_i} and e_{v_j} 3-tuples of e_{v_i}-e_{v_j} pair. For this purpose, we employ the TimeBank's corpus [Pustejovsky et al. 2003] with the annotations of events and non-events labels on verbs (or verbal phrases). In this corpus, there is a total of 5132 events and 2792 non-events annotations available for the verbs (or verbal phrases). Pustejovsky et al., (2003) have provided these annotations by closely following the linguistic definition of events. Consider the following example of a verbal event "applauded" from the TimeBank corpus:

The company's sales force **applauded** the shake up.[2]

Inspired by the work of Bethard and Martin (2006), we build a supervised classifier to predict the labels E or $\neg E$ on both of the 3-tuples of e_{v_i}-e_{v_j} pairs. We employ the TimeBank's annotations of verbal events and non-events and the features given in Table 3.4 to build the supervised classifier. After acquiring the labels from this classifier, a pair e_{v_i}-e_{v_j} can fall into one of the following four cases: (1) both e_{v_i} and e_{v_j} are events (i.e., the label E-E for the pair), (2) e_{v_i} is an event and e_{v_j} is not an event (i.e., the label E-¬E), (3) e_{v_i} is not an event and e_{v_j} is an event (i.e., the label ¬E-E), and (4) both e_{v_i} and e_{v_j} are not events (i.e., the label ¬E-¬E). We assume that if a e_{v_i}-e_{v_j} pair is assigned the label ¬E-¬E then it cannot encode causality because at least one of e_{v_i} and e_{v_j} needs to be an event according to the linguistic definition in order to encode causation.

We acquire the probabilities of the above four cases as follows:

- $\mathrm{P}(e_{v_i}\text{-}e_{v_j}, \text{E-E}) = P(e_{v_i}, \mathrm{E})P(e_{v_i}, \mathrm{E})$
- $\mathrm{P}(e_{v_i}\text{-}e_{v_j}, \text{E-}\neg\text{E}) = P(e_{v_i}, \mathrm{E})P(e_{v_i}, \neg\mathrm{E})$
- $\mathrm{P}(e_{v_i}\text{-}e_{v_j}, \neg\text{E-E}) = P(e_{v_i}, \neg\mathrm{E})P(e_{v_i}, E)$
- $\mathrm{P}(e_{v_i}\text{-}e_{v_j}, \neg\text{E-}\neg\text{E}) = P(e_{v_i}, \neg\mathrm{E})P(e_{v_i}, \neg\mathrm{E})$

For each of the above four cases, the $P(e_{v_i}, \mathrm{E})$ is provided by the supervised classifier for identifying events. These probabilities are provided to our model to help determine if an instance of e_{v_i}-e_{v_j} pair can encode causality or not.

3.3.2 Semantic Classes of Events

In this section, we focus on the semantic classes of events to identify the causal semantics of verbs. Pustejovsky et al. (2003) have organized the verbal events of TimeBank's corpus into the following seven semantic classes: (1) OCCURRENCE, (2) PERCEPTION, (3) ASPECTUAL, (4) STATE, (5) I_STATE, (6)

[2]This example is taken from Bethard and Martin (2006).

Feature Type	Description
Lexical Features	verb or verbal phrase, lemma of the verb, subject and words and lemmas of the subject and object of verb, affixes (first and last three characters of the verb).
Word Class Features	part-of-speech tag of the verb and the subject and object of the verb.
Semantic Class Features	Frequent sense of the verb or the head verb of the verbal phrase from WordNet. In order to obtain the head verb, we traverse from the last word of the verbal phrase and pick the very first verb as the head verb.
Hypernym Features	This feature is set to 1 if any sense of the verb or the head verb of the verbal phrase falls into one of the following three senses from WordNet: (1) {think, cogitate, cerebrate}, (2) {move, displace} and (3) {act, move}. These three senses were identified by Bethard and Martin (2006) as the most discriminative senses for identifying events. They acquired these senses using the following scheme: Using each of the WordNet's hierarchy, they classified all words falling in that hierarchy as events and all words falling outside that hierarchy as non-events. They also applied this rule in reverse – i.e., all words falling in that hierarchy as non-events and all words falling outside that hierarchy as events. They used this procedure by employing the training instances from the TimeBank's corpus. They found the above stated hierarchies with the highest F-scores on the training instances.

Table 3.4: The linguistic features introduced by Bethard and Martin (2006) to identify events and non-events and the semantic classes of events.

I_ACTION and (7) REPORTING (see section 2.2 for the definitions of these semantic classes). In this work, we argue that each verb can have its own causal semantics depending on the semantic class of event it is representing in an instance. For example, a verb representing a REPORTING event may just narrate or describe another event instead of encoding causality with it. Consider the following example from the TimeBank corpus in which an event represented by the verb "said" is just narrating another event represented by the verb "sent".

5. In another mediation effort, the Soviet Union **said** today it had **sent** an envoy to the Middle East on a series of stops to include Baghdad .

In this work, we execute a data intensive procedure (see Procedure 3.1) to automatically identify tendencies of each of the above semantic classes to encode a cause and non-cause relation. This method returns a set of semantic classes of events with the highest tendency to encode non-causation. We assume that the semantic classes belonging to this set have a low tendency to encode causation and the rest of the semantic classes have a high tendency to encode causation. We provide this information of the categorization of semantic classes to our model to make better predictions for the current task.

Procedure 3.1 given below takes a training corpus of e_{v_i}-e_{v_j} pairs with C and $\neg C$ labels and a set of semantic classes – i.e., SC={OCCURRENCE, PERCEPTION, ASPECTUAL, STATE, I_STATE, I_ACTION, REPORTING} as input. This procedure outputs a set $SC_{\neg c}$ which contains the semantic classes with the highest tendency to encode non-cause relations. We apply Procedure 3.1 on both $Explicit_{e_{v_i}\text{-}e_{v_j}}$ and $PDTB_{e_{v_i}\text{-}e_{v_j}}$ corpora. This yields the sets $SC_{\neg c}$ = {ASPECTUAL} and $SC_{\neg c}$ = {REPORTING, STATE} using $Explicit_{e_{v_i}\text{-}e_{v_j}}$ and $PDTB_{e_{v_i}\text{-}e_{v_j}}$ corpora, respectively. In order to execute Procedure 3.1, we need to predict the semantic classes of both events e_{v_i} and e_{v_j} of a e_{v_i}-e_{v_j} pair. We acquire these predictions by building a supervised classifiers on the TimeBank corpus with the features given in Table 3.4.

Input: Training corpus of e_{v_i}-e_{v_j} pairs, SC: {OCCURRENCE, PERCEPTION, ASPECTUAL, STATE, I_STATE, I_ACTION, REPORTING}

Output: $SC_{\neg c}$: Semantic classes of events with the highest tendency to encode a non-cause relation

1. Initialize T = ø

2. **for** *each training instance k with the label* $l \in \{C, \neg C\}$ **do**
   3. Identify the semantic class (sc) of both events of e_{v_i}-e_{v_j} pair
   4. Add the tuples (k, $sc_{e_{v_i}}$, l) and (k, $sc_{e_{v_j}}$, l) to the set T.

end

5. Calculate the tendency of each semantic class sc ∈ SC to encode a non-cause relation using the following score:

$$
\begin{aligned}
score(sc, \neg C) &= score_1(sc, \neg C) \times score_2(sc, \neg C) \\
score_1(sc, \neg C) &= \left(\frac{count(T, (*, sc, \neg C))}{count(T, (*, sc, *))} - \frac{count(T, (*, sc, C))}{count(T, (*, sc, *))}\right) \\
score_2(sc, \neg C) &= \left(\frac{count(T, (*, sc, \neg C))}{count(T, (*, *, \neg C))} - \frac{count(T, (*, sc, C))}{count(T, (*, *, C))}\right)
\end{aligned}
$$

where count(T,(m,n,o)) is the count of (m,n,o) tuples in the set T. We put * to show that we do not care for that value. For example, count(T, (*, *, $\neg C$)) is the count of tuples in T with the label $\neg C$.

6. Acquire a ranked list of the semantic classes w.r.t their tendencies to encode non-causation. This results in a $list_{sc}$ = [sc_1, sc_2, … sc_m] s.t. $score(sc_i, \neg C) > score(sc_{i+1}, \neg C)$. From this list we remove the class sc_i if either the $score_1(sc_i, \neg C) < 0$ or $score_2(sc_i, \neg C) < 0$.

7. Initialize $SC_{\neg C}$ = ø and $result_{sc_{-1}}$ =$result_{sc_0} = 0$

8. **while** *not the end of* $list_{sc}$ **do**
   9. Remove sc_i from the front of the $list_{sc}$
   10. Initialize the set $S_1 = SC_{\neg C} + \{sc_i\}$ and the set $S_2 = \{sc_{i+1}, sc_{i+2}, \ldots, sc_m\}$.
   11. **for** *each (k, sc, l)* $\in T$ **do**
      12. Predict the label $\neg C$ if sc ∈ S_1 and predict the label C if sc ∈ S_2.
   end
   13. Using the predictions from the step 12, calculate $result_{sc_i}$ = F1-score × accuracy.
   14. **if** $result_{sc_i} - result_{sc_{i-1}} < result_{sc_{i-1}} - result_{sc_{i-2}}$ **then**
      15. Ouput $SC_{\neg c}$
   else
      16. Go to step 8
   end

end

Procedure 3.1. A data intensive procedure to acquire a set of semantic classes of events with the highest tendency to encode a non-cause relation.

In step 3 of above procedure, we identify the semantic classes of both events e_{v_i} and e_{v_j} of a e_{v_i}-e_{v_j} pair. After the predictions of semantic classes for all events of the training instances of e_{v_i}-e_{v_j} pairs, we identify tendency of each semantic class sc $\in$ SC to encode a non-cause relation. This is done by computing the $score(sc, \neg C)$ in the step 5 of the procedure. This score has two components $score_1(sc, \neg C)$ and $score_2(sc, \neg C)$. $score_1(sc, \neg C)$ is greater than 0 only if the semantic class sc encodes non-cause relations more often than the cause ones. $score_2(sc, \neg C)$ is greater than 0 only if the percentage of total non-cause training instances with the semantic class sc is greater than the percentage of total cause training instances with the semantic class sc. We generate a list of semantic classes of events in descending order w.r.t the value of $score(sc, \neg C)$ for each sc $\in$ SC. This results in a ranked $\text{list}_{sc} = [\text{sc}_1, \text{sc}_2, \ldots \text{sc}_m]$ where $score(sc_i, \neg C) > score(sc_{i+1}, \neg C)$. We did not encounter a situation where $score(sc_i, \neg C) = score(sc_{i+1}, \neg C)$. From the list_{sc}, we remove the class sc_i if either the $\text{score}_1(\text{sc}_i, \neg C) < 0$ or $\text{score}_2(\text{sc}_i, \neg C) < 0$ because the class sc has a higher tendency to encode a cause relation than the non-cause one. For the list_{sc}, we determine the semantic class sc_i above which all the semantic classes have a tendency to encode non-causation –i.e., the $\text{SC}_{\neg c} = \{ \text{sc}_1, \text{sc}_2, \ldots, \text{sc}_{i-1}\}$. We identify sc_i using the steps 7 to 16 of Procedure 3.1. The main idea is to traverse the list_{sc} in order and predict the label $l \in \{C, \neg C\}$ for the tuples of T based on the semantic class sc. For example, if we reach sc_2 in the list_{sc}, then predicts $\neg C$ for all tuples of form (*,sc_1,*) and (*,sc_2,*) $\in$T and C for the rest of the tuples. On these predictions we calculate performance in terms of F-score $\times$ Accuracy. We keep on traversing the list_{sc} and stop where the performance gain is less than the performance gain achieved in the last step.

Procedure 3.1 yields the set $\text{SC}_{\neg c} = \{\text{ASPECTUAL}\}$ and $\text{SC}_{\neg c} = \{\text{REPORTING, STATE}\}$ using the training $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ and $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpora, respectively. The above results imply that ASPECTUAL, REPORTING and STATE events have the highest tendency to encode non-causation as compared with the semantic classes OCCURRENCE, PERCEPTION, I_STATE, I_ACTION. In the set $\text{SC}_{\neg c} = \{$REPORTING, STATE$\}$, REPORTING class has a high tendency to encode non-causation according to the $score(sc, \neg C)$ function of Procedure 3.1 as compared with STATE class. Following are some examples of ASPECTUAL, REPORTING and STATE events encoding the non-causal relations:

6. He **started** his athletic career as a swimmer with Olympic potential but **switched** to basketball after Hurricane Hugo damaged the pool in which he trained.

7. Although preliminary findings were **reported** more than a year ago, the latest results **appear** in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

8. Although traders rushed to **buy** futures contracts, many **remained** skeptical about the Brazilian

development , which couldn't be confirmed, analysts said.

Example (6) is taken from the Explicit$_{\mathrm{e_{v_i}\text{-}e_{v_j}}}$ corpus where the event e_{start} is the ASPECTUAL event encoding a non-cause relation with the event e_{switch} an OCCURRENCE event. Example (7) is taken from the PDTB$_{\mathrm{e_{v_i}\text{-}e_{v_j}}}$ corpus where the event e_{report} (a REPORTING event) is just explaining the facts i.e., preliminary findings and is encoding a non-cause relation with the event e_{appear} an OCCURRENCE event. In example (8) from the PDTB$_{\mathrm{e_{v_i}\text{-}e_{v_j}}}$ corpus, e_{remain} is a STATE event encoding a non-cause relation with the event e_{buy} an OCCURRENCE event.

Notice that for Procedure 3.1 we do not consider temporal directions of events. For example, what is the tendency of a REPORTING event $\mathrm{e_{v_i}}$ to encode a cause-effect relation with another event $\mathrm{e_{v_j}}$ if it temporally precedes (overlaps) the event $\mathrm{e_{v_j}}$? In current work we do not consider temporal directions because the temporal relations between events are not always available for the test sets and the current state-of-the-art model for identifying temporal relations [Do 2012] is far from achieving performance close to humans. However, in the future it is an interesting topic to explore the tendencies of the semantic classes of events to encode causation with respect to temporal directions.

After the acquisition of set $\mathrm{SC}_{\neg c}$, we classify the verbal events into two classes $\mathrm{C_{e_v}}$ and $\neg\mathrm{C_{e_v}}$ where the class $\mathrm{C_{e_v}}$ ($\neg\mathrm{C_{e_v}}$) contains the verbal events with a high (low) tendency to encode causation, respectively. For example, if the above procedure outputs the set $\mathrm{SC}_{\neg c}$ = {REPORTING, STATE} then the class $\mathrm{C_{e_v}}$ consists of all REPORTING and STATE events and the class $\neg\mathrm{C_{e_v}}$ consists of all other types of events. Using this rule, we classify the instances of verbal events of the TimeBank corpus into the classes $\mathrm{C_{e_v}}$ and $\neg\mathrm{C_{e_v}}$. Using the instances of TimeBank corpus with the labels $\mathrm{C_{e_v}}$ and $\neg\mathrm{C_{e_v}}$, we build a supervised classifier for these labels. We employ the linguistic features given in Table 3.4 to build this classifier. This classifier is then used to predict the labels on both events $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ of a $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair. According to the predictions of supervised classifier, each $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair can fall into one of the following four cases: (1) both $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ belong to the class $\mathrm{C_{e_v}}$ (i.e., the label $\mathrm{C_{e_v}\text{-}C_{e_v}}$ is assigned to the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair), (2) $\mathrm{e_{v_i}}$ belongs to the class $\mathrm{C_{e_v}}$ and $\mathrm{e_{v_j}}$ belongs to the class $\neg\mathrm{C_{e_v}}$ (i.e., the label $\mathrm{C_{e_v}\text{-}\neg C_{e_v}}$), (3) $\mathrm{e_{v_i}}$ belongs to the class $\neg\mathrm{C_{e_v}}$ and $\mathrm{e_{v_j}}$ belongs to the class $\mathrm{C_{e_v}}$ (i.e., the label $\neg\mathrm{C_{e_v}\text{-}C_{e_v}}$) and (4) both $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ belong to the class $\neg\mathrm{C_{e_v}}$ (i.e., the label $\neg\mathrm{C_{e_v}\text{-}\neg C_{e_v}}$).

We acquire the probabilities of the above four cases as follows:

- $\mathrm{P(e_{v_i}\text{-}e_{v_j}, C_{e_v}\text{-}C_{e_v})} = P(\mathrm{e_{v_i}, C_{e_v}})P(\mathrm{e_{v_i}, C_{e_v}})$
- $\mathrm{P(e_{v_i}\text{-}e_{v_j}, C_{e_v}\text{-}\neg C_{e_v})} = P(\mathrm{e_{v_i}, C_{e_v}})P(\mathrm{e_{v_i}, \neg C_{e_v}})$
- $\mathrm{P(e_{v_i}\text{-}e_{v_j}, \neg C_{e_v}\text{-}C_{e_v})} = P(\mathrm{e_{v_i}, \neg C_{e_v}})P(\mathrm{e_{v_i}, C_{e_v}})$

- $P(e_{v_i}\text{-}e_{v_j}, \neg C_{e_v}\text{-}\neg C_{e_v}) = P(e_{v_i}, \neg C_{e_v})P(e_{v_i}, \neg C_{e_v})$.

For the above four cases, $P(e_{v_i}, C_{e_v})$ and $P(e_{v_i}, \neg C_{e_v})$ are provided by the supervised classifier for identifying the labels C_{e_v} and $\neg C_{e_v}$. These probabilities provide information to our model about the tendency of the semantic class of a verb to encode causation. We assume that a e_{v_i}-e_{v_j} has a low tendency to encode causation if one of the following labels C_{e_v}-$\neg C_{e_v}$, $\neg C_{e_v}$-C_{e_v} and $\neg C_{e_v}$-$\neg C_{e_v}$ is assigned to it. In chapter 4, we propose our approach to incorporate the knowledge of causal semantics of verbs.

3.4 Summary

The task of knowledge acquisition for verb-verb pairs is critical for building our model for identifying causality. In this chapter, we have introduced methods to acquire the background knowledge and the knowledge of causal semantics of verbs. Our objective is to incorporate the knowledge of context with the above types of knowledge to identify causality. In this chapter, we started with proposing our method for automatically acquiring the training corpora of verb-verb pairs labeled with cause and non-cause relations. The automated derivation of a training corpus saves us from the trouble of manually annotating a large number of instances for the current task. We build a supervised classifier using the two training corpora $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ and $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ acquired for the current task. The supervised classifier provides predictions of the labels C and $\neg C$ using linguistic features extracted from the contexts of instances of e_{v_i}-e_{v_j} pairs. It also provides the probabilities of assignments of these labels which is termed as the knowledge of context in this thesis.

The knowledge of context allows our model to predict causality by merely depending on linguistic features. However, the models relying only on such features lack the additional sources of knowledge necessary to identify causality e.g., background knowledge. Therefore, we have introduced our methods to derive background knowledge. We aim to supply background knowledge to our model in terms of causal associations of verb-verb pairs. For this purpose, we have proposed a set of very carefully designed metrics. These metrics acquire causal associations by taking care of all types of contexts i.e., unambiguous, ambiguous and implicit. There are two forms in which we aim to supply background knowledge to our model. These two forms include the ranking scores of verb-verb pairs and the novel categorization of these pairs with respect to relation of causality. In order to identify causality in verb-verb pairs, it is important to understand the causal semantics of verbs. We extract this type of knowledge by focusing on the linguistic definition of events and the semantic classes of events represented by verbs.

Chapter 4

Identifying Causality in Verb-Verb Pairs

After the knowledge acquisition for the task of identifying causality in verb-verb pairs, our objective is to combine all types of knowledge to obtain optimal predictions for the current task. For this purpose, we take advantage of the learning and inference framework of Integer Linear Programming (ILP) for NLP [Roth and Yih 2004]. In this chapter, we present our model for identifying causality in verb-verb pairs and discuss performance of our model for the current task.

4.1 Model for Identifying Causality

In the framework of Integer Linear Programming for NLP [Roth and Yih 2004], various sources of knowledge are added in the form of hard and soft constraints to an integer linear program. The objective function of the integer linear program is then optimized in the presence of constraints to achieve the globally coherent predictions for the NLP tasks. In our approach, we begin with setting up an integer linear program with the knowledge of context and then incrementally add other types of knowledge to achieve progress over the model relying merely on linguistic features.

4.1.1 Knowledge on Context

As discussed in chapter 3, the supervised classifier for identifying causality using linguistic features provides the knowledge of context. This type of knowledge comprises the probabilities of assignments of the labels C and $\neg C$ to $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pairs. Using the knowledge of context we set up the following integer linear program:

$$Z_1 = \max \sum_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP} \sum_{l \in L_1} x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, l) P(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, l) \tag{4.1}$$

$$\sum_{l \in L_1} x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP \tag{4.2}$$

$$x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP, \quad \forall l \in L_1 \tag{4.3}$$

Here, $L_1 = \{C, \neg C\}$, EP is the set of all $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pairs. $x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, l)$ is a binary decision variable (4.3) which is set to 1 only if the label $l \in L_1$ is assigned to the pair $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$. Constraint 4.2 enforces the assignment of only one label out of $|L_1|$ choices to a pair $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$. In our model, we maximize the objective function Z_1 (4.1) subject to the constraints introduced above. This function assigns the label $l \in L_1$ to all $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}} \in EP$ depending on the probabilities of assignments of labels to these pairs. We acquire these probabilities through the supervised classifier for the labels C and $\neg C$.

4.1.2 Background Knowledge

In this section, we propose our methods to add background knowledge to our model. In chapter 3, we acquired background knowledge in terms of the causal associations of verb-verb pairs. In addition to this, we introduce two forms of background knowledge named as KB_1 and KB_2. We add background knowledge to our model in one of the two forms –i.e., either KB_1 or KB_2.

Background Knowledge of Form KB_1

For this form of background knowledge we set up the Ranking Scores (RS) for verb-verb pairs. The score $RS(\mathrm{v_i}\text{-}\mathrm{v_j})$ is the likelihood of the pair $\mathrm{v_i}$-$\mathrm{v_j}$ to encode a cause relation and the score 1.0-$RS(\mathrm{v_i}\text{-}\mathrm{v_j})$ is the likelihood of $\mathrm{v_i}$-$\mathrm{v_j}$ to encode a non-cause relation. We make the following changes to the integer linear program by using the score RS:

$$\begin{aligned} Z_{KB_1} = \max \sum_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP} & x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, C)(RS(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}})) \times P(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, C)) \\ & + x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, \neg C)((1.0 - RS(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}})) \times P(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, \neg C)) \end{aligned} \tag{4.4}$$

Here, the function $f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}})$ returns the verb-verb pair of the $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pair – i.e., $f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}) = \mathrm{v_i}$-$\mathrm{v_j}$. We replace the objective function Z_1 (4.1) with the function 4.4 to incorporate background knowledge. We optimize function 4.4 subject to the constraints introduced in the previous section. Our integer linear program now assigns the label l to a $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pair depending on the product of probabilities from the supervised classifier and the ranking score of the $\mathrm{v_i}$-$\mathrm{v_j}$ pair. We add a small value $\alpha = 0.01$ to the functions RS and 1-RS to avoid 0 values.

Note that we acquire the ranking scores of verb-verb pairs from the resource KB_c consisting of only $10,774$ verb-verb pairs (see section 3.2.3). Therefore it is possible that the $\mathrm{v_i}$-$\mathrm{v_j}$ pair of a $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ instance does not exist in KB_c. In this situation, one idea is to increase size of the resource KB_c to cover all English language verb-verb pairs but it may not be practical to try this approach. For example, there are $3,769$ lemmas of verbs in the VerbNet [Kipper et al. 2000] and to cover all possible verb-verb pairs we need to process $\binom{3769}{2} = 7100796$ pairs. It is not feasible to process this big number of verb-verb pairs because we need lots of resources and time for this purpose. Therefore, we adopt the following method in which we do not increase the size of KB_c. In this method, for a $\mathrm{v_i}$-$\mathrm{v_j} \notin KB_c$ we search for a $\mathrm{v_p}$-$\mathrm{v_q} \in KB_c$ which is semantically closest to the $\mathrm{v_i}$-$\mathrm{v_j}$ pair using the following steps:

1. If there exists verbs $\mathrm{v_k}$ and $\mathrm{v_l}$ s.t. $\mathrm{v_i}$-$\mathrm{v_k} \in KB_c$ and $\mathrm{v_l}$-$\mathrm{v_j} \in KB_c$ then identify the semantically closest pair to $\mathrm{v_i}$-$\mathrm{v_j}$ as follows:

$$\mathrm{v_p\text{-}v_q} = \underset{\{\mathrm{v_i\text{-}v_n}, \mathrm{v_m\text{-}v_j}\}}{\operatorname{argmax}} (sim(\mathrm{v_i\text{-}v_m}), sim(\mathrm{v_n\text{-}v_j})) \tag{4.5}$$

   where $\mathrm{v_m}$ and $\mathrm{v_n}$ are collected using the following:

$$\begin{aligned} \mathrm{v_n} &= \underset{\mathrm{v_k}:\mathrm{v_i\text{-}v_k} \in KB_c}{\operatorname{argmax}} sim(\mathrm{v_k\text{-}v_j}) \\ \mathrm{v_m} &= \underset{\mathrm{v_l}:\mathrm{v_l\text{-}v_j} \in KB_c}{\operatorname{argmax}} sim(\mathrm{v_i\text{-}v_l}) \end{aligned} \tag{4.6}$$

   Here, $sim(\mathrm{w_i\text{-}w_j})$ is the measure of semantic similarity between two words $\mathrm{w_i}$ and $\mathrm{w_j}$ using WordNet [Lin 1998] (for details of method of computation of semantic similarity, we refer the reader to [Lin 1998]). We compute the sim function only for the verbs which exist in WordNet. If we are not able to compute sim functions for more than 50% of the verb-verb pairs involved in equation 4.6 for collection of both v_m and v_n then go to step 3. Secondly, to predict the semantically closest pair with some confidence we assume that in KB_c there are at least 5 instances of the pairs with $\mathrm{v_i}$ or 5 instances of the pairs with $\mathrm{v_j}$. If this is not the case then go to step 3.

2. Else If there exists more than one verb $\mathrm{v_k}$ s.t. $\mathrm{v_i}$-$\mathrm{v_k} \in KB_c$ and there is no verb $\mathrm{v_l}$ s.t. $\mathrm{v_l}$-$\mathrm{v_j} \in KB_c$ then identify the semantically closest pair to $\mathrm{v_i}$-$\mathrm{v_j}$ as follows:

$$\mathrm{v_p\text{-}v_q} = \underset{\mathrm{v_i\text{-}v_k} \in KB_c}{\operatorname{argmax}} sim(\mathrm{v_k\text{-}v_j}) \tag{4.7}$$

   Again, if we are not able to compute sim functions for more than 50% of the verb-verb pairs involved

in the equation 4.7 and in KB_c there are less than 5 $\mathrm{v_i}$-$\mathrm{v_k}$ pairs then go to step 3.

3. We assume there exists no semantically closest pair $\mathrm{v_p}$-$\mathrm{v_q} \in KB_c$ for the pair $\mathrm{v_i}$-$\mathrm{v_j} \notin KB_c$. In this case our model depends on predictions from the supervised classifier relying merely on linguistic features.

For a $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pair with $\mathrm{v_i}$-$\mathrm{v_j} \notin KB_c$, we identify a pair $\mathrm{v_p}$-$\mathrm{v_q} \in KB_c$ that is semantically closest to the $\mathrm{v_i}$-$\mathrm{v_j}$ pair using the above method. We then use the $\mathrm{v_p}$-$\mathrm{v_q}$ pair to acquire background knowledge from the resource KB_c. In this work, the above method is referred to as the "Procedure of Semantic Mapping of Verb-Verb Pairs".

Background Knowledge of Form KB_2

For this form of background knowledge, we organize verb-verb pairs into three categories: Strongly Causal (S_c), Ambiguous (A_c) and Strongly Non-causal ($\neg S_c$). The category S_c ($S_{\neg c}$) contains the verb-verb pairs with the highest (least) tendency to encode a causal relation, respectively and A_c contains the verb-verb pairs with tendency to encode both types of relation depending on the context. Using these categories of verb-verb pairs, we add background knowledge as follows:

$$Z_{KB_2} = Z_1 + \sum_{\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP} \sum_{l \in L_2} x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), l) P(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), l) \tag{4.8}$$

$$x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, C) - x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), S_c) \geq 0 \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP \tag{4.9}$$

$$x_1(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}, \neg C) - x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), S_{\neg c}) \geq 0 \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP \tag{4.10}$$

$$\sum_{l \in L_2} x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), l) = 1 \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP \tag{4.11}$$

$$x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), l) \in \{0, 1\} \quad \forall\, \mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}} \in EP, \quad \forall l \in L_2 \tag{4.12}$$

where $L_2 = \{S_c, A_c, S_{\neg}c\}$ is the set of labels for three categories of verb-verb pairs. The function $f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}})$ returns the verb-verb pair of the $\mathrm{e_{v_i}}$-$\mathrm{e_{v_j}}$ pair – i.e., $f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}) = \mathrm{v_i}$-$\mathrm{v_j}$. $x_2(f_{vp}(\mathrm{e_{v_i}}\text{-}\mathrm{e_{v_j}}), l)$ is a binary decision variable (4.12) which is set to 1 only if the label $l \in L_2$ is assigned to the pair $\mathrm{v_i}$-$\mathrm{v_j}$. The constraint 4.11 enforces the assignment of only one label out of $|L_2|$ choices to a $\mathrm{v_i}$-$\mathrm{v_j}$ pair. For each category $l \in L_2$, if the $\mathrm{v_i}$-$\mathrm{v_j}$ pair belongs to l then we set its corresponding decision variable and the probability to

1 (i.e., $x_2(\text{v}_\text{i}\text{-v}_\text{j}, l) = 1$ and $P(\text{v}_\text{i}\text{-v}_\text{j}, l) = 1$) and 0 otherwise. Constraint 4.9 enforces that if a v_i-v_j pair belongs to the Strongly Causal category then the label C must be assigned to its corresponding $\text{e}_{\text{v}_\text{i}}$-$\text{e}_{\text{v}_\text{j}}$ pair. Constraint 4.10 enforces that if a v_i-v_j pair belongs to the Strongly Non-causal category then the label $\neg C$ must be assigned to its corresponding $\text{e}_{\text{v}_\text{i}}$-$\text{e}_{\text{v}_\text{j}}$ pair. For the ambiguous v_i-v_j pairs, the decision for the label C and $\neg C$ is taken from the supervised classifier. We maximize the objective function Z_{KB_2} subject to the constraints of the integer linear program introduced till now.

Here, we provide background knowledge using the above mentioned hard constraints. There is a motivation behind this idea which is explained as follows. The supervised classifier identifies causality for a $\text{e}_{\text{v}_\text{i}}$-$\text{e}_{\text{v}_\text{j}}$ pair by just focusing on the linguistic features extracted from the local context of the instance of this pair. However, our metrics introduced in the previous chapter (i.e., ECA, ICA and BCA) identify the causal associations of a v_i-v_j pair by employing a large number of instances of this pair. For an ambiguous v_i-v_j pair identified by our metric, we are not certain if this pair has more likelihood to encode cause or non-cause relation. On the other hand, for a strongly causal (non-causal) v_i-v_j pair, our metric is more certain about its tendency to encode a cause (non-cause) relation, respectively. Therefore, for the strongly causal and strongly non-causal pairs we pay more importance to the decisions obtained from our metrics and for the ambiguous pairs we rely on the supervised classifier.

Note that in order to acquire the background knowledge of form KB_2 for a v_i-$\text{v}_\text{j} \notin KB_c$, we depend on the procedure of semantic mapping of verb-verb pairs introduced above to determine the semantically closest pair v_p-$\text{v}_\text{q} \in KB_c$. The information about the category of v_p-v_q pair is then used to acquire background knowledge.

4.1.3 Knowledge of Causal Semantics of Verbs

In this work, we define causal semantics of verbs in terms of the linguistic definition of events and the semantic classes of events. In the following sections, we introduce our method to plug in the knowledge of causal semantics of verbs to our model.

Linguistic Definition of Events

In our model, we stress on the fact that a $\text{e}_{\text{v}_\text{i}}$-$\text{e}_{\text{v}_\text{j}}$ pair can encode a cause relation only if at least one of the 3-tuples i.e., $\text{e}_{\text{v}_\text{i}}$ and $\text{e}_{\text{v}_\text{j}}$ represent an event according to the linguistic definition of events. We incorporate this information using the following additions to ILP:

$$Z_3 = \sum_{\mathrm{e_{v_i}\text{-}e_{v_j}} \in EP} \sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \tag{4.13}$$

$$x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg c) - x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg\mathrm{E}\text{-}\neg\mathrm{E})) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.14}$$

$$x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, c) \leq \sum_{l \in L_3 - \{\neg\mathrm{E}\text{-}\neg\mathrm{E}\}} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.15}$$

$$\sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.16}$$

$$x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0, 1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_3 \tag{4.17}$$

Here $L_3 = \{\mathrm{E\text{-}E}, \mathrm{E\text{-}}\neg\mathrm{E}, \neg\mathrm{E\text{-}E}, \neg\mathrm{E}\text{-}\neg\mathrm{E}\}$. $x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l)$ is a binary decision variable (4.17) which is set to 1 only if the label $l \in L_3$ is assigned to the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair. Constraint 4.16 enforces that only one label out of $|L_3|$ choices can be assigned to a pair $\mathrm{e_{v_i}\text{-}e_{v_j}}$. For example, if both 3-tuples of the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ are not events then the label $\neg\mathrm{E}\text{-}\neg\mathrm{E}$ is assigned to it. Constraint 4.14 enforces that if both 3-tuples are not events then the pair $\mathrm{e_{v_i}\text{-}e_{v_j}}$ must encode a non-cause relation. Constraint 4.15 implies that if a $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair encodes a cause relation then atleast one of the 3-tuples of this pair must represent an event according to the linguistic definition of events.

We use the supervised classifier introduced in chapter 3 for the labels E and $\neg$E. The probabilities of assignments of the labels E and $\neg E$ to $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ are used to compute the probabilities of labels $l \in L_3$. For example, $P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg\mathrm{E}\text{-}\neg\mathrm{E}) = P(e_{v_i}, \neg\mathrm{E})P(e_{v_j}, \neg\mathrm{E})$. We use the sum of log of probabilities for such expressions. We employ both Naive Bayes and Maximum Entropy to acquire the probabilities of the supervised classifier. We use the following expression $\log(P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg\mathrm{E}\text{-}\neg\mathrm{E})) = log(P(e_{v_i}, \neg\mathrm{E})) + log(P(e_{v_j}, \neg\mathrm{E}))$ in the function Z_3 by employing the Maximum Entropy classifier. Using Naive Bayes classifier, we take the log of probabilities as follows:

$$\log(P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg\mathrm{E}\text{-}\neg\mathrm{E})) = \sum_{k=1}^{n} logP(f_k \mid \neg\mathrm{E}) + \sum_{k=1}^{n} logP(f_k \mid \neg\mathrm{E}) \tag{4.18}$$

In our initial experiments, we observed that the probabilities ($P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l)$) where $l \in L_3$ obtained through the Naive Bayes classifier help getting better performance for the target task of identifying causality in $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pairs as compared with Maximum Entropy. Therefore in our model we provide probabilities of assignments

of $l \in L_3$ obtained through NB classifier.

We add the function Z_3 (4.13) to the function Z_{KB_1} (4.4) if we use background knowledge in the form KB_1. This results in the following objective function:

$$Z_{KB_1} = Z_{KB_1} + Z_3 \tag{4.19}$$

Similarly, we add the function Z_3 (4.13) to Z_{KB_2} (4.8) if we use the background knowledge in the form KB_2. This results in the following objective function:

$$Z_{KB_2} = Z_{KB_2} + Z_3 \tag{4.20}$$

We maximize one of the above two objective functions –i.e., either function 4.19 or 4.20 subject to the constraints of the integer linear program introduced till now.

Semantic Classes of Events

We have proposed a data intensive procedure in section 3.3.2 of chapter 3 to identify the two classes of events – i.e., $\mathrm{C_{e_v}}$ and $\neg\mathrm{C_{e_v}}$. The class $\mathrm{C_{e_v}}$ ($\neg\mathrm{C_{e_v}}$) contains the semantic classes of events with high (low) tendency to encode causation, respectively. For example, after running the above mentioned procedure on $\mathrm{Explicit_{e_{v_i}\text{-}e_{v_j}}}$, we acquired the class $\neg\mathrm{C_{e_v}} = \{\mathrm{ASPECTUAL}\}$. $\mathrm{C_{e_v}}$ contains rest of the semantic classes. Similarly, with the $\mathrm{PDTB_{e_{v_i}\text{-}e_{v_j}}}$ training corpus the above mentioned procedure yields the class $\neg\mathrm{C_{e_v}} = \{\mathrm{REPORTING}, \mathrm{STATE}\}$. In section 3.3.2, we have discussed about a supervised classifier to assign the labels $\mathrm{C_{e_v}}$ and $\neg\mathrm{C_{e_v}}$ to the events. Using the probabilities of this classifier, we add the information about the semantic classes of events as follows:

$$Z_4 = \sum_{\mathrm{e_{v_i}\text{-}e_{v_j}} \in EP} \sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \tag{4.21}$$

$$\sum_{l \in L_4 - \{\mathrm{C_{e_v}\text{-}C_{e_v}}\}} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \leq x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C) \leq \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.22}$$

$$x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{C_{e_v}\text{-}C_{e_v}}) - x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, C)) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.23}$$

$$\sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \tag{4.24}$$

$$x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \quad \in EP \quad \forall l \in L_4 \tag{4.25}$$

Here $L_4 = \{\mathrm{C_{e_v}\text{-}C_{e_v}}, \mathrm{C_{e_v}\text{-}\neg C_{e_v}}, \mathrm{\neg C_{e_v}\text{-}C_{e_v}}, \mathrm{\neg C_{e_v}\text{-}\neg C_{e_v}}\}$. $x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l)$ is a binary decision variable (4.25) which is set to 1 only if the label $l \in L_4$ is assigned to the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair. Constraint 4.24 enforces that only one label out of $|L_4|$ choices can be assigned to a $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair. For example, if both events of the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair have a low tendency to encode a causal relation, then the label $\mathrm{\neg C_{e_v}\text{-}\neg C_{e_v}}$ is assigned to it. Constraint 4.22 enforces that if any one of two events of $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair has a low tendency to encode causation i.e., $l \in L_4 - \{\mathrm{C_{e_v}\text{-}C_{e_v}}\}$ then the $\neg C$ label must be assigned to the $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair. Constraint 4.23 implies that if a $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair encodes a cause relation then $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pair must be assigned the label $\mathrm{C_{e_v}\text{-}C_{e_v}}$ because both $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ have a tendency to encode causation.

The probabilities of assignments of the labels $\mathrm{C_{e_v}}$ and $\mathrm{\neg C_{e_v}}$ to $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ are used to compute the probabilities for the labels $l \in L_4$. For example, $P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{\neg C_{e_v}\text{-}\neg C_{e_v}}) = P(e_{v_i}, \mathrm{\neg C_{e_v}})P(e_{v_j}, \mathrm{\neg C_{e_v}})$. We use the sum of log of probabilities for such expressions. We employ both Naive Bayes and Maximum Entropy to acquire the probabilities of the supervised classifier. We use the following expression $\log(P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{\neg C_{e_v}\text{-}\neg C_{e_v}})) = log(P(e_{v_i}, \mathrm{\neg C_{e_v}}))+log(P(e_{v_j}, \mathrm{\neg C_{e_v}}))$ in the function Z_4 by employing the Maximum Entropy classifier. Using Naive Bayes classifier, we take the log of probabilities as follows:

$$\log(P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{\neg C_{e_v}\text{-}\neg C_{e_v}})) = \sum_{k=1}^{n} logP(f_k \mid \mathrm{\neg C_{e_v}}) + \sum_{k=1}^{n} logP(f_k \mid \mathrm{\neg C_{e_v}}) \tag{4.26}$$

In our initial experiments, we observed that the probabilities ($P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l)$) where $l \in L_4$ obtained through the Naive Bayes classifier help getting better performance for the target task of identifying causality in $\mathrm{e_{v_i}\text{-}e_{v_j}}$ pairs as compared with Maximum Entropy. Therefore in our model we provide probabilities of assignments of $l \in L_4$ obtained through NB classifier.

We add the function Z_4 (4.21) to the function Z_{KB_2} (4.19) if we use the background knowledge in the form KB_1. This results in the following objective function:

$$Z_{KB_1} = Z_{KB_1} + Z_4 \tag{4.27}$$

Similarly, we add the function Z_4 (4.21) to Z_{KB_2} (4.20) if we provide the background knowledge using the form KB_2. This results in the following objective function:

$$Z_{KB_2} = Z_{KB_2} + Z_4 \tag{4.28}$$

We maximize one of the above two objective functions –i.e., either function 4.27 or 4.28 subject to the constraints of the integer linear program introduced till now.

In order to assist the reader, we provide the complete integer linear programs using background knowledge KB_1 or KB_2 in the next two sections.

4.1.4 Integer Linear Program: ILP$_{KB_1}$

This section provides the integer linear program with background knowledge of form KB_1.

$$
\begin{aligned}
Z_{KB_1} \;=\; & \max \sum_{\mathrm{e_{v_i}\text{-}e_{v_j}} \in EP} x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, C)(RS(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}})) \times P(\mathrm{e_{v_i}\text{-}e_{v_j}}, C)) \\
+ \;& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C)((1.0 - RS(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}})) \times P(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C)) \\
+ \;& \sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) + \sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg c) - x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg \mathrm{E}\text{-}\neg \mathrm{E})) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, c) \leq \sum_{l \in L_3 - \{\neg \mathrm{E}\text{-}\neg \mathrm{E}\}} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_4 - \{\mathrm{C_{e_v}\text{-}C_{e_v}}\}} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \leq x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C) \leq \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{C_{e_v}\text{-}C_{e_v}}) - x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, C) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_1} x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_1 \\
& x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_3 \\
& x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_4
\end{aligned}
$$

4.1.5 Integer Linear Program: ILP$_{KB_2}$

This section provides the integer linear program with background knowledge of form KB_2.

$$
\begin{aligned}
Z_{KB_2} \;=\; & \max \sum_{\mathrm{e_{v_i}\text{-}e_{v_j}} \in EP} \sum_{l \in L_1} x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) + \sum_{l \in L_2} x_2(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), l) P(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), l) \\
+ \; & \sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) + \sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) P(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, C) - x_2(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), S_c) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C) - x_2(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), S_{\neg c}) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg c) - x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg \mathrm{E}\text{-}\neg \mathrm{E})) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, c) \leq \sum_{l \in L_3 - \{\neg \mathrm{E}\text{-}\neg \mathrm{E}\}} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_4 - \{\mathrm{C_{e_v}\text{-}C_{e_v}}\}} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \leq x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, \neg C) \leq \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, \mathrm{C_{e_v}\text{-}C_{e_v}}) - x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, C) \geq 0 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_1} x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_2} x_2(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_3} x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& \sum_{l \in L_4} x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) = 1 \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \\
& x_1(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_1 \\
& x_2(f_{vp}(\mathrm{e_{v_i}\text{-}e_{v_j}}), l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_2 \\
& x_3(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP, \quad \forall l \in L_3 \\
& x_4(\mathrm{e_{v_i}\text{-}e_{v_j}}, l) \in \{0,1\} \quad \forall\, \mathrm{e_{v_i}\text{-}e_{v_j}} \in EP \quad \forall l \in L_4
\end{aligned}
$$

4.2 Empirical Study

Our model employing various sources of knowledge needs to be evaluated against the baseline model which lacks the knowledge necessary for identifying causality. Therefore, in this empirical study we first assess performance of the supervised classifier which merely depends on the linguistics features introduced in section 3.1.2. In other words, we assess performance of our model relying only on the knowledge of context. After the assessment of supervised classifier, we present performance of our model by incrementally adding the background knowledge and the knowledge of causal semantics of verbs in our model.

4.2.1 Evaluation Data

Before presenting performance of our model for identifying causality, we need to provide details of the test sets we employ for this purpose. In this research, we have proposed a model to identify causal relations in both intra- and inter-sentential instances of e_{v_i}-e_{v_j} pairs. Therefore, we need a test set with both types of instances. To the best of our knowledge, Bethard and Martin (2008) were the first to introduce a data set of 1000 e_{v_i}-e_{v_j} pairs labeled with C and $\neg C$ relations. However, their data set consists of only intra-sentential instances of e_{v_i}-e_{v_j} pairs and in these instances two verbs are conjoined with the marker "and". Therefore, for this research we have generated a test set of both intra- and inter-sentential instances of e_{v_i}-e_{v_j} pairs and in these instances both events can appear in any context i.e., explicit and unambiguous, ambiguous and implicit contexts. In order to generate this test set, we randomly selected 100 verb-verb pairs from the list of $10,774$ verb-verb pairs in KB_c. For each pair, we randomly selected 3 intra- and 3 inter-sentential instances from the English Gigaword corpus and the "Hurricane Katrina" and "Iraq war" articles. This test set has 600 instances in it (with 50% intra- and 50% inter-sentential instances). In order to keep the development set different from the test set, we automatically traversed the development set to determine if any test instance is available in it or not. In case of finding any such test instance, we removed it from the development set to perform evaluation on unseen test instances. Two human annotators were asked to provide the label C or $\neg C$ for each instance. In order to make the annotations task easier for the inter-sentential instances, we randomly selected those inter-sentential instances for our test set in which the length of each of both sentences is at most 40 words. This helped boosting human inter-annotator agreement on our test set.

In addition to the above requirement of having both intra- and inter-sentential instances in a test set, we also need a high human inter-annotator agreement for the labels C and $\neg C$. For this purpose, we provide an objective notion of causality [Beamer and Girju 2009] to two human annotators for the assignments of above two labels to the e_{v_i}-e_{v_j} pairs. This notion is based on the Manipulation Theory of Causality [Woodward 2008] reproduced below from chapter 1.

> Manipulation theory of causality determines truth of the following two conditions to determine if a cause-effect relation is encoded between the two events **a** and **b** or not: (1) event **a** must temporally precede or overlap event **b** in time and (2) while keeping as many state of current affairs constant as possible, modifying event **a** must entail predictably modifying event **b** [Woodward 2008, Beamer and Girju 2009].

Based on the manipulation theory of causality, we gave the following guidelines to two human annotators for labeling the instances of e_{v_i}-e_{v_j} pairs with C and $\neg C$.

Test-set	Total	Test Instances	% C	% Agreement	Kappa
Test-set$_1$	600	536	23.69	89.33	0.72
Test-set$_2$	–	1000	27.10	77.80	0.55

Table 4.1: The total number of instances (Total), the number of total instances on which human annotators agreed and these instances are used for evaluation (Test Instances), the percentage of "Test Instances" with the label C (%C), the percentage of "Total" instances on which human annotators agreed to each other (% Agreement) and kappa value for the human inter-annotator agreement on the "Total" instances (Kappa). Bethard and Martin (2008) have not provided the total number of instances on which their human annotators applied the labels C and $\neg C$.

> Assign the label C to an instance of e_{v_i}-e_{v_j} pair only if the two truth conditions of the manipulation theory of causality are satisfied and no additive relation (list, continuation, opposition, exception, enumeration, temporal, and concession) can be recognized from the discourse markers or other elements of the context of instance. Otherwise, assign the label $\neg C$. Also, assign the label $\neg C$ if the annotator assumes that two events are not even relevant to each other in the current context.

We refer the reader to Appendix A of this thesis for details of application of these guidelines to the instances of e_{v_i}-e_{v_j} pairs.

In order to evaluate our model we employ two test sets (1) our test set generated using the above mentioned guidelines (named as Test-set$_1$) and (2) the data set of 1000 e_{v_i}-e_{v_j} pairs generated by Bethard and Martin (2008) (named as Test-set$_2$). Notice that in all instances of Test-set$_2$ two events are conjoined by the marker "and" and thus these events appear in ambiguous context. On the other hand in the instances of Test-set$_1$ two events can appear in any context i.e., explicit and unambiguous, ambiguous and implicit contexts. We have achieved 0.72 kappa value for the human-inter annotator agreement on Test-set$_1$ and this test set contains 23.69% causal instances (see Table 4.1). Similarly Bethard and Martin (2008) have reported 0.55 kappa value for the human-inter annotator agreement on Test-set$_2$ and this test set contains 27.10% causal instances (Table 4.1).

4.2.2 Assessment of the Knowledge of Context

In this section we evaluate performance of the supervised classifier relying merely on the linguistic features introduced in section 3.1.2. In order to build the supervised classifier we utilize one of the two training corpora Explicit$_{e_{v_i}\text{-}e_{v_j}}$ and PDTB$_{e_{v_i}\text{-}e_{v_j}}$. Using one of the above mentioned training corpora, we train the supervised classifier with both Naive Bayes (NB) and Maximum Entropy (MaxEnt) classification algorithms. The Test-set$_2$ from Bethard and Martin (2008) contains instances extracted from the Penn TreeBank (PTB). Therefore, we remove the test instances of the Test-set$_2$ that were found to be present in the PDTB$_{e_{v_i}\text{-}e_{v_j}}$

Test sets	Score	$\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$	$\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$
Test-set_1	Accuracy	31.52	42.53
	Precision	23.33	23.76
	Recall	82.67	64.56
	F-score	36.39	34.74
Test-set_2	Accuracy	35.50	56.80
	Precision	25.84	31.57
	Recall	70.84	50.92
	F-score	37.31	38.98

Table 4.2: The performance of supervised classifiers on Test-set_1 and Test-set_2 using the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ and $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ training corpora. The results are provided using NB classifier. Using the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus, MaxEnt classifier gives a very low F-score of around 20% on both test sets. Using the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus, MaxEnt classifier produces 34.13% (35.21%) F-score on Test-set_1 (Test-set_2), respectively.

corpus. We found 99 such instances in the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus with $5,277$ C and $16,640$ $\neg C$ instances. After removing the above mentioned 99 instances from the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus, we are left with a total of $5,258$ C and $16,560$ $\neg C$ instances. In order to avoid over-fitting towards the class $\neg C$, we employ equal number of instances of both labels –i.e., $5,258$ instances for each of both labels. For the purpose of training, we randomly selected $5,258$ instances from the $16,560$ non-cause instances of the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus.

Table 4.2 provides results on Test-set_1 and Test-set_2 using the supervised classifier. As shown in this table, F-score obtained with NB classifier is better than the F-score achieved using MaxEnt classifier. The results in Table 4.2 reveal that the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ (a large training corpus) results in a very high recall as compared with the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus. On the other hand, in comparison with the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus, $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ achieves a high accuracy by compromising recall. For Test-set_2, supervision from the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ results in 1.67% improvement in F-score achieved through the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$. The reason is the availability of gold-standard features for both $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ and Test-set_2. However, the gold-standard features are not always available to achieve a better result. For example, for the Test-set_1 with the non-availability of the gold-standard features we observe a drop in F-score from the classifier trained using the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus. On this test set, the supervision from $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ produces 1.65% drop in F-score achieved through the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus. Also notice that Test-set_1 (Test-set_2) contains 23.69% (27.01%) causal instances, receptively. Since the real distribution of labels is highly skewed it is easy to achieve a very high accuracy by compromising recall. In this situation, we need a model which is capable of providing better accuracy but not at the cost of recall. In the next sections we show that with the addition of more sources of knowledge (i.e., background knowledge and causal semantics of verbs), our model provides higher accuracy as well as recall as compared with the baseline of supervised classifier. Also we observe improvements in F-score over the baseline model.

In this research, our objective is to empirically evaluate performance of our model with the addition of novel sources of knowledge – i.e., background knowledge and the knowledge of causal semantics of verbs. For this purpose, we pick the current supervised models trained using the Explicit$_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ and PDTB$_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpora and add the above stated types of knowledge to these models. We choose the supervised classifiers trained using NB algorithm because MaxEnt algorithm gives lower F-scores as shown in Table 4.2.

4.2.3 Assessment of Background Knowledge

We provide background knowledge to our model in one of the two following forms: (1) KB_1 and (2) KB_2. In this section, the performance of our model is provided with the addition of each form of background knowledge. The form KB_1 employs the ranking scores of verb-verb pairs with respect to their likelihood to encode causation and the form KB_2 considers three categories of these pairs. We introduced three metrics (ECA, ICA and BCA) in section 3.2 to determine the scores of likelihood of causality in verb-verb pairs and based on these scores the background knowledge is derived for our model. Therefore, in this section we provide results using background knowledge derived via all of the above mentioned metrics. In addition to this, for comparison with prior work we also acquire and provide background knowledge to our model using the unsupervised state-of-the-art metric Cause-Effect Association (CEA) [Do et al. 2011]. We employ CEA in the following way to calculate the likelihood of causality in a v_i-v_j pair:

$$\frac{1}{\mid VP \mid} \sum_{I_{\text{v}_\text{i}\text{-v}_\text{j}} \in VP} (CEA(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}})) \tag{4.29}$$

where VP is the set of all unlabeled intra- and inter-sentential instances of the v_i-v_j pair. The score of the likelihood of causality in the v_i-v_j pair is computed by taking the average of CEA values over all instances $I_{\text{v}_\text{i}\text{-v}_\text{j}} \in VP$. Based on the above score from the equation 4.29, we identify the ranking scores and three categories of all verb-verb pairs. CEA [Do et al. 2011] (4.30) is an unsupervised score for identifying the causal association in a $\text{e}_{\text{v}_\text{i}}$-$\text{e}_{\text{v}_\text{j}}$ pair.

$$CEA(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}) = s_{pp}(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}) + s_{pa}(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}) + s_{aa}(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}) \tag{4.30}$$

Here, $s_{pp}(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}})$ is same as the score CD (3.2) for measuring causal dependency between two verbs. Do et al., (2011) also penalized the score CD with the function $Dist(\text{v}_\text{i}\text{-v}_\text{j})$. The $Dist(\text{v}_\text{i}\text{-v}_\text{j})$ function [Do et al. 2011] is a penalization factor which assumes that two verbs appearing in the same sentence have a high tendency to encode causation as compared with the verbs appearing in the adjacent sentences. $s_{pa}(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}})$ computes the average of PMI associations for the pairs v_i-subject$_{\text{v}_\text{j}}$, v_i-object$_{\text{v}_\text{j}}$ and the pairs

v_j-subject$_{v_i}$, v_j-object$_{v_i}$. Similarly, s_{aa}(e_{v_i}-e_{v_j}) computes the average of PMI associations for the pairs of arguments of verbs e.g., the pairs subject$_{v_i}$-subject$_{v_j}$, subject$_{v_i}$-object$_{v_j}$, etc. We refer the reader to Do et al. (2011) for details of the metric CEA.

Before acquiring performance of our model with the addition of background knowledge, notice that the verb-verb pairs of all instances of Test-set$_1$ $\in KB_c$. However, for Test-set$_2$ there are only 242 instances s.t. their verb-verb pairs exist in KB_c. Therefore, for the Test-set$_2$ we depend on the procedure of semantic mapping of verb-verb pairs to acquire background knowledge (see section 4.1.2 for details). Using this procedure, we are able to acquire background knowledge for 83.5% instances of the Test-set$_2$. Here, we introduce another test set named as "Test-set$_3$". It consists of 242 test instances of the Test-set$_2$ s.t. the verb-verb pairs of all these instances are available in KB_c. The Test-set$_3$ contains 26.03% causal instances. Now for all instances of Test-set$_1$ and Test-set$_3$ we can obtain the exact background knowledge from the resource KB_c. On the other hand, for the Test-set$_2$ we depend on the procedure of semantic mapping of verb-verb pairs to acquire the approximate background knowledge using semantically closest verb-verb pairs in KB_c.

In this section, we first provide performance of our model after the addition of background knowledge to the supervised classifier trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus (Tables 4.3 and 4.4) and then provide performance after the addition of background knowledge to the supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus (Tables 4.5 and 4.6).

Table 4.3 (4.4) provides performance of our model with the background knowledge of form KB_1 (KB_2), respectively. The results given in the Tables 4.3 and 4.4 reveal that the unsupervised metric CEA does not provide a better source of background knowledge as compared with the metrics ECA, ICA and BCA. This makes sense because the metric CEA acquires background knowledge with no supervision. Also, the metric CEA relies mainly on the PMI association measure to derive the causal associations in verb-verb pairs. The PMI measure is not capable to distinguish causality from any other type of correlation. However, the background knowledge extracted with the metrics ECA, ICA and BCA has proven to be useful because these measures incorporate supervision from the automatically generated training corpus of cause and non-cause relations to distinguish causality from any other type of correlation. In addition to this, our advanced metrics ICA and BCA provide better results as compared with ECA. In comparison with ECA, the metrics ICA and BCA take care of the ambiguous and implicit contexts of causal associations and thus provide better recall most of the times while maintaining accuracy, precision and F-score (see Tables 4.3 and 4.4). The background knowledge KB$_1$ from ICA achieves 5.17% (2.48%) (1.21%) improvement in F-score acquired from the supervised classifier on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively (see Table 4.3). Similarly, the

Test-set	Score	Explicit$_{e_{v_i}\text{-}e_{v_j}}$	+ CEA	+ ECA	+ ICA	+ BCA
Test-set$_1$	Accuracy	31.52	44.96	50.37	52.98	52.42
	Precision	23.33	25.14	26.91	29.50	29.35
	Recall	82.67	66.92	63.77	70.86	71.65
	F-score	36.39	36.55	37.85	41.66	41.64
Test-set$_2$	Accuracy	35.50	43.40	43.70	41.00	41.40
	Precision	25.84	24.78	27.18	27.50	27.59
	Recall	70.84	53.50	64.20	71.90	71.58
	F-score	37.31	33.87	38.19	39.79	39.83
Test-set$_3$	Accuracy	32.64	44.62	45.04	39.66	39.66
	Precision	25.49	22.90	28.12	27.07	26.81
	Recall	82.53	47.61	71.42	77.77	76.19
	F-score	38.95	30.92	40.35	40.16	39.66

Table 4.3: The performance of supervised classifier trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge of form KB_1 to the supervised classifier. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA.

Test-set	Score	Explicit$_{e_{v_i}\text{-}e_{v_j}}$	+ CEA	+ ECA	+ ICA	+ BCA
Test-set$_1$	Accuracy	31.52	43.47	43.84	47.57	47.01
	Precision	23.33	25.82	25.96	28.49	28.61
	Recall	82.67	74.01	74.01	80.31	82.67
	F-score	36.39	38.28	38.44	42.06	42.51
Test-set$_2$	Accuracy	35.50	39.20	40.40	38.00	37.80
	Precision	25.84	24.66	26.61	26.63	26.56
	Recall	70.84	60.51	68.26	73.43	73.43
	F-score	37.31	35.04	38.30	39.09	39.01
Test-set$_3$	Accuracy	32.64	38.01	40.49	37.60	35.21
	Precision	25.49	23.95	27.62	27.77	26.26
	Recall	82.53	63.49	79.36	87.30	82.53
	F-score	38.95	34.78	40.98	42.14	39.84

Table 4.4: The performance of supervised classifier trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge of form KB_2 to the supervised classifier. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA.

background knowledge KB$_1$ from ICA achieves 21.46% (5.5%) (7.02%) improvement in accuracy acquired from the supervised classifier on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively (see Table 4.3). These results are quite encouraging and reveal that the background knowledge is very important to achieve progress on this task. Our model achieves even better F-scores when the background knowledge is employed in the form KB_2 i.e., in terms of three categories of the verb-verb pairs. For example, the background knowledge KB_2 from ICA achieves 5.67% (1.78%) (3.19%) improvement in F-score obtained from the supervised classifier on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively (see Table 4.4). However, the background knowledge KB_2 results in less improvements in accuracy as compared with the background knowledge KB_1. The reason is that KB_1 considers the background knowledge in terms of the strict ranking scores of verb-verb pairs and KB_2 considers the background knowledge in terms of loosely defined categories of verb-verb pairs.

Test-set	Score	$\mathbf{PDTB}_{e_{v_i}\text{-}e_{v_j}}$	+ CEA	+ ECA	+ ICA	+ BCA
$\mathbf{Test\text{-}set}_1$	Accuracy	42.53	45.33	50.55	52.98	52.79
	Precision	23.76	25.14	26.84	29.23	29.41
	Recall	64.56	66.14	62.79	69.29	70.86
	F-score	34.74	36.44	37.64	41.12	41.57
$\mathbf{Test\text{-}set}_2$	Accuracy	56.80	46.70	47.30	44.60	44.70
	Precision	31.57	25.74	28.45	28.59	28.63
	Recall	50.92	51.29	62.36	69.74	69.74
	F-score	38.98	34.27	39.07	40.55	40.60
$\mathbf{Test\text{-}set}_3$	Accuracy	52.47	45.04	45.04	41.73	40.49
	Precision	28.33	23.07	28.12	27.84	27.62
	Recall	53.96	47.61	71.42	77.77	79.36
	F-score	37.15	31.08	40.35	41.00	40.98

Table 4.5: The performance of supervised classifier trained using the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge of form KB_1 to the supervised classifier. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA.

Now we provide performance of our model after the addition of background knowledge to the supervised classifier trained using the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus (Tables 4.5 and 4.6). With this setting, we also observe improvements in F-scores with the addition of background knowledge derived using the advanced metrics ICA and BCA. The addition of background knowledge KB_1 from ICA results in 6.38% (1.57%) (3.85%) improvement in F-score acquired from the supervised classifier on Test-set_1 (Test-set_2) (Test-set_3), respectively (see Table 4.5). Similarly, the addition of background knowledge KB_2 from ICA achieves 8.4% (1.95%) (4.91%) improvement in F-score acquired from the supervised classifier on Test-set_1 (Test-set_2) (Test-set_3), respectively (see Table 4.6). On Test-set_1, the model +ICA achieves more than 10% increase in accuracy over the supervised classifier (see Tables 4.5 and 4.6). However, we observe a trade off between accuracy and F-score given in Tables 4.5 and 4.6. On Test-set_2 and Test-set_3, the supervised classifier compromises F-score and recall to boost accuracy. On the other hand, the models +ICA and +BCA boost F-score at the cost of accuracy. Since Test-set_2 (Test-set_3) contains 27.10% (26.03%) causal instances, it is easy to achieve a very high accuracy by compromising recall. Therefore in this situation F-score and recall are better measures of evaluation. In comparison with the supervised classifier our models +ICA and +BCA provide encouraging results by not compromising recall and F-score in the favor of accuracy. An important observation from the results provided in Tables 4.3, 4.4, 4.5 and 4.6 is regarding the availability of exact and approximate background knowledge for the e_{v_i}-e_{v_j} pairs. Recall that the exact background knowledge is available for Test-set_1 and Test-set_3 and the approximate background knowledge is provided for Test-set_2. As expected, the improvements in F-score for Test-set_1 and Test-set_3 are mostly higher than the improvements in F-score for Test-set_2 (see Tables 4.3, 4.4, 4.5 and 4.6). However, the encouraging trend is the improvements in F-scores using both exact and approximate background knowledge.

Test-set	Score	$\textbf{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$	+ CEA	+ ECA	+ ICA	+ BCA
$\textbf{Test-set}_1$	Accuracy	42.53	46.64	49.81	52.79	51.11
	Precision	23.76	25.63	28.08	30.18	29.60
	Recall	64.56	66.14	71.65	75.59	77.16
	F-score	34.74	37.00	40.35	43.14	42.79
$\textbf{Test-set}_2$	Accuracy	56.80	48.80	47.50	46.90	45.50
	Precision	31.57	27.56	28.10	29.29	28.54
	Recall	50.92	54.61	60.14	67.89	67.52
	F-score	38.98	36.63	38.30	40.93	40.13
$\textbf{Test-set}_3$	Accuracy	52.47	43.80	43.80	44.21	39.25
	Precision	28.33	24.47	27.32	28.82	26.13
	Recall	53.96	55.55	69.84	77.77	73.01
	F-score	37.15	33.98	39.28	42.06	38.49

Table 4.6: The performance of supervised classifier trained using the $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus and the models with the addition of background knowledge of form KB_2 to the supervised classifier. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA.

In this work, we also compare performance of the supervised classifiers trained using the $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ and $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpora and the models with the addition of background knowledge (i.e., +CEA, +ECA, +ICA and +BCA) on a interpolated precision-recall curve (see Figure 4.1). Our objective is to observe performances of all models on various levels of recall. We combine Test-set_1 and Test-set_2 and for this combined test set we set up the interpolated precision-recall curve using the following scheme:

- A supervised classifier predicts the label C if $P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) \geq \gamma$ and $\neg C$ otherwise. Here $P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C)$ is the probability of assignment of the label C to the $\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}$ pair. We vary the threshold γ from 0.1 to 1.0 with the increments of 0.1 and observe the precision and recall of the supervised classifier obtained using each value of threshold (Figure 4.1).

- We supply the background knowledge of form KB_1 to the supervised classifier and observe performance of this model using the precision-recall curve (Figure 4.1). Using the objective function 4.4, the model with background knowledge predicts the label C if $RS(\text{v}_\text{i}\text{-v}_\text{j})P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) \geq (1 - RS(\text{v}_\text{i}\text{-v}_\text{j})P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, \neg C))$ and $\neg C$ otherwise. We compute the scores (4.31 and 4.32) and predict the label C if $S(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) \geq \gamma$ and $\neg C$ otherwise. We vary the threshold γ from 0.1 to 1.0 with the increments of 0.1 and observe the precision and recall of the model with background knowledge KB_1.

$$S(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) = \frac{RS(\text{v}_\text{i}\text{-v}_\text{j})P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C)}{RS(\text{v}_\text{i}\text{-v}_\text{j})P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) + (1 - RS(\text{v}_\text{i}\text{-v}_\text{j}))P(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, \neg C)} \tag{4.31}$$

$$S(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, \neg C) = 1 - S(\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}, C) \tag{4.32}$$

The precision values shown in the Figure 4.1 are interpolated precision values computed as follows:

$$IntPrecision(r) = max_{i \geq r} Precision(r) \tag{4.33}$$

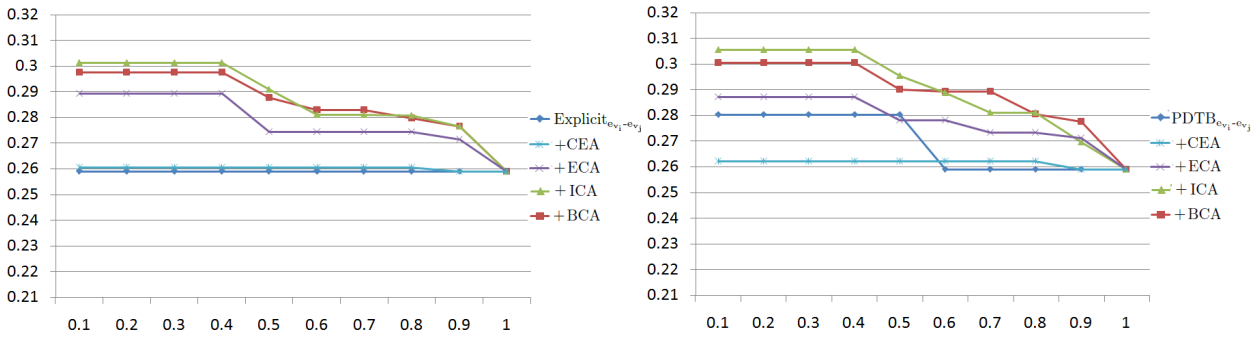

Figure 4.1: The interpolated precision-recall curves for the supervised classifier and the models with the addition of background knowledge of form KB_1. The supervised classifiers are trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ (shown on left) and PDTB$_{e_{v_i}\text{-}e_{v_j}}$ (shown on right) corpora. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA. The threshold γ increases in the increments of 0.1 from left to right and produces different precision and recall values for each of the above stated models.

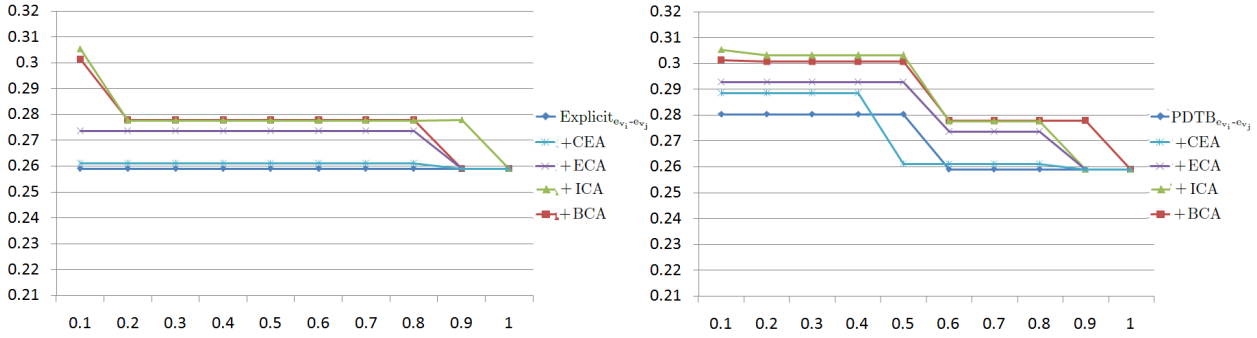

Figure 4.2: The interpolated precision-recall curves for the supervised classifier and the models with the addition of background knowledge of form KB_2. The supervised classifiers are trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ (shown on left) and PDTB$_{e_{v_i}\text{-}e_{v_j}}$ (shown on right) corpora. The background knowledge is acquired using one of the following four metrics: (1) CEA, (2) ECA, (3) ICA and (4) BCA. The threshold γ increases in the increments of 0.1 from left to right and produces different precision and recall values for each of the above stated models.

Figure 4.1 reveals that the models with background knowledge from the metrics ECA, ICA and BCA provide better precision values on almost all recall levels from 0.1 (1%) to 0.9 (90%) as compared with the baselines of supervised classifiers trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ and PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpora. ICA and BCA perform better on all recall levels as compared with ECA by taking care of explicit and unambiguous, ambiguous and implicit contexts of causal associations of verb-verb pairs. BCA provides better precision as compared with ICA on the recall values greater than 0.6. BCA performs better on high recall values

by combining both ECA and ICA metrics where ECA takes care of the explicit and unambiguous contexts of verb-verb pairs and ICA pays more attention to the ambiguous and implicit contexts of verb-verb pairs. As compared with the metrics ECA, ICA and BCA, precision of the supervised classifier trained using the $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus remains better than the model +CEA on the recall values from 0.1 to 0.5. It is evident that the unsupervised metric CEA does not provide a better source of background knowledge and thus results in drop of precision as compared with the advanced metrics ECA, ICA and BCA. An interesting observation from Figure 4.1 is regarding the performance of supervised classifier trained using the $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus. The $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus provides better precision on the recall values from 0.1 to 0.5 than the $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus. The $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus provides better supervision because of the following three reasons: (1) manual annotations of contingency and non-contingency relations, (2) gold-standard linguistic features and (2) explicit and unambiguous, ambiguous and implicit training instances. However, after the recall of 0.5 both $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ and $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpora provide similar results.

We also acquire interpolated precision-recall curve for the models using background knowledge of form KB_2 (see Figure 4.2). To generate the precision-recall curves for these models, we forcefully set the label C ($\neg C$) for the pairs $\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}$ with $\text{v}_\text{i}\text{-v}_\text{j} \in S_c(S_{\neg c})$ category, respectively. For the rest of the verb-verb pairs, we acquire predictions from the supervised classifier. Figure 4.2 reveals that on all recall values the models +ICA and +BCA provide better precision values than the supervised classifiers. On top of the supervised classifier trained via $\text{Explicit}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus, the background knowledge KB_1 from ICA and BCA produce better precision values than the background knowledge KB_2 from these metrics (see Figure 4.1 and 4.2). KB_2 provides lower precision values because it considers naive categorization of verb-verb pairs and KB_1 provides background knowledge in the form of strict ranking scores of verb-verb pairs. On top of the supervised classifier trained using the $\text{PDTB}_{\text{e}_{\text{v}_\text{i}}\text{-e}_{\text{v}_\text{j}}}$ corpus, the background knowledge KB_1 from BCA produces better precision values than the background knowledge KB_2 from BCA on the recall values 0.6 to 0.9. KB_1 and KB_2 from BCA produce almost same precision values for the recall values from 0.1 to 0.4.

As it is evident from Figures 4.1 and 4.2, the metrics ICA and BCA provide a better source of background knowledge. Also the background knowledge of form KB_1 provides better precision values as compared with the background knowledge of form KB_2. Therefore from now onward, we use background knowledge of form KB_1 derived through the metrics ICA (denoted by $KB_{1_{ICA}}$) and BCA (denoted by $KB_{1_{BCA}}$). In the next section, we assess performance of our model by adding the knowledge of causal semantics of verbs.

4.2.4 Assessment of the Causal Semantics of Verbs

We incorporate the knowledge of causal semantics of verbs by providing information about the linguistic definition of events and the semantic classes of events with a high and low tendency to encode causation. In order to add information regarding the semantic classes of events we set the class $\neg C_{e_v}$ {REPORTING} or {REPORTING, STATE} or {ASPECTUAL}. The approach to derive and incorporate information about the classes C_{e_v} and $\neg C_{e_v}$ in our model is explained in sections 3.3.2 and 4.1.3.

In this section, we first provide results by adding the knowledge of causal semantics of verbs to the model employing the supervised classifier trained via $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus and the background knowledge $KB_{1_{ICA}}$ or $KB_{1_{BCA}}$ (Tables 4.7 and 4.8). In particular, Tables 4.7 and 4.8 show results for the following models:

- The supervised classifier trained via $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus (i.e., the column $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$).
- The model with the supervised classifier and background knowledge (i.e., the column $+KB_{1_{ICA}}$ or $+KB_{1_{BCA}}$).
- The model with the supervised classifier, background knowledge and information about the linguistic definition of events (i.e., the column +LD).
- The model with the supervised classifier, background knowledge, information about the linguistic definition of events and the semantic classes of events (i.e., the column $+\neg C_{e_v} = \{R\}$ or $+\neg C_{e_v} = \{R, S\}$ or $+\neg C_{e_v} = \{A\}$ where R (S) (A) stands for REPORTING (STATE) (ASPECTUAL), respectively.)

As it is shown in Tables 4.7 and 4.8, information about the linguistic definition of events (i.e., the model +LD) brings only minute changes in the F-score and accuracy for Test-set$_2$. This reveals that the Riaz and Girju (2010)'s assumption that the main verbs normally represent events is mostly correct.

Now, we examine performance of our model after addition of the semantic classes of events with a high and low tendency to encode causation. Using Procedure 3.1 given in section 3.3.2, the manually annotated $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus yields the class $\neg C_{e_v}$={REPORTING, STATE} where REPORTING events have a high tendency to encode non-causation than STATE events. Similarly, with the $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ corpus we obtain $\neg C_{e_v}$={ASPECTUAL}. In this work we test our model by setting $\neg C_{e_v}$={REPORTING} or $\neg C_{e_v}$={REPORTING, STATE} or $\neg C_{e_v}$={ASPECTUAL}. The results in Tables 4.7 and 4.8 reveal that performance of the models with $\neg C_{e_v} = \{\text{ASPECTUAL}\}$ drops in most of the cases. However, the models with $\neg C_{e_v} = \{\text{REPORTING}\}$ and $\neg C_{e_v} = \{\text{REPORTING, STATE}\}$ bring improvements in F-score as well accuracy in most of the cases. This validates the output of Procedure 3.1 with the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus that the REPORTING and STATE events have the highest tendency to encode non-causation. For example, the

Test-set	S	Explicit$_{e_{v_i}\text{-}e_{v_j}}$	+ $KB_{1_{ICA}}$	+LD	+¬C$_{e_v}$ = {R}	+¬C$_{e_v}$ = {R, S}	+¬C$_{e_v}$ = {A}
Test-set$_1$	A	31.52	52.98	52.98	58.20	58.58	54.10
	P	23.33	29.50	29.50	32.10	31.51	29.83
	R	82.67	70.86	70.86	68.50	63.77	69.29
	F	36.39	41.66	41.66	43.71	42.18	41.70
Test-set$_2$	A	35.50	41.00	41.50	44.80	46.30	42.50
	P	25.84	27.50	27.69	28.54	28.75	26.89
	R	70.84	71.90	71.95	69.00	66.40	65.31
	F	37.31	39.79	39.90	40.38	40.13	38.10
Test-set$_3$	A	32.64	39.66	39.66	43.38	43.80	40.90
	P	25.49	27.07	27.07	28.23	28.14	24.68
	R	82.53	77.77	77.77	76.19	74.60	61.90
	F	38.95	40.16	40.16	41.20	40.86	35.29

Table 4.7: The performance of supervised classifier trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge (i.e., + $KB_{1_{ICA}}$) and the knowledge of causal semantics of verbs. The knowledge of causal semantics of verbs is incorporated by adding information about the linguistic definition of events (i.e., +LD) and the semantic classes of events with a high and low tendency to encode causation (i.e., +¬C$_{e_v}$ ={(R)EPORTING} OR +¬C$_{e_v}$ ={(R)EPORTING, (S)TATE} OR +¬C$_{e_v}$ ={(A)SPECTUAL}). The performance is provided in terms of the following (S)cores: (A)ccuracy, (P)recision, (R)ecall and (F)-score.

Test-set	S	Explicit$_{e_{v_i}\text{-}e_{v_j}}$	+ $KB_{1_{BCA}}$	+LD	+¬C$_{e_v}$ = {R}	+¬C$_{e_v}$ = {R, S}	+¬C$_{e_v}$ = {A}
Test-set$_1$	A	31.52	52.42	52.42	57.83	58.02	53.35
	P	23.33	29.34	29.34	32.00	31.43	29.15
	R	82.67	71.65	71.65	69.29	65.35	67.61
	F	36.39	41.64	41.64	43.78	42.45	40.75
Test-set$_2$	A	35.50	41.40	41.90	45.30	46.80	42.90
	P	25.84	27.59	27.79	28.70	28.91	26.99
	R	70.84	71.58	71.58	68.63	66.05	64.94
	F	37.31	39.83	40.04	40.47	40.22	38.13
Test-set$_3$	A	32.64	39.66	39.66	43.38	43.80	40.90
	P	25.49	26.81	26.81	27.97	27.87	24.35
	R	82.53	76.19	76.19	74.60	73.01	60.31
	F	38.95	39.66	39.66	40.69	40.35	34.70

Table 4.8: The performance of supervised classifier trained using the Explicit$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge (i.e., + $KB_{1_{BCA}}$) and the knowledge of causal semantics of verbs. The knowledge of causal semantics of verbs is incorporated by adding information about the linguistic definition of events (i.e., +LD) and the semantic classes of events with a high and low tendency to encode causation (i.e., +¬C$_{e_v}$ ={(R)EPORTING} OR +¬C$_{e_v}$ ={(R)EPORTING, (S)TATE} OR +¬C$_{e_v}$ ={(A)SPECTUAL}). The performance is provided in terms of the following (S)cores: (A)ccuracy, (P)recision, (R)ecall and (F)-score.

Test-set	**S**	**PDTB**$_{e_{v_i}\text{-}e_{v_j}}$	+ $KB_{1_{ICA}}$	**+LD**	+$\neg C_{e_v} = \{R\}$	+$\neg C_{e_v} = \{R, S\}$	+$\neg C_{e_v} = \{A\}$
Test-set$_1$	A	42.53	52.98	53.17	58.39	58.76	54.47
	P	23.76	29.23	29.33	31.95	31.34	29.75
	R	64.56	69.29	69.29	66.92	62.20	67.71
	F	34.74	41.12	41.21	43.25	41.68	41.34
Test-set$_2$	A	56.80	44.60	45.10	47.40	48.60	46.00
	P	31.57	28.59	28.81	29.40	29.51	27.98
	R	50.92	69.74	69.74	67.15	64.57	63.09
	F	38.98	40.55	40.77	40.89	40.50	38.77
Test-set$_3$	A	52.47	41.73	41.73	45.45	45.86	42.97
	P	28.33	27.84	27.84	29.09	29.01	25.49
	R	53.96	77.77	77.77	76.19	74.60	61.90
	F	37.15	41.00	41.00	42.10	41.77	36.11

Table 4.9: The performance of supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge (i.e., + $KB_{1_{ICA}}$) and the knowledge of causal semantics of verbs. The knowledge of causal semantics of verbs is incorporated by adding information about the linguistic definition of events (i.e., +LD) and the semantic classes of events with a high and low tendency to encode causation (i.e., +$\neg C_{e_v}$ ={(R)EPORTING} OR +$\neg C_{e_v}$ ={(R)EPORTING, (S)STATE} OR +$\neg C_{e_v}$ ={(A)SPECTUAL}). The performance is provided in terms of the following (S)cores: (A)ccuracy, (P)recision, (R)ecall and (F)-score.

model with $\neg C_{e_v} = \{R\}$ brings 5.22% (3.3%) (3.72%) improvements in accuracy over the model +LD on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively (see Table 4.7). These improvements are not at the cost of F-score as demonstrated by the 2.05% (0.48%) (1.04%) improvements in F-score over the model +LD on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively.

We also provide results by adding the knowledge of causal semantics of verbs to the model employing the supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the background knowledge $KB_{1_{ICA}}$ or $KB_{1_{BCA}}$ (Tables 4.9 and 4.10). The performance of the models given in the Tables 4.9 and 4.10 show the similar patterns of improvements as we have observed for the models given in the Tables 4.7 and 4.8. The model with $\neg C_{e_v} = \{R\}$ brings more than 2% (0.12%) (1%) improvements in accuracy over the model +LD on Test-set$_1$ (Test-set$_2$) (Test-set$_3$), respectively (see Tables 4.9 and 4.10).

For this research, we performed some experiments to identify the sources of knowledge which produce best performance for Test-set$_1$ and Test-set$_2$. For this purpose, we combined both Test-set$_1$ and Test-set$_2$ and executed our model by choosing one option from each of the following sources of knowledge:

- **Training data:** Explicit$_{e_{v_i}\text{-}e_{v_j}}$ or PDTB$_{e_{v_i}\text{-}e_{v_j}}$
- **Background knowledge:** $KB_{1_{ICA}}$ or $KB_{1_{BCA}}$ or $KB_{2_{ICA}}$ or $KB_{2_{BCA}}$
- **Linguistic definition of events:** LD
- **Semantic classes of events:** $\neg C_{e_v} = \{R\}$ or $\neg C_{e_v} = \{R, S\}$ or $\neg C_{e_v} = \{A\}$

Test-set	S	**PDTB**$_{e_{v_i}\text{-}e_{v_j}}$	+ $KB_{1_{BCA}}$	**+LD**	+$\neg C_{e_v} = \{R\}$	+$\neg C_{e_v} = \{R, S\}$	+$\neg C_{e_v} = \{A\}$
Test-set$_1$	A	42.53	52.79	52.79	58.20	58.39	53.73
	P	23.76	29.41	29.41	32.10	31.53	29.90
	R	64.56	70.86	70.86	68.50	64.56	66.92
	F	34.74	41.57	41.57	43.71	42.37	40.60
Test-set$_2$	A	56.80	44.70	45.20	47.60	48.80	46.10
	P	31.57	28.63	28.85	29.49	29.61	28.03
	R	50.92	69.74	69.74	67.15	64.57	63.09
	F	38.98	40.60	40.82	40.99	40.60	38.81
Test-set$_3$	A	52.47	40.49	40.49	44.21	44.62	41.73
	P	28.33	27.62	27.62	28.82	28.74	25.31
	R	53.96	79.36	79.36	77.77	76.19	63.49
	F	37.15	40.98	40.98	42.06	41.73	36.19

Table 4.10: The performance of supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus and the models with the addition of background knowledge (i.e., + $KB_{1_{BCA}}$) and the knowledge of causal semantics of verbs. The knowledge of causal semantics of verbs is incorporated by adding information about the linguistic definition of events (i.e., +LD) and the semantic classes of events with a high and low tendency to encode causation (i.e., +$\neg C_{e_v}$ ={(R)EPORTING} OR +$\neg C_{e_v}$ ={(R)EPORTING, (S)STATE} OR +$\neg C_{e_v}$ ={(A)SPECTUAL}). The performance is provided in terms of the following (S)cores: (A)ccuracy, (P)recision, (R)ecall and (F)-score.

Notice that using the above mentioned four sources of knowledge we executed our model for 32 settings and obtained best F-score (i.e., 42.04%) using the following options: (1) PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus, (2) $KB_{1_{BCA}}$, (3) LD and (4) $\neg C_{e_v} = \{R\}$. These four options represent the model PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$+LD+$\neg C_{e_v} = \{R\}$. The second best model is PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{ICA}}$+LD+$\neg C_{e_v} = \{R\}$ which produces 41.34% F-score.

Figure 4.3 presents the interpolated precision-recall curves of the best model PDTB$_{e_{v_i}\text{-}e_{v_j}}$ + $KB_{1_{BCA}}$ + LD + $\neg C_{e_v} = \{R\}$, the model PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$ and the supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus. Note that the interpolated precision-recall curve of the model PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$ is same as the curve of the model +BCA of Figure 4.1. In the previous section we have explained method for obtaining the interpolated precision-recall curve of the models PDTB$_{e_{v_i}\text{-}e_{v_j}}$ and PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$. In order to acquire precision-recall curve for the model PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$+LD+$\neg C_{e_v} = \{R\}$, we execute the ILP program of this model and observe the event pairs p_i that are forcefully labeled with C or $\neg C$ as a result of constraints employing LD and $\neg C_{e_v} = \{R\}$. For the pairs p_i we acquire the labels produced by ILP program and for the rest of the pairs we obtain predictions using the model PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$.

Figure 4.3 reveals that the addition of more knowledge (i.e., LD and $\neg C_{e_v} = \{R\}$) helps achieve more precision over PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB_{1_{BCA}}$ on almost all recall levels except the recall values 0.7, 0.9 and 1.0 where the precision values of both models match each other. In the next section, we present detailed error analysis of the best performing model with some discussion on how this model can be further improved to achieve more progress on the current task.

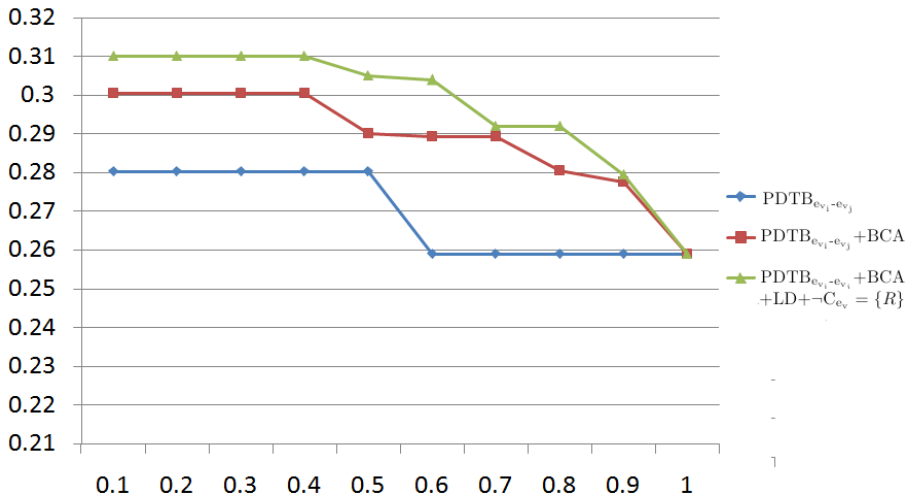

Figure 4.3: The interpolated precision-recall curves for the supervised classifier trained using the PDTB$_{e_{v_i}\text{-}e_{v_j}}$ corpus, the models PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB1_{BCA}$ and PDTB$_{e_{v_i}\text{-}e_{v_j}}$+$KB1_{BCA}$+LD+$\neg C_{e_v} = \{R\}$. The threshold γ increases in the increments of 0.1 from left to right and produces different precision and recall values for each of the above stated models.

4.2.5 Error Analysis and Discussion

In order to perform error analysis, we choose the best performing model PDTB$_{e_{v_i}\text{-}e_{v_j}}$ + $KB1_{BCA}$ + LD + $\neg C_{e_v} = \{R\}$ which produces 30.41% precision, 68.09% recall and 42.04% F-score. We randomly selected 100 false positives and 50 false negatives from the predictions of this model. In the rest of this section, we provide frequent types of errors made by the above mentioned model which results in false positive and false negative predictions.

False Positives

After the analysis of false positives, we observed the following frequent types of errors:

- About 47% instances of the false positives are encoding expansion, continuation or list relations and are mistakenly identified as causal instances by the best performing model. Consider the following two examples of such type of instances:

  1. But sales in the oil-patch state of Texas **surged** 12.9% and sales in South Carolina **jumped** 10.6% in the period, the New York trade group said.
  2. Taiwan's USI Far East Corp., a petrochemical company , initialed the agreement with an unidentified Japanese contractor to build a naphtha cracker , according to Alson Lee , who heads the Philippine company set up to **build** and **operate** the complex.

Notice that in example (1) two events "sales surged 12.9%" and "sales jumped 10.6%" are synonymous events i.e., both events are representing similar state of affairs. The background knowledge (i.e., $KB_{1_{BCA}}$) currently predicts that the pair surge-jump has a high tendency to encode a cause relation than the non-cause one. In order to handle such mistakes we need to incorporate a mechanism on top of our model to identify synonymous events. For example, our model should know that surge and jump are nearly synonyms. In addition to this, there is more evidence available in example (1) which can help filter this example from the false positives. For example, notice that both verbs surge and jump appear in the similar context i.e., {sales **verb** PERCENTAGE}. This provides additional evidence to assign the label $\neg C$ to example (1). Example (1) is an easy case of verb-verb pair encoding a non-cause relation. Consider example (2) where the pair "build-operate" does not consist of synonymous verbs and can encode both cause or non-cause relations depending on the context. A model needs to process this example more deeply to identify a non-cause relation in two events. Note that there are purpose (or cause) relations in the pairs "set up-build" and "set up-operate" as identified by the preposition "to". The events e_{build} and $e_{operate}$ conjoined by the marker "and" are effects of the same event e_{set-up} and are encoding a non-cause relation. A model can simply predict the label $\neg C$ for the events $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ following a structure $\mathrm{e_{v_i}} \leftarrow \mathrm{e_{v_k}} \rightarrow \mathrm{e_{v_j}}$ where $\mathrm{e_{v_k}}$ causes both $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$. This type of reasoning is not always correct because we know that the events $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ can influence each other even if they are effects of the same event $\mathrm{e_{v_k}}$. In this situation we can learn and incorporate the general tendencies of the events. For example, two events $\mathrm{e_{v_i}}$ and $\mathrm{e_{v_j}}$ representing effects of the same event $\mathrm{e_{v_k}}$ may always have a high tendency to encode non-causality if these events are conjoined by the marker "and". From the above examples, it is evident that the model for identifying causality needs to deeply process natural language instances to distinguish expansion, continuation or list relations from the causal relations.

- In about 28% instances of the false positives two events are not even directly relevant and are mistakenly identified as causal instances. Consider the following two examples of such type of instances:

  3. US Airways is **using** a similar system on its Airbus aircraft to eliminate all non-precision approaches at every runway of every airport where the airline's Airbus aircraft **flies**.

  4. Traders expected a rise of only 50 billion to 85 billion cubic feet because cold weather in the U.S. was **thought** to have boosted demand for heating fuels more. The increase **put** the nation 's gas storage within 8 percent of where it was a year ago at this time , when inventories were considered sufficient.

Note that in example (3) the events e_{use} and $_{fly}$ are not encoding causality. In fact, these events are not directly relevant to each other in the given context. In this example the event e_{fly} is explaining a fact about some airports. Therefore, in order to identify causality our model first needs to identify if two events are directly relevant or not in the current context. If the two events are not directly relevant then there is a high tendency of encoding non-causality. Similarly in example (4) the event e_{think} is directly relevant to the event e_{boost} and it does not have a direct relation with the event e_{put}. Instead the event e_{boost} encodes a cause relation with the event e_{put}. Our current model does not have a mechanism to identify if two events are directly relevant or not and this leads to lots of false positives in the predictions.

- For about 16% instances of the false positives our model fails to identify REPORTING events. Consider the following example where a REPORTING event is just describing another event instead of encoding causation with it:

  5. And on the West Coast , evidence today **shows** that a monster quake in 1700 ruptured 500 miles of the ground from Puget Sound south and **sent** a huge tsunami that flooded coastal Japan .

  In example (5) e_{show} (a REPORTING event) is just describing the event e_{send} instead of encoding a cause relation with it. Our model fails to identify e_{show} as a REPORTING event because in the TimeBank corpus of events there is a total of 9 instances of the verb "show" and only 2 instances are of REPORTING class. Therefore, in the future we need more training data for the semantic classes of events to avoid mistakes in identifying these semantic classes. The TimeBank corpus contains only 7924 instances of verbal events and thus it can result in wrong predictions of the semantic classes as demonstrated by the example (5). In this situation, we can utilize other resources with the semantic classes of verbs. For example, WordNet provides 15 semantic classes of verbs and we can acquire a large number of instances of these classes from the WordNet's glosses/examples. Some of the above mentioned 15 classes are VERB. Body, VERB. Communication, VERB.Cognition, etc. Here, the VERB.Communication closely maps to the REPORTING events. For the current task, we can acquire the instances of verbs from the WordNet and organize the 15 semantic classes of verbs into the categories C_{e_v} and $\neg C_{e_v}$. The fine-grained 15 semantic classes of verbs and their large number of instances from the WordNet may help achieve better results for the current task.

- In the rest of the 9% instances of the false positives, two events are either encoding comparison or temporal only relations and are mistakenly identified as causal relations. As discussed above, the model for identifying causality needs to deeply process natural language instances to distinguish non-cause

relations (i.e., expansion, continuation, list, comparison and temporal only) from the causal relations.

False Negatives

After the analysis of false negatives, we observed the following frequent case of errors:

- In about 88% instances of the false negatives, the verb-verb pairs apparently encode a non-cause relation but some facts about these pairs allow them to encode causality. Consider the following example of a such pair:

  6. **Leave** it alone and unlocked and you might return to find a stranger **sitting** in the driver's seat just getting the feel of the car.

  Apparently the pair "leave-sit" seems to be a non-causal pair and in our test set there is a total of 5 instances of this pair with the label $\neg C$. Example (6) is the only causal instance of this pair in our test set. The background knowledge $KB_{1_{BCA}}$ predicts that this pair has a high tendency to encode non-causality. Using the background knowledge $KB_{1_{BCA}}$, our model predicts correct label (i.e., $\neg C$) on 83.33% instances of the pair "leave-sit". In order to predict the label C for example (6) our model needs to have more specific knowledge about the pair "leave-sit". For example, our model should know that "if you **leave** some space **X** then somebody can **sit** on the space **X**". Also our model must have information that "it" and "the driver's seat" are more or less referring to the same entities in example (6). After acquiring the above mentioned information we need to validate if the above rule (i.e., **Leave** some space **X** $\rightarrow$ somebody **sit** on **X**) is satisfied in the context of example (6) or not.

- In about 12% instances of the false negatives, the verb-verb pairs apparently seem to encode causality with a high tendency and the background knowledge $KB_{1_{BCA}}$ has failed to identify the causal connections. For example, the pair "want-push" seems to encode a causal association but $KB_{1_{BCA}}$ considers this pair as a non-causal pair. There is a total 2 cause and 4 non-cause instances of this pair in the test set. So our model does not reduce recall a lot by considering the pair "want-push" as non-causal pair. But in future we need to incorporate more sources of knowledge to determine the accurate tendency of each pair to encode causation.

4.3 Conclusion

The model introduced in this chapter integrates the knowledge of context with the background knowledge and the knowledge of causal semantics of verbs for identifying causality in verb-verb pairs. The above types

of knowledge are identified by deeply analyzing the semantics of verbs. Using the learning and inference framework of ILP, our model is capable of providing optimal predictions on verb-verb pairs in the presence of rich sources of knowledge. The detailed empirical assessment of our model supports the argument that the background knowledge and the knowledge of causal semantics of verbs are critically important to achieve progress on the current task. The assessment of the knowledge of context has brought interesting trends of performance with respect to the nature of training corpus employed in the supervised learning. The supervised classifier trained via a massive training corpus (i.e., $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$) yields a very high recall as compared with the classifier trained via a small manually annotated corpus (i.e., $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$). On the other hand, the training of supervised classifier with the $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$ corpus brings a high accuracy but at the cost of recall. These trends highlight that the performance of supervised classifier is sensitive to the size of training corpus and yields trade off between accuracy and recall. However, our model tries to find balance between accuracy and recall by incorporating additional sources of knowledge. In the current work, we supply the background knowledge to our model in terms of causal associations of verb-verb pairs. The results provided in the section 4.2.3 have revealed that the advanced metrics ICA and BCA identify the causal associations in a better way as compared with the metrics ECA and CEA. Also we observed that ICA and BCA provide the better sources of background knowledge which allow our model to produce better precision values than the supervised classifier on almost all recall values from 10% to 90%. The use of knowledge of causal semantics of verbs has also produced encouraging results for the current task. Though, the performance of our best performing model is quite better than the baseline of supervised classifier, there are still several research problems that need to be tackled to bring more improvements in results. In section 4.2.5 we have identified some of these research problems while providing detailed error analysis of our best performing model.

Chapter 5

Knowledge Acquisition for Verb-Noun Pairs

The causal relation in a verb-noun pair is characterized by the semantics of its participants –i.e, verb and noun. In this chapter, we exploit the semantics of nouns, verbs and verb frames in depth to derive the knowledge important for identifying causality. In particular, for the current task we identify causality in verb-noun_phrase (denoted by v-np) pairs. Consider the following examples of Cause (C) and Non-Cause ($\neg C$) relations of v-np pairs:

1. At least 1,833 people **died** in the **hurricane**.

2. **The explosion occurred** in the city's main business area.

Here in example (1) the noun_phrase "the hurricane" represents a cause event and the verb "die" represents the effect of "hurricane". However in example (2) the noun_phrase "The explosion" does not encode a cause relation with the verb "occurred". For an instance of v-np pair, the verb (v) can represent either a cause event or an effect event and the same is the case with noun_phrase (np). In addition to this, the v and np can appear anywhere in the sentence and in any order. We consider only relations between the main verbs and the noun_phrases for the current task.

Following the task of identifying causality in verb-verb pairs, we consider the broad notion of causality for the current task. According to this notion, the contingency semantic relations i.e., cause, reason, explanation, purpose, result, etc. are considered as causal relations and any other type of relation or no relation falls into the class of non-causal relations.

Figure 5.1 shows the structure of our model for identifying causality in verb-noun_phrase pairs. As shown in Figure 5.1, our model takes as input a set VNP of instances of v-np pairs and generates the labels C or $\neg C$ on all the instances of this set. In our model, the component "Identification of Causality via Linguistic Features" is a supervised classifier which identifies causality by exploiting linguistic features. These features are extracted from the contexts of instances of v-np pairs. This component assigns the labels C or $\neg C$ on the instances of v-np pairs and provides the probabilities of these assignments. We use the term "knowledge of context" for these probabilities. To the best of our knowledge, there exists no

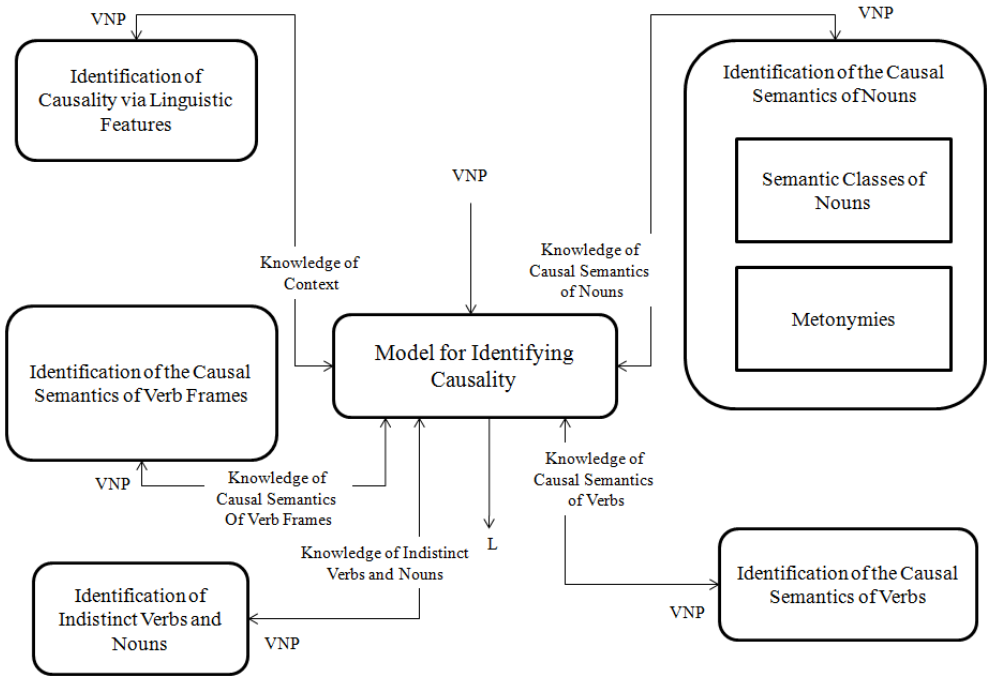

Figure 5.1: The model for identifying causality for a set VNP of instances of v-np pairs – i.e., VNP = {v-np | v is the main verb and np is the noun_phrase}. The output of this model is the set K of instances of v-np pairs with the assignments of labels C or $\neg C$ – i.e., L= $\{(\text{v-np}, l) \mid$ v is the main verb and np is the noun_phrase and $l \in \{C, \neg C\}\}$.

previously proposed supervised classifier for identifying causality in verb-noun_phrase pairs. We introduce a supervised classifier for the current task and this classifier serves as the baseline model for our approach. After the acquisition of knowledge of context, the components "Identification of the Causal Semantics of Nouns", "Identification of the Causal Semantics of Verbs" and "Identification of the Causal Semantics of Verb Frames" identify the knowledge of causal semantics of nouns, verbs and verb frames. We also introduce a component "Identification of Indistinct Verbs and Nouns" for extracting the knowledge of indistinct verbs and nouns –i.e., the knowledge of verbs and nouns which do not represent distinct state of affairs. In our model we integrate all of the above stated types of knowledge via Integer Linear Programming (ILP) framework for NLP [Roth and Yih 2004]. An example of a causality detection model implemented using this framework is provided in chapter 4.

5.1 Identification of Causality via Linguistic Features

In this section, we introduce the component of our model for identifying causality via linguistic features. This component is composed of a supervised classifier which extracts linguistic features from a training corpus of v-np pairs to learn and identify causality.

Label	Frame Elements
C	Cause, Purpose, Reason, Explanation, Required situation, Purpose of event, Negative consequences, Resulting action, Internal cause, Result, External cause, Effect, Cause of shine, Purpose of goods, Response action, Enabled situation, Grinding cause, Trigger
$\neg C$	Place, Speed, Driver, Attribute, Time, Path, Manner, Duration, Means, Attribute, Activity, Group, Protagonist, Difference, Process, Content, Executioner, Amount of progress, Treatment, Sender, Holding location, Food, Produced food

Table 5.1: Some examples of the assignments of frame elements of FrameNet to the labels C and $\neg C$.

5.1.1 Acquisition of Training Corpus

In order to set up a supervised classifier for the current task, we acquire the training corpus of v-np pairs by leveraging the annotations for all verbs in the FrameNet corpus [Baker et al. 1998]. For example, consider the following FrameNet's annotation for the verb "dying":

3. A campaign has started to try to cut the rising number of children **dying** [$_{Cause}$ from **solvent abuse**].

In this annotation the labeled element "from solvent abuse" is an instance of the frame element "Cause". We remove the preposition "from" from the above annotation to acquire an instance of v-np pair. In our model, we collect all FrameNet's annotations for verbs such that the labeled elements do not contain any verb and must contain at least one noun to represent a relation at the v-np level. If a labeled element contains a verb then this may not encode a relation at the v-np level. For example, consider the following annotation for the verb "died" where the labeled element contains a verb "fell" in it and the causal relation is encoded between the verbs "died" and "fell".

4. A fitness fanatic **died** [$_{Cause}$ when 26 stone of weights **fell** on him as he exercised].

The annotations acquired above are labeled with the 729 distinct frame elements of the FrameNet. We manually assigned the labels C and $\neg C$ to all of these frame elements to generate the training corpus for the current task. Some examples of the frame elements to which we assigned the labels C and $\neg C$ are shown in Table 5.1. These manual assignments are then used to label the FrameNet's annotations with C and $\neg C$. For example, using the assignments given in Table 5.1, we apply the label C to example (3) above with the pair "died-solvent abuse" and label $\neg C$ to example (5) below with the pair "occurred-the Demilitarized Zone".

5. The most serious border incident for several years was reported to have **occurred** on May 22 [$_{Place}$ in **the Demilitarized Zone**].

We assigned the label C ($\neg C$) to 27 (658) frame elements acquired through the FrameNet's annotations. For the rest of the 44 frame elements, we were not certain about the assignments of labels C and $\neg C$ and thus we do not include the FrameNet's annotations with these frame elements in our training corpus. We refer the reader to Appendix B for the complete list of frame elements with the assignments of labels C and $\neg C$ and the 44 frame elements with no assignments. Using the above method, We have acquired a total of $4,141$ C and $77,119$ $\neg C$ instances of v-np pairs for training. In order to avoid over-fitting towards the label $\neg C$, we employ an equal number of C and $\neg C$ instances for training by randomly selecting $4,141$ instances out of $77,119$ total $\neg C$ instances. We use the notation $\text{FNET}_{\text{v-np}}$ to refer to this training corpus. We build a supervised classifier using the $\text{FNET}_{\text{v-np}}$ training corpus and the linguistic features given in the next section.

5.1.2 Linguistic Features

In this section, we propose the following linguistic features to build a supervised classifier for the current task. We use example (3) to elaborate the following list of features (see Table 5.2).

- **Verb:** verb, lemma and part-of-speech tag of the verb.
- **Noun_Phrase:** noun_phrase, lemmas of all words of the noun_phrase, head noun of the noun_phrase and its lemma and part-of-speech tag of the head noun of the noun_phrase. In order to obtain the head noun of a noun_phrase, we traverse from the last word of the noun_phrase and pick the very first noun as the head noun.
- **Context Words:** Lemmas of all words between the verb and the noun_phrase.
- **Semantic Features:** We adopted this feature from Girju (2003) to capture the semantics of nouns. The 9 noun hierarchies of WordNet i.e., entity, psychological feature, abstraction, state, event, act, group, possession, phenomenon are used as this feature. Each of these hierarchies is set to 1 if any sense of the head noun of the noun_phrase lies in that hierarchy, otherwise set to 0.
- **Structural Features:** This feature is applied by considering both subject (i.e., sub_in_np) and object (i.e., obj_in_np) of the verb. For example, for a v-np pair, the variable sub_in_np is set to 1 if the subject_v $\in$ np, set to 0 if the subject_v $\notin$ np and set to -1 if the subject_v is not available in an instance. The subject and object of a verb are its core arguments and may sometime be part of the event represented by their verb. Therefore, these arguments may have high tendency to encode non-causation with their verb. For example, in the example (2) above, "explosion" is the subject of the verb "occurred" and is part of the event represented by the verb "occur".

Feature Type	Examples
Verb	dying, die, VBG
Noun_Phrase	solvent, abuse, solvent, abuse, abuse, abuse, NN
Context Words	from
Semantic Features	entity=1, psychological feature=1, abstraction=1, state=0, event=1, act=1, group=0, possession=0, phenomenon=0
Structural Features	sub_in_np=-1, obj_in_np=-1
Pairs	die-abuse
Example: A campaign has started to try to cut the rising number of children **dying** from **solvent abuse**	

Table 5.2: The instances of linguistics features employed by a supervised classifier for identifying causality in verb-noun_phrase pairs.

- **Pairs:** The following pairs "verb-head noun of the noun_phrase", "subject$_{verb}$-head noun of the noun_phrase" and "object$_{verb}$-head noun of the noun_phrase" are used to capture relations. The above pairs are taken with the lemmas of words.

We obtain the probabilities of assignments of labels C and $\neg C$ using both Naive Bayes (NB) and Maximum Entropy (MaxEnt) classification algorithms.

5.2 Identification of the Causal Semantics of Nouns

In order to identify causality in verb-noun_phrase pairs, our model needs to have knowledge of causal semantics of nouns. We identify this type of knowledge in terms of the semantic classes of nouns with a high and low tendency to encode causation (see section 5.2.1) and the metonymies associated with nouns (see section 5.2.2). For example, a named entity such as ORGANIZATION or LOCATION has the least tendency to encode causation unless it is involved in a metonymy. Consider the following examples:

6. Sandy **hit Cuba** as a Category 3 hurricane.

7. **The United States** has **killed** Osama bin Laden and has custody of his body.

8. Sex, drugs, and **Vietnam** have **haunted** Bill Clinton's campaign.

Here, in example (6) the pair "hit-Cuba" is a non-causal pair and in examples (7) and (8) the "kill-the United States" and "haunted-Vietnam" pairs are encoding causality where the metonymic readings are associated with the noun_phrases "the United States" and "Vietnam". In example (7), "the United States" is referring to an event "raid in Abbottabad on May 2, 2011 by the United States" rather than merely referring to a country. Similarly, in example (8) "Vietnam" refers to an event of "war in Vietnam" rather than referring to a country.

Label	Frame Elements
C_{np}	Characterization, Event, Goal, Purpose, Cause, Internal Cause, External cause, Result, Means, Reason, Phenomena, Coordinated Event, Action, Activity, Circumstances, Desired Goal, Explanation, Issue, Stimulus
$\neg C_{np}$	Artist, Performer, Duration, Time, Place, Distributor, Area, Path, Direction, Sub-region Frequency, Body part, Degree, Angle, Fixed location, Path shape, Addressee, Interval, Experiencer

Table 5.3: Some examples of the assignments of frame elements of FrameNet to the classes C_{np} and $\neg C_{np}$.

5.2.1 Semantic Classes of Nouns

For the current task, we organize nouns into two semantic classes with a high and low tendency to encode causation. In particular, we identify two classes of noun_phrases named as C_{np} and $\neg C_{np}$ where the class C_{np} ($\neg C_{np}$) contains the noun_phrases with a high (low) tendency to encode causation, respectively. For example, the class C_{np} consists of the noun_phrases which normally represent events, conditions, states, phenomena, processes and thus have a high tendency to encode causation. Similarly, the class $\neg C_{np}$ consists of the noun_phrases that normally do not encode causation unless a metonymy is associated with them e.g., any noun_phrase representing a location (see examples (6), (7) and (8)).

We leverage the annotations of FrameNet corpus to identify the above mentioned two classes of noun_phrases. For this purpose, we consider only those annotations in which the labeled elements do not contain any verb and must contain at least one noun. These annotations roughly represent instances of noun_phrases. We manually examined the inventory of frame elements acquired from these annotations and assign these frame elements to the classes C_{np} and $\neg C_{np}$. For this purpose, we follow the definitions of these two classes stated above in this section. For example, the frame element "Place" is assigned to the class $\neg C_{np}$ because it has the least tendency to encode causation unless a metonymy is associated with it. Table 5.3 shows some examples of the assignments of frame elements to the classes C_{np} and $\neg C_{np}$. Using FrameNet's annotations, we have acquired 936 distinct frame elements which we assigned to the classes C_{np} and $\neg C_{np}$. We have assigned 355 (524) frame elements to the class C_{np} ($\neg C_{np}$), respectively. For the rest of the 57 frame elements, we were not certain about the assignments of these frame elements to two classes of noun_phrases. We refer the reader to Appendix C for the full list of frame elements with assignments to the classes C_{np} and $\neg C_{np}$ and 57 frame elements with no assignments. We use these assignments of frame elements to apply the labels of C_{np} and $\neg C_{np}$ to the annotations of FrameNet corpus. Using the above method, we have acquired 52,706 C_{np} and 94,841 $\neg C_{np}$ instances. We use the term $FNET_{np}$ to refer to the corpus of these instances. We employ the $FNET_{np}$ corpus to build a supervised classifier to predict the labels C_{np} and $\neg C_{np}$. On getting input of an instance of verb-noun_phrase pair, this classifier predicts the label C_{np} or $\neg C_{np}$ on the noun_phrase.

This information is then provided to our model to make better predictions for the current task of identifying causality. Following are two examples of classes C_{np} and $\neg C_{np}$ from the $FNET_{np}$ corpus:

9. Each year, we help thousands of people who face **tremendous obstacles** (C_{np}).

10. And there are a lot of people who face these challenges **every day of their lives** ($\neg C_{np}$).

In FrameNet, for example (9) the frame element "Issue" is assigned to the expression "tremendous obstacles" and for example (10) the frame element "Frequency" is assigned to the expression "every day of their lives". Using Table 5.3 we assign the class C_{np} and $\neg C_{np}$ to examples (9) and (10).

Using the instances of $FNET_{np}$ corpus, we build a supervised classifier for the classes C_{np} and $\neg C_{np}$. In addition to above mentioned corpus obtained from FrameNet, we also employed WordNet to extract more training instances of the classes C_{np} and $\neg C_{np}$. For this purpose, we followed the approach similar to Girju and Moldovan (2002) and adopted some senses of WordNet shown in Table 5.4. For example, considering Table 5.4, we assigned the label $\neg C_{np}$ to any noun concept whose all senses in WordNet lie in the semantic hierarchy originated by the sense {time period, period of time, period}. Note that with this approach we consider a relatively unambiguous noun concept with all its senses lying in the hierarchy {time period, period of time, period}. Following this scheme, we extracted the instances of noun concepts $\in$ WordNet from the English Gigaword corpus and assigned the labels C_{np} and $\neg C_{np}$ to these instances by using the assignments of senses from Table 5.4. Girju and Moldovan (2002) have used similar scheme to rank the noun_phrases according to their tendencies to encode causation. In comparison to them, we use the WordNet's senses to increase the size of the training corpus $FNET_{np}$. In addition to this, we build an automatic classifier using this corpus to predict the semantic classes of noun_phrases. We employ the term $FNET\text{-}WNET_{np}$ to refer to the training corpus with instances of noun_phrases acquired using FrameNet and WordNet. In the training corpus $FNET\text{-}WNET_{np}$, there are $280,212$ instances of the noun_phrases (50% belonging to each of C_{np} and $\neg C_{np}$ classes). After removing the instances of $FNET_{np}$'s corpus from the $FNET\text{-}WNET_{np}$'s corpus, we are left with 87,400 C_{np} and 45,265 $\neg C_{np}$ instances of noun_phrases. We use the term $WNET_{np}$ to refer to the training corpus with these instances. We evaluate our model by providing the information of the semantic classes of noun_phrases acquired using both training corpora of noun_phrases i.e., $WNET_{np}$ and $FNET\text{-}WNET_{np}$. Note that $WNET_{np}$'s training corpus contains instances of the relatively unambiguous noun_phrases. On the other hand, $FNET\text{-}WNET_{np}$ training corpus contains instances of both ambiguous and unambiguous noun_phrases.

In order to build the supervised classifier, we employ the following list of features:

- **Lexical Features:** All words, lemmas of all words of the noun_phrase, the head noun of the noun_phrase,

Label	WordNet Senses of Nouns
C_{np}	{act, deed, human action, human activity}, {phenomenon}, {state}, {psychological feature}, {event}, {causal agent, cause, causal agency}
$\neg C_{np}$	{time period, period of time, period}, {measure, quantity, amount}, {group, grouping}, {organization, organisation}, {time unit, unit of time}, {clock time, time}

Table 5.4: The assignments of WordNet's senses of nouns to the classes C_{np} and $\neg C_{np}$.

first two (three) (four) letters of the head noun of the noun_phrase, last two, (three) (four) letters of the head noun of the noun_phrase.

- **Syntatic Features:** part-of-speech tags of all words of the noun_phrase and the head noun of the noun_phrase.

- **Semantic Features:** Frequent sense of the head noun of the noun_phrase.

We employ both NB and MaxEnt supervised classification algorithms to predict the semantic classes of noun_phrases (i.e., the classes C_{np} and $\neg C_{np}$).

In addition to using the supervised classifier for the classes C_{np} and $\neg C_{np}$, we also apply a named entity recognizer [Finkel et al. 2005] to identify the seven types of named entities (i.e., LOCATION, PERSON, ORGANIZATION, DATE, TIME, MONEY, PERCENT). On getting an instance of verb-noun_phrase, if the noun_phrase is identified as a named entity then we assume that it belongs to the class $\neg C_{np}$ unless a metonymy is associated with it. In the next section, we introduce our method to determine the association of metonymies with noun_phrases.

5.2.2 Metonymies

A part of the task of metonymy resolution is to determine if a literal or a non-literal (figurative) sense is associated with a natural language expression [Markert and Nissim 2009]. In this section, we present our approach to identify the association of metonymic readings with noun_phrases. For example, our approach to identify the semantic classes of nouns described in section 5.2.1 assigns the class $\neg C_{np}$ to the noun_phrases "the United States" of example (7). However, this noun_phrase has a non-literal sense associated with it and this leads to a causal relation in "kill-the United States" pair. Therefore, in this work we develop an approach to determine if the metonymic reading is associated with a noun_phrase or not to support the identification of causality. We assume that a noun_phrase $\in \neg C_{np}$ with the metonymic reading associated with it can encode causation as compared with the noun_phrase $\in \neg C_{np}$ with no metonymic reading.

Previously, researchers have employed the violations of the hand-annotated selectional restrictions associated with the subject and object of verbs to identify metonymies [Markert and Nissim 2009]. For example, a country cannot kill someone but an event or action launched by a country can result in killing of someone. Thus a metonymic reading is associated with "the United States" expression of example (7). For the SemEval-2007 task of metonymy resolution, a manually annotated data set was released with metonymic and non-metonymic readings to provide training for this task. This data set contains only instances of locations and organizations with and without metonymic readings. To the best of our knowledge, there is no other data set publicly available for the task of metonymy resolution. In this work, we employ any type of noun_phrase instead of just locations and organizations. Therefore, we do not employ the SemEval-2007 data set of metonymic readings for the current task. Instead, we propose a method which identifies the association of metonymy with any type of noun_phrase using the FrameNet corpus. In our method, we leverage the verb frames and the prepositions acquired from the annotations of FrameNet corpus to identify the association of metonymies with noun_phrases.

Verb Frames

In the first part of our method, we depend on verb frames for the task of metonymy resolution. In particular, we learn the rules of language with respect to verb frames and exploit the violations of these rules to identify the association of non-literal sense with the noun_phrases.

In order to learn the rules of language with respect to verb frames, we extract all the annotations of FrameNet for the verbs in which the labeled elements do not contain any verb in it. We impose this restriction because we predict the association of metonymies with noun_phrases. We assume that a labeled element with no verb in it roughly represents a noun_phrase expression[1]. We use the assignments of frame elements to the classes $\mathrm{C_{np}}$ and $\neg\mathrm{C_{np}}$ discussed in section 5.2.1 to learn the rules of language. For the current method, we are employing $48,370\,\mathrm{C_{np}}$ and $66,428\,\neg\mathrm{C_{np}}$ annotations for the metonymy resolution task.

In FrameNet, annotations are provided using the actual semantic sense of language expressions rather than literal sense. Consider the following annotation where the label "Cause" is assigned to the expression "axe and chainsaw" rather than its literal sense –i.e., tools. Therefore, considering the FrameNet's annotations, we learn the rules of language for the metonymy resolution.

11. Too often [$_{Cause}$ axe and chainsaw] **destroy** primary rainforest .

We use the annotations of the FrameNet to populate a knowledge base of verb frames given in Table 5.5.

[1]Note that, we can also place the restriction that the labeled elements must contain a noun in it. But this restriction reduces the total number annotations to a great extent

Verb (V)	**Grammatical Relation (GR)**	**Count**$_{C_{np}}$	**Count**$_{\neg C_{np}}$
kill	nsubj	1	0
kill	dobj	0	1
FN Annotations: [$_{Stimulus}$Pissed off Angelus] just kills [$_{Experiencer}$me].			
C_{np} **and** $\neg C_{np}$ **Labeling:** [$_{C_{np}}$Pissed off Angelus] just kills [$_{\neg C_{np}}$me].			

Table 5.5: The fields of a knowledge base of verb frames. The FrameNet (FN) annotations are provided in this table along with the labels for the semantic classes of nouns (i.e., C_{np} and $\neg C_{np}$). These annotations are used to populate the fields of knowledge base i.e., Verb, Grammatical Relation, $Count_{C_{np}}$ and $Count_{\neg C_{np}}$. $Count_{C_{np}}$ ($Count_{\neg C_{np}}$) is the count of the class C_{np} ($\neg C_{np}$) associated with the verb frame of form {v, gr} where v is the verb and gr is the grammatical relation with respect to the verb v.

This knowledge base contains the fields of Verb (V), Grammatical Relation (GR), $Count_{C_{np}}$ and $Count_{\neg C_{np}}$. The grammatical relation is the dependency relation [Marneffe et al. 2006] of a labeled element with respect to a verb. $Count_{C_{np}}$ and $Count_{\neg C_{np}}$ are the counts of the verb frame of form {v, gr} with labels C_{np} and $\neg C_{np}$. In the verb frame {v, gr} v is the verb and gr is the grammatical relation with respect to the verb v. Next, we define our method to populate the above stated fields of knowledge base. For this purpose, consider the following FrameNet's annotations:

12. [$_{Stimulus}$Pissed off Angelus] just kills [$_{Experiencer}$me].

The above annotations with the labels of C_{np} and $\neg C_{np}$ are as follows:

13. [$_{C_{np}}$Pissed off Angelus] just kills [$_{\neg C_{np}}$me].

Using the examples (12) and (13), we can populate the fields of knowledge base as demonstrated in Table 5.5. In example (12), the verb "kill" has the grammatical relation (or dependency relation) "nsubj" with the word "pissed" as identified by the collapsed dependency parser [Marneffe et al. 2006]. Thus the verb frame {kill, nsubj} has the $Count_{C_{np}} = 1$ and $Count_{\neg C_{np}} = 0$. Similarly, the verb "kill" has the grammatical relation "dobj" with the word "me" and thus the verb frame {kill, dobj} has the $Count_{C_{np}} = 0$ and $Count_{\neg C_{np}} = 1$. If two or more words of a labeled element has the dependency relations with the verb then we choose the very first relation in the text order to populate the knowledge base of verb frames.

Using the current state of the knowledge base given in Table 5.5, we now introduce our method to identify the association of metonymic readings with noun_phrases. Consider the noun_phrase "The United States" of example (7) which has the grammatical relation "nsubj" with the verb "kill". Our supervised classifier predicts the label $\neg C_{np}$ for this noun_phrase. However, in the current state of knowledge base of verb frames (see Table 5.5) P({kill, nsubj}, C_{np}) > P({kill, nsubj}, $\neg C_{np}$). But, our prediction of $\neg C_{np}$ for the "The United States" violates the above probabilities (i.e., the regular tendency of the verb frame {kill, nsubj}).

Considering this violation, we predict the association of metonymy with "The United States". We identify the association of metonymies with the noun_phrases and supply this knowledge along with semantic classes of nouns to our model for identifying causality in verb-noun_phrase pairs. We noticed that for some of the {v, gr} verb frames the sum of counts of the semantic classes of noun_phrases was less than 5 –i.e., $\text{Count}_{C_{\text{np}}}$ $+\text{Count}_{\neg C_{\text{np}}} < 5$. In order to identify metonymies with confidence, we ignore these verb frames in our current approach.

Prepositions

In the second part of our method, we identify the tendencies of prepositions to encode causal relations and use the violation of these tendencies to identify metonymies. In order to learn these tendencies, we use the training corpus $\text{FNET}_{\text{v-np}}$ of v-np pairs (see section 5.1.1). We select only those instances from this corpus in which a preposition appears between a verb and a noun_phrase and no other main verb appears between them. We populate the sets T_{pr} and PR with these training instances as follows. For each training instance i, we acquire the preposition pr and its label l (i.e., $l \in \{C, \neg C\}$) and add the tuple (i, pr, l) in the set T_{pr} and add pr in the set PR. Using the set T_{pr} and PR as input, we execute Procedure 5.1 given below. This procedure outputs a set PR_c which contains the prepositions with the highest tendency to appear in causal instances or the highest tendency to encode a causal relation.

Procedure 5.1 is similar to procedure 3.1 introduced in section 3.3.2 to acquire the set of semantic classes of events with the highest tendency to encode a non-causal relation. Here, Procedure 5.1 outputs the set {for, by} where the prepositions "for" and "by" have the highest tendency to encode a causal relation. This output of Procedure 5.1 contains the preposition "for" which is normally used to encode a purpose relation and the preposition "by" normally encodes a cause relation.

Using Procedure 5.1, we identify the tendency of each preposition pr $\in$ PR to encode a causal relation. This is done by computing the $score(\text{pr}, C)$ in the step 2 of the procedure. As introduced in section 3.3.2, this score has two components $score_1(\text{pr}, C)$ and $score_2(\text{pr}, C)$. The $score_1(\text{pr}, C)$ is greater than 0 only if the preposition pr encodes the causal relations more often than the non-cause ones. The $score_2(\text{pr}, C)$ is greater than 0 only if the percentage of total causal training instances with the preposition pr is greater than the percentage of total non-causal training instances with the preposition pr. We generate a list of the prepositions in descending order w.r.t the value of $score(\text{pr}, C)$ for each pr $\in$ PR. This results in a ranked $\text{list}_{\text{pr}} = [\text{pr}_1, \text{pr}_2, \ldots \text{pr}_m]$ where $score(\text{pr}_i, C) > score(\text{pr}_{i+1}, C)$. From the list_{pr}, we remove the preposition pr_i if either the $\text{score}_1(\text{pr}_i, C) < 0$ or $\text{score}_2(\text{pr}_i, C) < 0$ because the preposition pr has a high tendency to encode a non-cause relation than the cause one. For the list_{pr}, we determine the preposition pr_i above

which the preposition has the highest tendency to encode causation –i.e., the $\text{PR}_c = \{ \text{pr}_1, \text{pr}_2, \ldots, \text{pr}_{i-1}\}$. We identify pr_i using the steps 4 to 13 of procedure 5.1. The main idea is to traverse the list_{pr} in order and predict the label $l \in \{C, \neg C\}$ for the tuples of T_{pr} based on the preposition pr. For example, if we reach pr_2 in the list_{pr}, then predicts C for all tuples of form (*,pr_1,*) and (*,pr_2,*) $\in \text{T}_{\text{pr}}$ and $\neg C$ for the rest of the tuples. On these predictions we calculate the performance in terms of F-score $\times$ accuracy. We keep on traversing the list_{pr} and stop where the performance gain is less than the performance gained in the last step.

Input: T_{pr}, PR

Output: PR_c: Set of prepositions with the highest tendency to encode a causal relation.

1. **for** *each* pr $\in$ PR **do**
   2. Calculate the tendency of pr to encode a causal relation as follows:

$$
\begin{aligned}
score(\mathrm{pr}, C) &= score_1(\mathrm{pr}, C) \times score_2(\mathrm{pr}, C) \\
score_1(\mathrm{pr}, C) &= \left(\frac{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, C))}{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, *))} - \frac{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, \neg C))}{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, *))}\right) \\
score_2(\mathrm{pr}, C) &= \left(\frac{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, C))}{count(\mathrm{T_{pr}}, (*, *, C))} - \frac{count(\mathrm{T_{pr}}, (*, \mathrm{pr}, \neg C))}{count(\mathrm{T_{pr}}, (*, *, \neg C))}\right)
\end{aligned}
$$

   where count(T_{pr},(m,n,o)) is the count of (m, n, o) tuples in the set T_{pr}. We use * to show the values for which we do not care. For example count(T_{pr},(*,*,C)) is the total count of causal instances in the set T_{pr}.

end

3. Acquire the ranked list of prepositions with respect to the scores calculated in the step 2 – i.e., $\mathrm{list_{pr}} = [\mathrm{pr}_1, \mathrm{pr}_2, \ldots, \mathrm{pr}_m]$ where $score(\mathrm{pr}_i, C) > score(\mathrm{pr}_{i+1}, C)$. From the $\mathrm{list_{pr}}$, we remove the preposition pr_i if either the $\mathrm{score}_1(\mathrm{pr}_i, C) < 0$ or $\mathrm{score}_2(\mathrm{pr}_i, C) < 0$.

4. Initialize $PR_c = \emptyset$ and $\mathrm{result_{pr_{-1}}} = \mathrm{result_{pr_0}} = 0$

5. **while** *not the end of* $\mathrm{list_{pr}}$ **do**
   6. Remove pr_i from the front of the $\mathrm{list_{pr}}$
   7. Initialize the set $\mathrm{PR}_1 = \mathrm{PR}_c + \{\mathrm{pr}_i\}$ and the set $\mathrm{PR}_2 = \{\mathrm{pr}_{i+1}, \mathrm{pr}_{i+2}, \ldots, \mathrm{pr}_m\}$.
   8. **for** *each tuple (k,* pr*, l)* $\in T_{pr}$ **do**
      9. Predict the label C if pr $\in \mathrm{PR}_1$ and predict the label $\neg C$ if pr $\in \mathrm{PR}_2$.
   end
   10. Using the predictions from the step 8, calculate the $\mathrm{result_{pr_i}}$ = F-score $\times$ accuracy.
   11. **if** $result_{\mathrm{pr}_i} - result_{\mathrm{pr}_{i-1}} < result_{\mathrm{pr}_{i-1}} - result_{\mathrm{pr}_{i-2}}$ **then**
      12. Ouput PR_c
   else
      13. Go to step 5
   end

end

Procedure 5.1 A procedure to acquire a set of prepositions with the highest tendency to encode a causal relation.

Now, after the acquisition of set PR_c from Procedure 5.1, we introduce our method to identify the association of metonymic reading using the following example:

14. All weapon sites in Iraq were **destroyed** by **the United States**.

In example (14), the noun phrase "the United States" refers to an event of "attack on Iraq" which results in encoding of causality in "destroy-the United States" pair. The supervised classifier assigns the label $\neg C_{np}$ to "the United States" but the verb "destroyed" is connected with "the United Stated" via preposition "by". According to Procedure 5.1 the preposition "by" has a high tendency to encode causation and thus we assume that the noun_phrase "the United States" may encode causation with the verb "destroyed". Therefore, there is a possibility that the noun_phrase "the United States" has the metonymic reading attached with it which leads to encoding of causality in "destroyed-the United States". Using this method we predict metonymies only for the v-np instances where a preposition appears between a verb and a noun_phrase and there appears no other verb between them.

We apply both methods of metonymy resolution on noun_phrases and assume that the metonymy is associated with a noun_phrase if at least one of the above methods (i.e., the methods using verb frames and prepositions) predicts the association of metonymy with the noun_phrase.

5.3 Identification of the Causal Semantics of Verbs

As demonstrated in section 4.2.4, the knowledge of causal semantics of verbs can contribute a lot towards the identification of causality. We also incorporate this type of knowledge for the current task. In particular, we determine the semantic classes of events with a high and low tendency to encode causation using the constructions of verb-noun_phrase pairs. For this purpose, we apply Procedure 3.1 on the $FNET_{v\text{-}np}$ corpus with $4,141$ C and $77,119$ $\neg C$ instances in it. Using the supervised classifier introduced in section 3.3.2, we identify the semantic classes of events represented by the verbs of v-np pairs of $FNET_{v\text{-}np}$ corpus. After this step, we automatically acquire the semantic classes of events with a high and low tendency to encode causation via Procedure 3.1. For the current task, we have derived the following classes C_{e_v}={OCCURRENCE, PERCEPTION, ASPECTUAL, STATE, I_ACTION} and $\neg C_{e_v}$= {REPORTING, I_STATE} using the $FNET_{v\text{-}np}$ corpus where the class C_{e_v} ($\neg C_{e_v}$) contains the semantic classes of events with a high (low) tendency to encode causation, respectively. In the next chapter, we evaluate the performance of our model by using the output of Procedure 3.1 on $FNET_{v\text{-}np}$ training corpus.

5.4 Identification of the Causal Semantics of Verb Frames

In order to identify causality in verb-noun_phrase pairs, we assume that it is important for a model to have information of the verb frames which support encoding of causality. Consider the following two examples to understand this type of knowledge:

15. **The Great Storm of October 1987** almost totally **destroyed** the eighty year old pinetum at Nymans Garden in Sussex.

16. **The explosion occurred** in the city's main business area.

Here in example (15) the verb frame {destroy, subject} encodes a cause relation i.e., the verb "destroy" has a cause relation with its subject. On the other hand, in example (16) the verb frame {occur, subject} encodes a non-cause relation. The above two examples reveal that both "destroy" and "occur" have their own semantics to encode causation with respect to their subjects. In the current section, we propose our method to identify tendencies of the verb frames of form {v, gr} to encode causation. In the verb frame {v, gr} v is the verb and gr is the grammatical relation with the verb v e.g., the frames {destroy, subject} and {occur, subject} from the examples (15) and (16).

In addition to above, each grammatical relation may have different tendency to encode causation in general. For example, how likely it is for a subject or object of any verb to encode causation with its verb? Consider the following two examples:

17. The hurricane surge protection failures **prompted a lawsuit**.

18. They **provided weather forecasts**.

Here in examples (17) and (18) the direct objects "a lawsuit" and "weather forecasts" encode non-cause relations with the verbs "prompted" and "provided", respectively. In fact the direct objects of the above two examples are part of the events represented by their verbs and thus encode non-cause relations with their verbs. In our model, we also identify tendencies of the frames of form {*, gr} to encode causation. For example, the frame {*, direct object} represents the direct object of any verb. Specifically, using the frames of form {*, direct object}, we identify how much tendency a direct object of any verb has to encode causation with its verb.

Now, we introduce our method to identify tendencies of the frames of form {v, gr} and {*, gr} to encode causation. We leverage all the annotations of FrameNet for verbs to acquire the above type of knowledge. After acquiring these annotations, we apply the labels C and $\neg C$ to the frame elements as introduced in

Verb (V)	**Grammatical Relation (GR)**	**Count$_C$**	**Count$_{\neg C}$**
destroy	nsubj	1	0
destroy	advmod	0	1
destroy	dobj	0	1
FN Annotations: [$_{Cause}$ The Great Storm of October 1987] [$_{Degree}$ almost totally] destroyed [$_{Undergoer}$ the eighty year old pinetum at Nymans Garden in Sussex].			
Labels C and $\neg C$: [$_C$ The Great Storm of October 1987] [$_{\neg C}$ almost totally] destroyed [$_{\neg C}$ the eighty year old pinetum at Nymans Garden in Sussex].			

Table 5.6: The fields of a knowledge base of verb frames with respect to the labels C and $\neg C$. The FrameNet (FN) annotations are provided in this table along with the labels C and $\neg C$. These annotations are used to populate the fields of knowledge base i.e., Verb, Grammatical Relation, Count$_C$ and Count$_{\neg C}$. Count$_C$ (Count$_{\neg C}$) is the count of verb frames (i.e., {v, gr}) of the labels C ($\neg C$), respectively.

the section 5.1.1. Using the annotations of FrameNet with labels C and $\neg C$, we build a knowledge base of verb frames given in Table 5.6. From the FrameNet corpus, we have acquired 7,156 and 114,898 instances of labels C and $\neg C$, respectively for the current purpose. Table 5.6 shows he fields of the knowledge base i.e.,Verb (V), Grammatical Relation (GR), Count$_C$ and Count$_{\neg C}$. Here the grammatical relation (gr) is the dependency relation with the verb (v). Count$_C$ (Count$_{\neg C}$) is the count of verb frames (i.e., {v, gr}) of labels C ($\neg C$), respectively. We employ the following annotations of FrameNet for the verb "destroyed" to populate the fields of knowledge base given in Table 5.6.

19. [$_{Cause}$ The Great Storm of October 1987] [$_{Degree}$ almost totally] destroyed [$_{Undergoer}$ the eighty year old pinetum at Nymans Garden in Sussex].

The above annotations with the labels C and $\neg C$ are as follows:

20. [$_C$ The Great Storm of October 1987] [$_{\neg C}$ almost totally] destroyed [$_{\neg C}$ the eighty year old pinetum at Nymans Garden in Sussex].

We acquire the collapsed dependency tree [Marneffe et al. 2006] of example (19) to populate the fields of knowledge base (see Table 5.6). Using examples (19) and (20), it is demonstrated in Table 5.6 that the word "storm" of the expression "The Great Storm of October 1987" has the grammatical (dependency) relation "nsubj" with the verb "destroy" and for the verb frame {destroy, nsubj} the count$_C$ = 1 and count $_{\neg C}$ = 0. If a labeled element does not have any dependency relation with a verb, then we do not consider that labeled element for our purpose. Similarly, if more than one word of a labeled element have the dependency relations with a verb then we just consider the very first relation. For example, in the dependency tree of example (19), the verb "destroyed" has the dependency relations dobj, prep_at and prep_in with the words "pinetum", "Garden" and "Sussex", respectively but we consider only dobj relation between "destroyed" and "the eighty year old pinetum at Nymans Garden in Sussex" to populate the knowledge base.

We populate the fields of knowledge base of verb frames using all the annotations of FrameNet. We use the notation $KB_{\{\text{V,GR}\}}$ to refer to this knowledge base of verb frames. We noticed that for some of the verb frames {v, gr} the Count_C+$\text{Count}_{\neg C} < 5$. For the current task, we ignore such verb frames with a very low count. Using $KB_{\{\text{V,GR}\}}$, we compute the following probabilities:

$$\begin{aligned} P(\{\text{v,gr}\}, C) &= \frac{\text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, C)}{\text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, C) + \text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, \neg C)} \\ P(\{\text{v,gr}\}, \neg C) &= 1.0 - P(\{\text{v,gr}\}, C) \end{aligned} \tag{5.1}$$

Here, $P(\{\text{v,gr}\}, l)$ is the probability of the verb frame $\{\text{v,gr}\}$ to encode the label $l \in \{C, \neg C\}$. The function f_{count} takes as input the knowledge base $KB_{\{\text{V,GR}\}}$, the verb frame $\{\text{v,gr}\}$, and the label l. This function outputs the value of Count_l for the verb frame $\{\text{v,gr}\}$ available in the knowledge base $KB_{\{\text{V,GR}\}}$. For example, the function f_{count} with the inputs of the knowledge base $KB_{\{\text{V,GR}\}}$ of Figure 5.6, the verb frame {destroy, nsubj} and the label C provides the output 1. If a verb frame {v, gr} $\notin KB_{\{\text{V,GR}\}}$ then $\text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, C) = \text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, \neg C) = 0$ and in this case we set the $P(\{\text{v,gr}\}, C) = P(\{\text{v,gr}\}, \neg C) = 0.0$. By setting both these probabilities to 0, we inform our model that the information about the verb frame {v, gr} is not available to us. The above probabilities are the tendencies of the verb frames of form {v, gr} to encode the causal and non-causal relations.

We also compute the probabilities of the frames of form {*, gr} to encode the causal and non-causal relations. For this purpose, we populate the knowledge base $KB_{\{\text{V,GR}\}}$ using the $\text{FNET}_{\text{v-np}}$ corpus with the equal number of labels C and $\neg C$. After populating the knowledge base, we compute the following probabilities:

$$\begin{aligned} P(\{*, \text{gr}\}, C) &= \frac{\sum_{\text{v} \in KB_{\{\text{V,GR}\}}} \text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, C)}{\sum_{\text{v} \in KB_{\{\text{V,GR}\}}} \text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, C) + \text{f}_{\text{count}}(KB_{\{\text{V,GR}\}}, \{\text{v,gr}\}, \neg C)} \\ P(\{*, \text{gr}\}, \neg C) &= 1.0 - P(\{*, \text{gr}\}, C) \end{aligned} \tag{5.2}$$

We supply the probabilities from equations 5.1 and 5.2 to our model to provide the knowledge of causal semantics of verb frames.

Let us consider the following example to understand how the knowledge of causal semantics of verb frames can help identifying causality:

21. In 1698 a flood **destroyed the buildings** in Jamestown.

In above example, "destroyed" has a dependency relation "dobj" with the noun_phrase "the buildings". From $KB_{\{\mathrm{V,GR}\}}$ derived using all the FrameNet's annotations we have the following probabilities available: $P(\{\text{destroy, dobj}\}, \neg C) = 100\%$ and $P(\{*, \text{dobj}\}, \neg C) = 71.41\%$. However, using the training corpus of v-np pairs (i.e., randomly collected equal number of C and $\neg C$ instances from $\mathrm{FNET_{v\text{-}np}}$ corpus), we have the following probabilities available: $P(\text{destroy}|C) = 100\%$ and $P(\text{building}|C) = 28.57\%$. Using these probabilities, our supervised classifier has a high tendency to label example (21) with C but the $P(\{\text{destroy, dobj}\}, \neg C)$ and $P(\{*, \text{dobj}\}, \neg C)$ can help correct the wrong prediction of supervised classifier. We use smoothed probabilities to train the supervised classifier for the current task. In chapter 6, we explain our method to incorporate the knowledge of causal semantics of verb frames using probabilities of frames {v, gr} and {*, gr}.

5.5 Identification of Indistinct Verbs and Nouns

Each causal relation is characterized by two roles i.e., cause and its effect. In example (1), the noun "hurricane" is cause and the verb "died" is its effect. However, a verb-noun_phrase pair may not encode causality when a verb and a noun_phrase represent the same state of affairs. Consider the following instance:

22. Colin Powell **presented** further evidence in his **presentation**.

Here the verb "presented" and the noun_phrase "presentation" represent same event of "presenting" and thus encode a non-cause relation with each other. In our model, we determine if for a verb-noun_phrase pair, the verb and noun_phrase represent distinct or same (indistinct) state of affairs to make predictions accordingly. For this purpose, we employ the following scheme of lexical matching:

- We use NOMLEX [Macleod et al. 1998] to transform a verb into its corresponding nominalization and use the following text segments for the lexical matching.

  $Text_{\mathrm{v}}$ = [Subject] nominlized_v [Object][2]

  $Text_{\mathrm{np}}$ = All nouns of np

  If applicable, we select the noun phrases containing subject and object of v and select only nouns from these noun phrases to put in $Text_{\mathrm{v}}$.

- We remove stopwords and duplicate words from $Text_{\mathrm{v}}$ and $Text_{\mathrm{np}}$ and take lemmas of all words. If the subject or object (or both) belongs to np then we remove these arguments from $Text_{\mathrm{v}}$. Using

[2]The arguments of subject and object of a verb are parts of the verbal event (Riaz and Girju 2010). Therefore, we use these core arguments along with a verb for the lexical matching with a noun phrase.

$Text_{\text{v}}$ we add all the words of this text segment to a set T_{v}, Similarly, using $Text_{\text{np}}$ we add all the words of this text segment to a set T_{np}.

We determine the probabilities of the verb and noun_phrase representing indistinct ($\equiv$) and distinct ($\equiv$) state of affairs as follows:

$$\begin{aligned} P(\text{v} \equiv \text{np}) &= \frac{|T_{\text{v}} \bigcap T_{\text{np}}|}{|T_{\text{v}} \bigcup T_{\text{np}}|} \\ P(\text{v} \not\equiv \text{np}) &= 1.0 - P(\text{v} \equiv \text{np}) \end{aligned} \tag{5.3}$$

We supply the above probabilities (5.3) to our model to provide the knowledge of indistinct verbs and nouns.

5.6 Summary

In this chapter, we have discussed our methods for knowledge acquisition to identify causality in verb-noun_phrase pairs. The methods introduced in this chapter derive the knowledge of context, causal semantics of nouns, verbs and verb frames and the knowledge of indistinct verbs and nouns. We aim to provide these types of knowledge to our model to achieve a better performance for the current task.

Our model begins with identifying causality by extracting linguistic features from the instances of verb-noun_phrase pairs. These features are employed in the framework of supervised learning to identify causality. In section 5.1, we have introduced a supervised classifier which predicts the labels of cause and non-cause relations on verb-noun_phrase pairs using these features. This classifier also provides probabilities of assignments of above labels and these probabilities are used to provide the knowledge of context to our model.

On top of the knowledge of context, we want to add the knowledge of causal semantics of nouns, verbs and verb frames. We define the causal semantics of nouns in terms of two classes of noun_phrases with a high and low tendency to encode causation. We aim to supply information about the semantic classes of nouns to our model along with information regarding the association of metonymies with noun_phrases. The association of metonymies with noun_phrases can result in encoding of causality. In this chapter, we have proposed methods to identify the semantic classes of nouns and the association of metonymies with the noun_phrases.

We argue that the model for identifying causality should also have knowledge of verb frames which support encoding of causality. For example, we determine how likely it is for the subject/object of a verb to

encode causality with its verb. For this purpose, we have introduced a knowledge base of verb frames and this knowledge base is used to identify tendencies of the verb frames to encode causation. Our objective is to provide this information to achieve a better performance for the current task.

The causal relations in verb-noun_phrase pairs can be encoded only if the verb and noun_phrase represent distinct state of affairs. For example, in a verb-noun_phrase causal instance the verb and noun_phrase occupy distinct roles of causation i.e., cause and its effect. We have proposed a method to determine if a verb and noun_phrase of an instance represent the distinct or indistinct state of affairs to support better predictions of the cause and non-cause relations.

Chapter 6

Identifying Causality in Verb-Noun Pairs

In our model, we intend to incorporate the knowledge of context, causal semantics of nouns, verbs, verb frames and the knowledge of indistinct verbs and nouns for identifying causality in verb-noun pairs. For the current task we identify causality in verb-noun_phrase pairs. In chapter 5, we have introduced our methods to acquire the above types of knowledge. After the process of knowledge acquisition, our objective is to incorporate and evaluate the above types of knowledge for the current task.

6.1 Model for Identifying Causality

Our model for identifying causality is developed in the framework of Integer Linear Programming (ILP) for NLP [Roth and Yih 2004]. An instance of this framework is introduced in the chapter 4 to identify causality in verb-verb pairs. Using the ILP framework, in the following sections we incrementally add each type of knowledge stated above to our model.

6.1.1 Knowledge of Context

The component of our model for identifying causality via linguistic features provides the knowledge of context. This type of knowledge is available in the form of probabilities of assignments of labels C and $\neg C$ to the verb-noun_phrase pairs. These probabilities are provided by the supervised classifier for identifying causality in the verb-noun_phrase pairs (see section 5.1). We set up the following integer linear program using the knowledge of context:

$$Y_1 = \max \sum_{\text{v-np} \in \text{VNP}} \sum_{k \in K_1} w_1(\text{v-np}, k) P(\text{v-np}, k) \tag{6.1}$$

$$\sum_{k \in K_1} w_1(\text{v-np}, k) = 1 \quad \forall \, \text{v-np} \in \text{VNP} \tag{6.2}$$

$$w_1(\text{v-np}, k) \in \{0, 1\} \quad \forall\, \text{v-np} \in \text{VNP}, \quad \forall k \in K_1 \tag{6.3}$$

Here, $K_1 = \{C, \neg C\}$, VNP is the set of all v-np pairs. $w_1(\text{v-np}, k)$ is a binary decision variable (6.3) set to 1 only if the label $k \in K_1$ is assigned to the v-np pair. The constraint 6.2 enforces that only one label out of $|K_1|$ choices can be assigned to a v-np pair. In particular, we maximize the objective function Y_1 (6.1) assigning the labels $k \in K_1$ to the v-np pairs depending on the probabilities of assignments (i.e., $P(\text{v-np}, k)$) subject to the constraints introduced above. In our experiments, we obtain these probabilities using both NB and MaxEnt classifier.

6.1.2 Knowledge of Causal Semantics of Nouns

We incorporate the knowledge of causal semantics of nouns in our model by using the predictions of the supervised classifier for classes C_{np} and $\neg\text{C}_{\text{np}}$. Here, the class C_{np} ($\neg\text{C}_{\text{np}}$) contains the noun_phrases with a high (low) tendency to encode causation, respectively. On getting the set VNP of all v-np pairs, the above classifier predicts the labels C_{np} and $\neg\text{C}_{\text{np}}$ for the noun_phrase of each v-np $\in$ VNP. In addition to the above, we identify the association of metonymies with the noun_phrases. For each v-np $\in$ VNP, if our metonymy resolver from section 5.2.2 predicts the association of metonymy then add the v-np pair to a set M. We make the following additions in the integer linear program to incorporate the information regarding the semantic classes of noun_phrases and the metonymies:

$$Y_2 = Y_1 + \sum_{\text{v-np} \in \text{VNP}} \sum_{k \in K_2} w_2(\text{f}_{\text{np}}(\text{v-np}), k) P(\text{f}_{\text{np}}(\text{v-np}), k) \tag{6.4}$$

$$w_1(\text{v-np}, \neg C) - w_2(\text{f}_{\text{np}}(\text{v-np}), \neg\text{C}_{\text{np}}) \geq 0 \quad \forall\, \text{v-np} \in \text{VNP} - M \tag{6.5}$$

$$\sum_{k \in K_2} w_2(\text{f}_{\text{np}}(\text{v-np}), k) = 1 \quad \forall\, \text{v-np} \in \text{VNP} - M \tag{6.6}$$

$$w_2(\text{f}_{\text{np}}(\text{v-np}), k) \in \{0, 1\} \quad \forall\, \text{v-np} \in \text{VNP} - M, \quad \forall k \in K_2 \tag{6.7}$$

Here $K_2 = \{\text{C}_{\text{np}}, \neg\text{C}_{\text{np}}\}$, $\text{f}_{\text{np}}(\text{v-np})$ is a function which returns np of the v-np pair. M is the set of those v-np pairs with which the metonymic readings are associated. $w_2(\text{f}_{\text{np}}(\text{v-np}), k)$ is a binary decision variable (6.7) set to 1 only if the label $k \in K_2$ is assigned to the np. Constraint 6.6 enforces that only one label out of $|K_2|$ choices can be assigned to a np. Constraint 6.5 enforces that if an np belongs to the semantic class

$\neg C_{np}$ then its corresponding v-np pair must be assigned the label $\neg C$. Here, we apply this hard constraint to our model and this filters lots of false positives. However, the association of the metonymic readings can result in false negatives as demonstrated by example (7) of chapter 5. The addition of information regarding metonymies in the form of set M helps avoiding as many false negatives as possible. Using the above integer linear program, we maximize the objective function Y_2 (6.4) subject to the constraints introduced till now. For each v-np pair, we predict the semantic class of the np using the supervised classifier for the labels $k \in K_2$ and set the probabilities – i.e., $P(f_{np}(\text{v-np}), k) = 1, P(f_{np}(\text{v-np}), \{K_2\} - \{k\}) = 0$ if the label $k \in K_2$ is assigned to the np.

6.1.3 Knowledge of Causal Semantics of Verbs

In this section, we introduce our approach to integrate the knowledge of causal semantics of verbs to our model. Using the supervised classifier introduced in section 3.3.2, we predict the class C_{e_v} or $\neg C_{e_v}$ on the verb (or verbal phrase) of an instance of v-np pair. Here the class C_{e_v} ($\neg C_{e_v}$) contains the semantic classes of verbal events with a high (low) tendency to encode causation, respectively. We add the knowledge of causal semantics of verbs as follows:

$$Y_3 = Y_2 + \sum_{\text{v-np} \in \text{VNP}} \sum_{k \in K_3} w_3(f_v(\text{v-np}), k) P(f_v(\text{v-np}), k) \tag{6.8}$$

$$w_1(\text{v-np}, \neg C) - w_3(f_v(\text{v-np}), \neg C_{e_v}) >= 0 \quad \forall \text{ v-np} \in \text{VNP} \tag{6.9}$$

$$w_3(f_v(\text{v-np}), C_{e_v}) - w_1(\text{v-np}, C) >= 0 \quad \forall \text{ v-np} \in \text{VNP} \tag{6.10}$$

$$\sum_{k \in K_3} w_3(f_v(\text{v-np}), k) = 1 \quad \forall \text{ v-np} \in \text{VNP} \tag{6.11}$$

$$w_3(f_v(\text{v-np}), k) \in \{0, 1\} \quad \forall \text{ v-np} \in \text{VNP}, \quad \forall k \in K_3 \tag{6.12}$$

Here $K_3 = \{C_{e_v}, \neg C_{e_v}\}$. $f_v(\text{v-np})$ is the function which returns verb v of the v-np pair. $w_3(f_v(\text{v-np}), k)$ is a binary decision variable (6.12) set to 1 only if the label $k \in K_3$ is assigned to the v. Constraint 6.11 implies that only one label out of $|K_3|$ choices can be assigned to a verb v. Constraint 6.9 enforces that if a verbal event represented by v belongs to the class $\neg C_{e_v}$ then its corresponding v-np pair must be assigned the label $\neg C$. Similarly, the constraint 6.10 enforces that if a v-np pair encodes causality then its verb v

represents a verbal event of class $\mathrm{C}_{\mathrm{e_v}}$ –i.e., the verb v has the tendency to encode a causal relation. We maximize the objective function Y_3 (6.8) subject to the constraints introduced till now.

6.1.4 Knowledge of Causal Semantics of Verb Frames

Using the probabilities of the frames of form {v, gr} and {*, gr} to encode the causation (equations 5.1 and 5.2), we provide the knowledge of causal semantics of verb frames. Here, in above frames v is the verb and gr is the grammatical relation with respect to the verb v e.g., {destroy, subject} is an instance of the frame {v, gr} and {*, subject} is an instance of {*, gr}. We introduced our methods in the section 5.4 to compute the probabilities $P(\{\mathrm{v}, \mathrm{gr}\}, k)$ and $P(\{*, \mathrm{gr}\}, k)$ (equation 5.1 and 5.2) where $k \in \{C, \neg C\}$. We make the following additions in the integer linear program to incorporate the knowledge of causal semantics of verb frames:

$$Y_4 = Y_3 + \sum_{\substack{\text{v-np} \in \text{VNP} \wedge \\ \mathrm{f}_{\mathrm{g}_1}(\text{v-np}) \in KB_{\{\mathrm{V,GR}\}} \wedge \\ \mathrm{f}_{\mathrm{np}}(\text{v-np}) \in \mathrm{C}_{\mathrm{np}}}} \sum_{k \in K_1} w_4(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) P(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) P(\mathrm{f}_{\mathrm{g}_2}(\text{v-np}), k) \tag{6.13}$$

$$w_4(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) \leq w_1(\text{v-np}, k) \qquad \forall k \in K_1, \forall^{\substack{\text{v-np} \in \text{VNP} \wedge \mathrm{f}_{\mathrm{g}_1}(\text{v-np}) \in KB_{\{\mathrm{V,GR}\}} \wedge \\ \mathrm{f}_{\mathrm{np}}(\text{v-np}) \in \mathrm{C}_{\mathrm{np}}}} \tag{6.14}$$

$$w_1(\text{v-np}, k) \leq w_4(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) \qquad \forall k \in K_1, \forall^{\substack{\text{v-np} \in \text{VNP} \wedge \mathrm{f}_{\mathrm{g}_1}(\text{v-np}) \in KB_{\{\mathrm{V,GR}\}} \wedge \\ \mathrm{f}_{\mathrm{np}}(\text{v-np}) \in \mathrm{C}_{\mathrm{np}}}} \tag{6.15}$$

$$\sum_{k \in K_1} w_4(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) = 1 \qquad \forall^{\substack{\text{v-np} \in \text{VNP} \wedge \mathrm{f}_{\mathrm{g}_1}(\text{v-np}) \in KB_{\{\mathrm{V,GR}\}} \\ \wedge \mathrm{f}_{\mathrm{np}}(\text{v-np}) \in \mathrm{C}_{\mathrm{np}}}} \tag{6.16}$$

$$w_4(\mathrm{f}_{\mathrm{g}_1}(\text{v-np}), k) \in \{0, 1\} \qquad \forall k \in K_1, \forall^{\substack{\text{v-np} \in \text{VNP} \wedge \mathrm{f}_{\mathrm{g}_1}(\text{v-np}) \in KB_{\{\mathrm{V,GR}\}} \wedge \\ \mathrm{f}_{\mathrm{np}}(\text{v-np}) \in \mathrm{C}_{\mathrm{np}}}} \tag{6.17}$$

Here, $KB_{\{\mathrm{V,GR}\}}$ is the knowledge base of verb frames (see section 5.4) and $\mathrm{f}_{\mathrm{g}_1}(\text{v-np})$ is the function which returns the frame of form {v, gr} where v is the verb and gr is the grammatical relation of the np with

v. Recall that grammatical relation is the dependency relation of the verb v with the np. f_{g_2}(v-np) returns the frame of form {*, gr} where gr is defined above. The functions f_{g_1}(v-np) and f_{g_2}(v-np) return NULL value if there is no grammatical relation between v and np in an instance. The above additions in ILP are only applicable for the v-np pairs with f_{g_1}(v-np) $\in KB_{\{V,GR\}}$ and np is identified as member of the class C_{np} because we have already filtered the cases of np$\in \neg C_{np}$ in the section 6.1.2. $w_4(f_{g_1}(\text{v-np}), k)$ is a binary decision variable (6.17) set to 1 only if the label $k \in K_1$ is assigned to f_{g_1}(v-np). Constraint 6.16 enforces that only one label out of $|K_1|$ choices can be assigned to f_{g_1}(v-np). We add the knowledge of causal semantics of verb frames in the form of constraints 6.14 and 6.15. These constraints enforce the predictions of the supervised classifier for the labels C and $\neg C$ (see section 6.1.1) to be consistent with the predictions based on the probabilities of the verb frames to encode the Cause (C) and Non-cause ($\neg C$) relations. In other words the value of the decision variable w_1 must match the value of the variable w_4. In the objective function 6.13 we add a small value $\alpha = 0.001$ with the $P(f_{g_1}(\text{v-np}), k)$ because we observed that in $KB_{\{V,GR\}}$ the counts for some verb frames {v, gr} are very low. In such cases, we rely on $P(f_{g_2}(\text{v-np}), k)$ –i.e., information about the tendencies of general frames of form *, gr. We maximize the objective function 6.13 subject to the constraints introduced till now.

6.1.5 Knowledge of Indistinct Verbs and Nouns

We argue that a verb-noun_phrase can not encode causality if the verb and the noun_phrase represent indistinct state of affairs. For example, if the verb v and the noun_phrase np represent the same (indistinct) events then they are part of the same event and thus encode non-causation. We supply this type of knowledge to our model by making the following additions to the integer linear program:

$$Y_5 = Y_4 + \sum_{\text{v-np}\in \text{VNP}} \sum_{k \in K_5} w_5(\text{v-np}, k) P(\text{v-np}, k) \tag{6.18}$$

$$w_1(\text{v-np}, \neg C) - w_5(\text{v-np}, \equiv) \geq 0 \quad \forall\, \text{v-np} \in \text{VNP} \tag{6.19}$$

$$\sum_{k \in K_5} w_5(\text{v-np}, k) = 1 \quad \forall\, \text{v-np} \in \text{VNP} \tag{6.20}$$

$$w_5(\text{v-np}, k) \in \{0, 1\} \quad \forall\, \text{v-np} \in \text{VNP}, \forall k \in K_5 \tag{6.21}$$

Here $K_5 = \{\equiv, \not\equiv\}$ where the label $\equiv$ ($\not\equiv$) represents indistinct (distinct) state of affairs, w_5(v-np, k) is

the decision variable (6.21) set to 1 only if the label $k \in K_5$ is assigned to a v-np pair. The constraint 6.20 enforces that only one label out of $|K_5|$ choices can be assigned to a v-np pair. The constraint 6.19 enforces that if a verb v and a noun_phrase np represent indistinct state of affairs then their corresponding v-np pair must be assigned the label $\neg C$. We maximize the objective function Y_5 (6.18) subject to the constraints introduced till now.

6.2 Empirical Study

In this empirical study our objective is to determine how much each type of knowledge extracted for the current task contribute towards identifying causality. In this section, we first introduce a test set we employ for evaluation and then assess performance of our model by incrementally adding the knowledge of context, causal semantics of nouns, verbs, verb frames and indistinct verbs and nouns.

6.2.1 Evaluation Data

To the best of our knowledge, there is no test set available with the verb-noun_phrase pairs annotated for the cause and non-cause relations. Do et al. (2011) have previously studied verb-noun pairs but by considering a small list of predefined nouns representing events. Therefore, we need to generate a test set to perform empirical study for the current task. In order to generate a test set, we collected three wiki articles on the topics of Hurricane Katrina, Iraq War and Egyptian Revolution of 2011. We applied a part-of-speech tagger, syntactic parser and a dependency parser on all sentences of these three articles (Toutanova et al., 2003; Klein and Manning 2003; Marneffe et al., 2006). We extracted all verb-noun_phrase (v-np) pairs from each sentence of these articles. Using the syntactic parser of each sentence, we extracted only low level noun_phrases i.e., a noun_phrase should not have any other noun_phrase embedded in it. For each of the above three articles, we selected first 500 instances of v-np for evaluation. We asked two human annotators to apply the labels of C or $\neg C$ to a total of 1500 v-np pairs. We provided the same annotation guidelines to our human annotators as we adopted for the task of identifying causality in verb-verb pairs. Table 6.1 shows the human inter-annotator agreement and the percentage of causal relations in our test set (named as Test-set$_{\text{v-np}}$. We have achieved 0.64 kappa value for the human inter-annotator agreement and there is 11.86% causal instances in Test-set$_{\text{v-np}}$.

Test-set	Total	Test Instances	% C	% Agreement	Kappa
Test-set$_{\text{v-np}}$	1500	1365	11.86	0.91	0.64

Table 6.1: The total number of instances (Total), the number of total instances on which two human annotators agreed and these instances are used for evaluation (Test Instances), the percentage of "Test Instances" with the label C (%C), the percentage of "Total" instances on which human annotators agreed to each other (% Agreement) and kappa value for the human inter-annotator agreement on the "Total" instances (Kappa).

6.2.2 Assessment of the Knowledge of Context

We start with assessing of our model by relying only on the knowledge of context. The knowledge of context is composed of the probabilities of assignments of labels C and $\neg C$ derived by the supervised classifier (see section 5.1). We obtain the results of the supervised classifier using both NB and MaxEnt classification algorithms (see Table 6.2). We implemented the NB algorithm using the add-α smoothing for the current task. For MaxEnt algorithm, we employed the MALLET toolkit (McCallum 2002) to acquire predictions for the current task.

Score	NB	MaxEnt
Accuracy	28.86	61.46
Precision	13.52	19.46
Recall	92.59	71.60
F-score	23.60	30.60

Table 6.2: The performance of the supervised classifier using both NB and MaxEnt classification algorithms on the Test-set$_{\text{v-np}}$.

The results given in Table 6.2 reveal that the MaxEnt classifier achieves a very high accuracy as well as F-score over the NB classifier. However, NB gives a very high recall by considering each type of linguistic feature independently of others. The supervised classifier serves as the baseline model for the current task. The accuracy of 61.46% by the MaxEnt classifier is very high as compared with the F-score of 30.60%. There is need to incorporate more knowledge into our model to boost the current F-score as well as accuracy. In addition to this, with the incorporation of more knowledge we intend to manage the trade off between the precision and recall in a better way. In the rest of this chapter, we use the term SC to refer to the supervised classifier identifying causality via linguistic features.

6.2.3 Assessment of the Knowledge of Causal Semantics of Nouns

In this section, we first supply the information about the semantic classes of nouns with a high and low tendency to encode causation and then add the information of the association of metonymies with the

Score	**SC**	**+NER**	**+SCN$_{\neg M_{WNET_{np}}}$**	**+SCN$_{\neg M_{FNET\text{-}WNET_{np}}}$**
Accuracy	28.86	48.13	62.12	71.86
Precision	13.52	17.34	21.23	26.18
Recall	92.59	89.50	80.86	75.30
F-score	23.60	29.05	33.63	38.85
Accuracy	61.46	67.61	75.53	80.73
Precision	19.46	22.22	26.88	32.02
Recall	71.60	69.13	61.72	55.55
F-score	30.60	33.63	37.45	40.63

Table 6.3: The performance of the supervised classifier (i.e., SC) and the model after the addition of information of semantic classes of nouns. The column NER represents the model in which the semantic classes of nouns are identified by merely relying on NER. The term (SCN$_{\neg M}$) is used to refer the model with information of semantic classes of nouns but the information regarding Metonymies is not yet available. The information of the semantic classes of nouns is acquired using a NER and a supervised classifier for the labels C_{np} and $\neg C_{np}$ trained via either $WNET_{np}$ or FNET-$WNET_{np}$ corpus. The first (second) row of the table presents results over the supervised classifier (SC) executed using NB (MaxEnt) classification algorithms, respectively.

noun_phrases.

For the current model, we predict the labels C_{np} and $\neg C_{np}$ for the noun_phrases using the named entity recognizer (Finkel et al., 2005) and a supervised classifier trained via either $WNET_{np}$ or FNET-$WNET_{np}$ corpus (see section 5.2.1). The training corpus of $WNET_{np}$ consists of instances of unambiguous nouns (or noun_phrases) with the labels C_{np} and $\neg C_{np}$. These unambiguous nouns (or noun_phrases) are extracted from WordNet. These nouns are unambiguous because all of their senses originate from the same semantic hierarchy. However, the training corpus FNET-$WNET_{np}$ contains instances of both ambiguous and unambiguous noun_phrases. This training corpus is the representative of the real data set with both ambiguous and unambiguous instances of noun_phrases.

Table 6.3 provides the performance of our model with the addition of information of semantic classes of nouns. We use the term "SCN$_{\neg M}$" to refer to the model with information of the semantic classes of nouns but the information regarding metonymies is not yet available to the model. With the addition of information of semantic classes of nouns, our model gains both accuracy and F-score to a great extent. The prediction of semantics classes of nouns via classifier relying on the FNET-$WNET_{np}$ training corpus provides lots of improvements in performance as compared with the models relying only on NER and the training corpus $WNET_{np}$ for this purpose.

As it is revealed in Table 6.3, the model +SCN$_{\neg M_{FNET\text{-}WNET_{np}}}$ gains 15.25% (10.03%) F-score over the supervised classifier build using NB (MaxEnt) classification algorithms, respectively. Similarly, the model +SCN$_{\neg M_{FNET\text{-}WNET_{np}}}$ gains 43% (19.27%) accuracy over the supervised classifier implemented via NB (MaxEnt) algorithms, respectively. These improvements in both accuracy and F-score are quite encouraging but

Score	**SC**	+SCN$_{\neg M}$	+SCN$_{M_1}$	+SCN$_{M_{1_{GR}}}$	+SCN$_{M_{1_{GR}}+M_2}$
Accuracy	28.86	71.86	71.35	71.42	71.64
Precision	13.52	26.18	26.29	26.34	27.54
Recall	92.59	75.30	78.39	78.39	85.18
F-score	23.60	38.85	39.37	39.44	41.62
Accuracy	61.46	80.73	80.65	80.73	81.02
Precision	19.46	32.02	32.41	32.52	34.09
Recall	71.60	55.55	58.02	58.24	64.19
F-score	30.60	40.63	41.59	41.68	44.53

Table 6.4: The performance of the supervised classifier (SC) and the model after the addition of information of semantic classes of nouns with no knowledge of metonymies (SCN$_{\neg M}$), information of semantic classes of nouns and metonymies derived via method M_1 (SCN$_{M_1}$), information of semantic classes of nouns and metonymies derived via method $M_{1_{GR}}$ (SCN$_{M_{1_{GR}}}$) and information of semantic classes of nouns and metonymies derived via methods $M_{1_{GR}}$ and M_2 (SCN$_{M_{1_{GR}}+M_2}$)

the recall drops by around 16% on both NB and MaxEnt supervised classifiers. We have observed that the model with information of semantic classes of nouns helps reducing lots of false positives and this leads to significant raise in accuracy, precision and F-Score. But our model does not currently have information regarding the association of metonymies and this leads to lots of false negatives in the predictions of the model +SCN$_{\neg M_{\text{FNET-WNET}_{\text{np}}}}$.

Next, we add information regarding metonymies to avoid as many false negatives as possible. Our objective is to recover the recall while not reducing accuracy, precision and F-score. Table 6.4 provides results after the addition of information of metonymies associated with the noun_phrase. In the section 5.2.2, we have introduced two methods for metonymy resolution. One of these methods employing the verb frames is referred with the name M_1 and the other method employing prepositions is referred with the name M_2. For the method M_1, we also evaluate the performance of our model when the metonymies are identified by depending only on the core grammatical (dependency) relations of subject, object and agent. This method is referred with the name $M_{1_{GR}}$ where GR = {csubj, csubjpass, nsubj, nsubjpass, xsubj, dobj, iobj, pobj, agent}

Table 6.4 shows that the addition of information of metonymies allows the model +SCN$_{M_{1_{GR}}+M_2}$ to achieve 2.77% (3.9%) gain in F-score over +SCN$_{\neg M}$ with NB (MaxEnt) supervised classifier, respectively. In fact the method $M_{1_{GR}}$ of metonymy resolution recovers 3.09% (2.69%) recall over +SCN$_{\neg M}$ with NB (MaxEnt) supervised classifier, respectively. Even more, the method $M_{1_{GR}}$ + M_2 recovers more than 8% recall over +SCN$_{\neg M}$ and these improvements in recall are not at the cost of precision, accuracy, F-score. In fact, the method $M_{1_{GR}}$ + M_2 allows the model +SCN$_{M_{1_{GR}}+M_2}$ to boost the precision of the model by more than 1% over +SCN$_{\neg M}$.

Score	**SC**	$+SCN_M$	$+SCN_M + \neg C_{e_v} = \{R\}$	$+SCN_M + \neg C_{e_v} = \{R, I\_S\}$
Accuracy	28.86	71.64	73.26	73.62
Precision	13.52	27.54	28.63	28.84
Recall	92.59	85.18	83.95	83.33
F-score	23.60	41.62	42.70	42.85
Accuracy	61.46	81.02	81.46	81.46
Precision	19.46	34.09	34.78	34.78
Recall	71.60	64.19	64.19	64.19
F-score	30.60	44.53	45.11	45.11

Table 6.5: The performance of the supervised classifier (SC) and the model after the addition of knowledge of causal semantics of nouns (SCN_M where $M = M_{1_{GR}} + M_2$), knowledge of causal semantics of verbs with $\neg C_{e_v} = \{R\}$ and $\neg C_{e_v} = \{R, I\_S\}$.

6.2.4 Assessment of the Knowledge of Causal Semantics of Verbs

In order to evaluate the performance of our model with the knowledge of causal semantics of verbs, we set up the model with one of the settings for the class $\neg C_{e_v}$ (see section 5.3).

- $\neg C_{e_v} = \{\text{REPORTING}\}$ denoted by $\neg C_{e_v} = \{R\}$
- $\neg C_{e_v} = \{\text{REPORTING, I\_STATE}\}$ denoted by $\neg C_{e_v} = \{R, I\_S\}$

Table 6.5 presents the performance of our model with the addition of information of semantic classes of verbs with a high and low tendency to encode causation. The results in Table 6.5 validate our assumption that there are some semantic classes of verbal events with the least tendency to encode causation. For example, with adding the knowledge of causal semantics of verbs by setting $\neg C_{e_v} = \{\text{REPORTING, I\_STATE}\}$, the model $+SCN_M + \neg C_{e_v} = \{R, I\_S\}$ gains 1.23% (0.58%) F-score over the model $+SCN_M$ with NB (MaxEnt) supervised classifier, respectively. Similarly, performance is gained with respect to accuracy. The model $+SCN_M + \neg C_{e_v} = \{R, I\_S\}$ gains 1.98% (0.44%) F-score over the model $+SCN_M$ with NB (MaxEnt) supervised classifier, respectively.

6.2.5 Assessment of the Knowledge of Causal Semantics of Verb Frames

This section provides the performance of our model with the addition of causal semantics of verb frames. We use the term VF to refer to this type of knowledge. Table 6.6 provides the performance of models when the knowledge VF is added to the models $+SCN_M$ and $+SCN_M + \neg C_{e_v} = \{R, I\_S\}$. The results in this table reveal that the knowledge of causal semantics of verb frames provides improvement in both accuracy and F-score. The model $+SCN_M$+VF brings 0.81% (0.49%) gain in F-score over the $+SCN_M$ with NB (MaxEnt) classifier, respectively. Similarly, we observe improvements in the accuracy with the addition of

Score	$+SCN_M$	$+SCN_M + \neg C_{e_v} = \{R, I\_S\}$	$+SCN_M$**+VF**	$+SCN_M + \neg C_{e_v} = \{R, I\_S\}$**+VF**
Accuracy	71.64	73.62	72.96	74.87
Precision	27.54	28.84	28.39	29.84
Recall	85.18	83.33	83.95	82.71
F-score	41.62	42.85	42.43	43.86
Accuracy	81.02	81.46	81.39	81.90
Precision	34.09	34.78	34.66	35.49
Recall	64.19	64.19	64.19	64.19
F-score	44.53	45.11	45.02	45.71

Table 6.6: The performance of the model after the addition of knowledge of causal semantics of nouns (SCN_M), causal semantics of verbs with $\neg C_{e_v} = \{R, I\_S\}$ and causal semantics of verb frames (VF).

Score	**SC**	$+SCN_M + \neg C_{e_v} = \{R, I\_S\}$**+VF**	$+SCN_M + \neg C_{e_v} = \{R, I\_S\}$**+VF+IVN**
Accuracy	28.86	74.87	75.75
Precision	13.52	29.84	30.57
Recall	92.59	82.71	82.09
F-score	23.60	43.86	44.55
Accuracy	61.46	81.90	82.63
Precision	19.46	35.49	36.65
Recall	71.60	64.19	63.58
F-score	30.60	45.71	46.50

Table 6.7: The performance of the supervised classifier (SC), the model after the addition of knowledge of causal semantics of nouns, verbs, and verb frames and the knowledge of indistinct verbs and nouns (IVN).

the information VF. The encouraging trend is that the knowledge of both causal semantics of verbs and verb frames help gaining 2.24% (1.18%) improvement in F-score over the model $+SCN_M$ using NB (MaxEnt) classifiers, respectively. The model $+SCN_M + \neg C_{e_v} = \{R, I\_S\}$+VF also gains 3.23%(0.88%) accuracy over the model $+SCN_M$ using NB (MaxEnt) classifiers, respectively. These results establish the fact that more sources of knowledge are required to achieve progress on the current task.

6.2.6 Assessment of the Knowledge of Indistinct Verbs and Nouns

Using this type of knowledge, we inform our model about the indistinct state of affairs. For example, if a verb and a noun_phrase represent the same event in an instance then they may not encode causality. We provide the performance of our model after the addition of information of indistinct verbs and nouns (denoted by IVN) (see Table 6.7). The Table 6.7 displays the overall progress our model has gained by employing the novel sources of knowledge. The model with the addition of information of indistinct state of affairs gains in terms of both accuracy and F-score.

Overall the performance achieved by our model as compared with the baseline of supervised classifier is quite encouraging. The improvement trends in both accuracy and F-score show that the interesting and

accurate sources of knowledge are critically important for identifying causality. Through the process of acquiring and integrating the right types of knowledge, our model has achieved more than 15% F-score and 20% accuracy over the supervised classifiers relying merely on linguistic features (see Table 6.7).

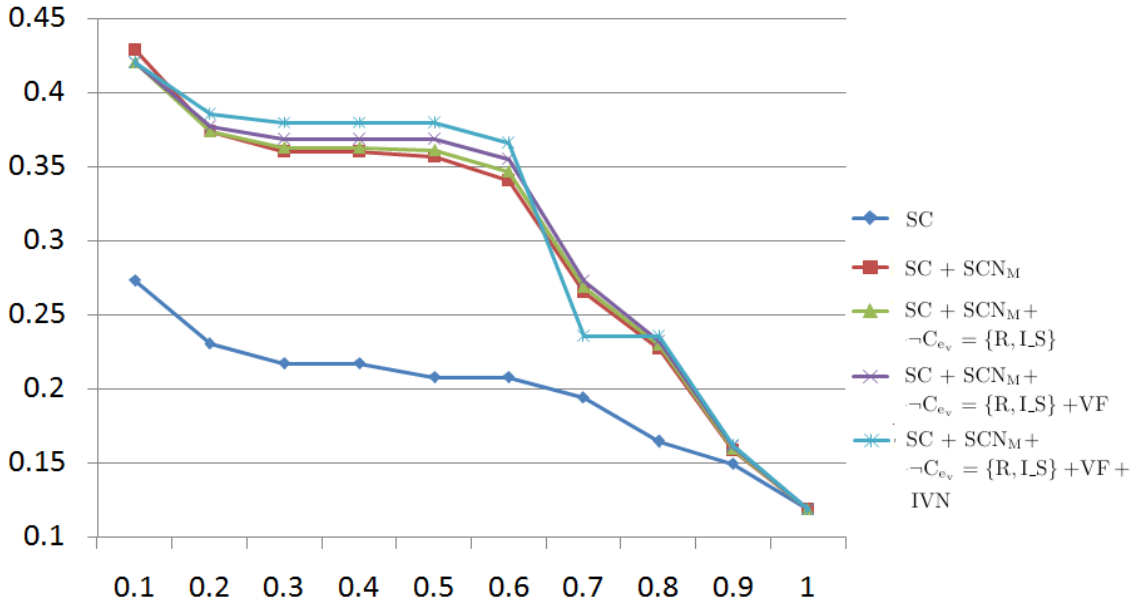

Figure 6.1: The interpolated precision-recall curves for the supervised classifier (SC) and the models SC + SCN$_M$, SC + SCN$_M$ + ¬C$_{e_v}$ = {R, I_S}, SC + SCN$_M$ + ¬C$_{e_v}$ = {R, I_S} + VF and SC + SCN$_M$ + ¬C$_{e_v}$ = {R, I_S} + VF + IVN. The threshold γ increases in the increments of 0.1 from left to right and produces different precision and recall values for each of the above stated models.

In chapter 4, we evaluated performance of our model for identifying causality using interpolated precision-recall curve. For this empirical study, we also generate interpolated precision-recall curves for the following models:

- **Supervised classifier (SC):** As described in chapter 4, a supervised classifier predicts the label C if $P(\text{v-np}, C) \geq \gamma$ and $\neg C$ otherwise. Here $P(\text{v-np}, C)$ is the probability of assignment of the label C to the v-np pair. We vary the threshold γ from 0.1 to 1.0 with the increments of 0.1 and observe the precision and recall of the supervised classifier obtained using each value of threshold (see Figure 6.1). We use supervised classifier build using MaxEnt classification algorithm which produces better results than the NB classifier.

- **SC + SCN$_M$:** In order to acquire precision-recall curve for this model, we execute the ILP program of this model and observe the pairs p_i that are forcefully labeled with $\neg C$ as a result of constraints. For the pairs p_i we acquire labels produced by the ILP program and for the rest of pairs we obtain predictions using the supervised classifier.

- **SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$:** We follow the same method as we use for the model SC + SCN_M to generate precision-recall curve.

- **SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF:** We execute the ILP program of this model and observe the pairs p_i that are forcefully labeled with C or $\neg C$ as a result of constraints. For the pairs p_i we acquire labels produced by the ILP program and for the rest of pairs we obtain predictions using the supervised classifier.

- **SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF + IVN:** We follow the same method as we use for the model SC + SCN_M+$\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$+VF to generate precision-recall curve.

Figure 6.1 reveals that the model SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF produces better results on almost all recall levels from 0.1 to 0.9. The model SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF + IVN produces best precision values on the recall values from 0.15 to 0.65 but its performance drops on the recall values from .65 to 0.8. On the recall values from 0.65 to 0.8, the model with knowledge of indistinct verbs and nouns (i.e., IVN) mistakenly assigns $\neg C$ labels to the causal v-np pairs and thus reduces precision values. The knowledge of causal semantics of nouns, verbs and verb frames allows us to produce quite better precision values than the baseline of supervised classifier. Though we have gained lots of progress over the baseline, there are still various research issues that need to be tackled to achieve more on the current task. In the next section, we identify these research issues by providing detailed error analysis for the model SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF + IVN.

6.2.7 Error Analysis and Discussion

In order to perform error analysis, we choose the model SC + SCN_M + $\neg\text{C}_{\text{e}_\text{v}} = \{\text{R}, \text{I\_S}\}$ + VF + IVN which produces 36.65% precision and 63.58% recall (46.50% F-score) (see Table 6.7). We randomly selected 100 false positives and 50 false negatives from the predictions of this model. In the rest of this section, we provide frequent types of errors made by the above mentioned model which results in false positive and false negative predictions.

False Positives

After the analysis of false positives, we observed the following frequent types of errors:

- For about 35% instances of the false positives our model makes wrong prediction due to lack of information about verb frames. Consider the following example to understand this type of error:

1. The system was **upgraded** to **tropical storm status** on the morning of August 24 and at this point, the storm was given the name Katrina.

On example (1), the supervised classifier predicts the label C mainly based on the word "storm" which appears more frequently in causal training instances. It is true that the event "storm" has a high tendency to cause various other events but its relation with the verb "upgrade" is non-causal in the current context. In FrameNet, there is no annotation available for the verb "upgrade" and thus our model is not able to correct prediction of supervised classifier based on the causal semantics of verb frames. In addition to the above, there are some verb frames of form {v, gr} $\in$ $\text{KB}_{\{\text{V,GR}\}}$ for which the annotations available are less than 5. In this situation, we need to incorporate some abstractions on the verb frames. For example, in $\text{KB}_{\{\text{V,GR}\}}$ we can keep verb frames of form {Semantic class of verbal event, gr} and identify the tendencies of these frames to encode cause or non-cause relations. Using this approach, our model will be able to take information about the causal semantics of verb frames for example (1). We can also try other types of abstractions on verb frames e.g., {sense of the verb v, gr}.

- On 32% instances of the false positives, our model makes errors on identifying the class $\neg\text{C}_{\text{np}}$ and on identifying the association of metonymy. Consider the following two examples to understand these errors:

  2. The hurricane surge protection failures in New Orleans are **considered** the worst civil engineering disaster in **U.S. history** and prompted a lawsuit against the U.S. Army Corps of Engineers (USACE), the designers and builders of the levee system as mandated by the Flood Control Act of 1965.

  3. The military junta , headed by effective head of state Mohamed Hussein Tantawi , announced on 13 February that the constitution would be suspended , both houses of parliament dissolved , and that the military would **rule** for **six months** until elections could be held .

  Our model fails to identify the correct class for the noun_phrase "U.S history" (i.e., the class $\neg\text{C}_{\text{np}}$) in example (2). This phrase represents time and it should be labeled with the class $\neg\text{C}_{\text{np}}$. Therefore, our model needs more training data of the semantic classes of nouns to filter such errors from the false positives. Consider example (3) where our model correctly assigns the label $\neg\text{C}_{\text{np}}$ to the noun_phrase "six months" but the metonymy resolver makes an error by predicting the association of metonymy with this noun_phrase due to the preposition *for*. In this work, we merely depend on verb frames and

prepositions to identify association of metonymies. In future, researchers need to study the task of metonymy resolver in more depth to handle above types of errors.

- In 21% instances of the false positives, our model predicts the label C when a verb and a noun_phrase are not even directly relevant to each other. Consider the following example:

    4. At least 1,833 people **died** in the hurricane and subsequent floods , making it the deadliest U.S. hurricane since the 1928 Okeechobee hurricane ; total property damage was estimated at $81 billion (2005 USD), nearly triple the damage brought by **Hurricane Andrew** in 1992.

  Example (4) is a complex example of the pair "died-Hurricane Andrew" where two highlighted events appear at a large distance in this example. Also notice that two events are not directly relevant to each other. The event "died" happened as a result of Hurricane Katrina in 2005 and Hurricane Andrew happened in 1992. Our current model is not able to identify if two events are happening in the current frame of time or not and thus failed to assign the label $\neg C$ to example (4).

- In the rest of 11% instances the verb-noun_phrase pair either encodes a temporal only or comparison relations. Consider the following example:

    5. It was **followed** by longer phase of **fighting**.

  In our test set there are few cases of temporal only and comparison relations in the false positives. In example (5), the verb "followed" and the noun_phrase "fighting" have a temporal only relation. Our model needs to have information that the verb "followed" provides information about time and its relation with "fighting" should be labeled with $\neg C$.

False Negatives

After the analysis of false negatives, we observed the following frequent types of errors:

- For about 77% of the false negatives, our model fails to identify causality because of lack of background knowledge. Consider the following example of such instance:

    6. In response to **mounting pressure**, Mubarak announced he had not **intended** to seek re-election in September.

  Example (6) encodes a cause relation between the event "mounting pressure" and the verb "had not intended". Our model needs to have background knowledge that "**pressure** on some person $\mathbf{X} \rightarrow \mathbf{X}$

back out". In addition to this, our model should also be able to identify if the above rule is satisfied in the context of example (6) or not. In our model for identifying causality in verb-verb pairs we provide background knowledge in terms of causal associations of verb-verb pairs. Similarly it is critical to add a component of background knowledge in terms of causal associations of verb-noun pairs to the current model.

- On 23% instances of the false positives, our model fails to correctly label noun_phrases with C_{np}. Consider the following example of such instances:

  7. **Hurricane Katrina** formed over the Bahamas on August 23, 2005 and crossed southern Florida as a moderate Category 1 hurricane, **causing** some deaths and flooding there before strengthening rapidly in the Gulf of Mexico .

  For example (7), our model fails to identify the class C_{np} for the noun_phrase "Hurricane Katrina" because in FrameNet there is no annotation available for the noun "Katrina" and word "Katrina" does not exist in WordNet. We need more training data to label the noun_phrase "Hurricane Katrina" with C_{np}. In addition to the above, the knowledge of causal semantics of verb frames cannot be utilized to label example (7) with C. From FrameNet we have $P(\{\text{cause,nsubj}\},C) > P(\{\text{cause, nsubj}\},\neg C)$. These probabilities cannot be utilized for example (7) because in this example there is no dependency relation between the noun_phrase "Hurricane Katrina" and the verb "causing". In this situation, we should take help from the unlabeled instances of the verb "cause" to identify how frequently "Hurricane Katrina" (or "Hurricane" or "Katrina") appears as subject of the verb "cause" and use such information along with above probabilities to identify causality in example (7).

6.3 Conclusion

Using multiple sources of knowledge, we have build a model for identifying causality in verb-noun_phrase pairs. This model integrates the knowledge of context, causal semantics of nouns, verbs and verb frames and the knowledge of indistinct verbs and nouns by taking advantage of the Integer Linear Programming framework. The supervised classifier which merely depends on linguistic features serves as a baseline model for our current task. In this chapter, we have shown that this supervised classifier lacks the knowledge necessary to identify causality and thus results in a very low F-score and accuracy as compared to our model employing various types of knowledge discussed above. In particular, we deeply analyze the semantics of the participants of a verb-noun_phrase pair i.e., verbs and nouns to identify and utilize the above types of

knowledge for the current task. Though we have gained lots of progress on this task over the baseline model but there is still room for improvement. In section 6.2.7, we have provided detailed error analysis of our model with discussion on how to make further progress on the current task.

Chapter 7

Conclusions

The automated recognition of causal information in text has lot of significance in NLP because the progress achieved on this problem is very important for various applications such as question answering, generation of coherent ordering of events and prediction of future or consequences of current actions [Girju 2003, Chklovski and Pantel 2004, Barzilay et al. 2002, Radinsky and Horvitz 2013]. Despite the importance of this task in NLP, previously researchers have studied this hard problem by mainly relying on linguistic features [Girju 2003, Bethard and Martin 2008, Sporleder and Lascarides 2008, Pitler et al. 2009]. The main motivation of this research was to drive and plug in more sources of knowledge on the supervised classifiers employing linguistic features to lead to a better performance on identifying causality. Now with the development of sophisticated machine learning models [Roth and Yih 2004, Ando and Zhang 2005a], it is possible to incorporate more and more useful sources of knowledge to tackle the natural language tasks with a high performance [Clarke 2008, Chan and Roth 2010, Do et al. 2011, Roth and Yih 2007, Ando and Zhang 2005b]. In this thesis, we have proposed methods to go deeper into the semantics of verbs, nouns and verb frames to acquire the following four types of knowledge (1) background knowledge, (2) causal semantics of nouns, (3) causal semantics of verbs, and (4) causal semantics of verb frames. Benefiting the ILP framework for NLP [Roth and Yih 2004], the above types of knowledge are integrated in our models to make optimal predictions for the tasks of identifying causality in verb-verb and verb-noun pairs. In the rest of this chapter, we briefly summarize the current research and discuss some future research directions.

7.1 Summary

In chapters 3 and 4 of this thesis, we have introduced models for the identification of causality in verb-verb pairs. For this task, our model identifies causality for both intra- and inter-sentential instances of e_{v_i}-e_{v_j} pairs where e_{v_i} and e_{v_j} are events represented by the verbs v_i and v_j. Also, in these instances two verbs can appear in any context i.e, unambiguous, ambiguous or implicit context. Our approach for the current task consists of two steps –i.e., knowledge acquisition and the integration of all types of knowledge. We

have proposed methods to acquire the background knowledge and the knowledge of causal semantics of verbs. These two types of knowledge are combined with the knowledge of context to identify causality using the ILP framework for NLP. We have used the term knowledge of context for the probabilities of assignments of cause and non-cause relations acquired through a supervised classifier. In order to set up a supervised classifier for the verb-verb pairs, we have proposed a method to acquire two training corpora –i.e., $\text{Explicit}_{e_{v_i}\text{-}e_{v_j}}$ and $\text{PDTB}_{e_{v_i}\text{-}e_{v_j}}$. Using the training corpus and the linguistic features, the supervised classifier learns and identifies causality. On top of the knowledge of context from the supervised classifier, we add the background knowledge and the knowledge of causal semantics of verbs. In our model, we supply the background knowledge using the causal associations in verb-verb pairs. For this purpose, we have proposed a set of metrics to identify the likelihood of causality in these pairs. These metrics identify causal associations by taking care of explicit and unambiguous, ambiguous and implicit contexts of causal associations. The background knowledge is provided to our model in two forms. These two forms involve the ranking scores of the verb-verb pairs and the organization of these pairs into three categories with respect to relation of causality. In our approach, the knowledge of causal semantics of verbs is described in terms of the linguistic definitions of events represented by verbs and the semantic classes of events with a high and low tendency to encode causation. In chapter 3, we have introduced our methods to acquire the causal semantics of verbs. The detailed empirical analysis of our model in chapter 4 has revealed that with the incorporation of the above two types of knowledge our model gains lots of progress in performance over the baseline model relying merely on the linguistic features employed in the framework of supervised learning.

In the second part of this thesis i.e., chapters 5 and 6, we have studied the task of identifying causality in verb-noun pairs. Particularly, our model identifies causality in verb-noun_phrase pairs where verb and noun_phrase can appear anywhere in a sentence. In order to identify causality for these pairs, we have introduced a model which acquires and integrate various types of knowledge necessary for this task. For the knowledge acquisition step for this task, we have proposed methods to acquire the knowledge of context, the knowledge of causal semantics of nouns, verbs, verb frames and indistinct verbs and nouns. In order to derive the knowledge of context, we have introduced a supervised classifier which takes supervision from a training corpus we extract from the FrameNet annotations and depends on the linguistic features for learning and identifying causality. This supervised classifier serves as the baseline for the current task. We combine the knowledge of context with the knowledge of causal semantics of nouns and verbs to achieve better results for the current task. The knowledge of causal semantics of nouns is comprised of the semantic classes of nouns and the information regarding metonymies. Using the FrameNet annotations and WordNet senses, we identify the semantic classes of nouns with a high and low tendency to encode causation. In addition to

this we have proposed two methods employing verb frames and prepositions to identify the association of metonymies with the noun phrases. The knowledge of causal semantics of verbs involve the semantic classes of events with a high and low tendency to encode causation as defined earlier in this section. Similarly, we have introduced methods to acquire the knowledge of causal semantics of verb frames. In particular, this type of knowledge provides information about the tendencies of verb frames to encode causation. The last type of knowledge we used for the current task is about the indistinct state of affairs represented by verbs and nouns. Using this knowledge, we inform our model that the indistinct verbs and nouns cannot encode a cause relation. The detailed empirical analysis of our model for the current task has revealed that the use of above types of knowledge along with the supervised classifier employing linguistic features allows us to perform with a very high performance over the baseline.

7.2 Future Work

In this work, we have identified some novel types of knowledge for identifying causality in verb-verb and verb-noun pairs. The empirical evaluation of our models have shown that each of these types of knowledge helps achieve a better performance over the knowledge poor baselines. Though the trends of improvements are encouraging there are still several research directions which can be addressed in future for the problem of identifying causality.

The knowledge base of causal associations of verb-verb pairs (KB_c) has proven to be a rich source of background knowledge. In the future, researchers can develop other types of knowledge bases using the semantic classes of verbs. For example, we should identify how the semantic classes of verbs behave with respect to each other. Researchers can strive to determine if the I_ACTION-OCCURRENCE pair has a high or low tendency to encode causation. This type of information may help identifying causality for the v_i-v_j $\notin KB_c$. For example, for a v_i-v_j $\notin KB_c$, we can utilize information regarding the pairs of semantic classes of verbal events e_{v_i} and e_{v_j}. Researchers can also go deeper in this direction and identify the tendencies of verbs to encode causation with respect to the semantic classes. For example, we can determine and utilize information regarding the likelihood of the pair "investigate-OCCURRENCE" to encode causation in our model for identifying causality. In future, one should also consider the utilization of background knowledge for the verb-noun pairs. For this purpose, researchers should acquire causal associations in verb-noun pairs using a large number of unlabeled instances of these pairs. However, here one needs to be careful because nouns do not always represent events. A better approach would be to find causal associations in verb-noun pairs where noun $\in C_{np}$ class.

In this thesis, we have focused only on two linguistic constructions and acquired the novel sources of knowledge relating to these constructions. Certainly, in the future we can follow the current direction of acquiring more sources of knowledge for identifying causality between discourse segments. Each discourse segment is a big chunk of text as compared with verbs and nouns and thus the model for identifying causality will require knowledge relating to all components of the discourse segment. For example, a discourse segment is composed of verbs, nouns, adjectives, etc. Thus, we need information about the causal semantics of nouns, verbs and adjectives, etc to identify causality between two discourse segments. We have acquired the causal semantics of nouns and verbs in this work. In future, the researchers can determine if there exists some classes of adjectives with a high and low tendency to encode causation. Also, it is a matter of research to identify the best way to employ the causal semantics of nouns and verbs for the discourse segments.

Appendix A

Human Annotations

This appendix explains the annotation guidelines we provide to our human annotators to label instances of verb-verb and verb-noun pairs with Cause (C) and Non-Cause ($\neg C$) labels. We provide an objective notion of causality [Beamer and Girju 2009] to two human annotators for the assignments of above two labels to verb-verb and verb-noun pairs. This notion is based on the Manipulation Theory of Causality [Woodward 2008] given below:

> Manipulation theory of causality determines truth of the following two conditions to determine if a cause-effect relation is encoded between the two events **a** and **b** or not: (1) event **a** must temporally precede or overlap event **b** in time and (2) while keeping as many state of current affairs constant as possible, modifying event **a** must entail predictably modifying event **b** [Woodward 2008, Beamer and Girju 2009].

Based on the manipulation theory of causality, we gave the following guidelines to two human annotators for labeling the instances of verb-verb and verb-noun pairs with C and $\neg C$.

> Assign the label C to an instance of **a-b** pair only if the two truth conditions of the manipulation theory of causality are satisfied and no additive relation (list, continuation, opposition, exception, enumeration, temporal, and concession) can be recognized from the discourse markers or other elements of the context of instance. Otherwise, assign the label $\neg C$. Also, assign the label $\neg C$ if the annotator assumes that **a** and **b** are not even relevant to each other or both are representing same state of affairs in the current context.

A.1 Verb-Verb Pairs

In this section, we provide details of annotations for the following examples of verb-verb pairs:

1. Deputies spotted the truck parked at the home of the suspect's father and **called** for assistance. The Border Patrol agents and others **responded**. (C)

The label C is assigned to the pair "called-responded" because this pair satisfies the truth conditions of manipulation theory of causality i.e., (1) the event e_{call} temporally precedes the event $e_{respond}$ and (2) while keeping as many state of current affairs constant if we can modify/control the event e_{call} then we can predictably modify/control the event $e_{respond}$.

2. Traders expected a rise of only 50 billion to 85 billion cubic feet because cold weather in the U.S. was **thought** to have boosted demand for heating fuels more. The increase **put** the nation's gas storage within 8 percent of where it was a year ago at this time , when inventories were considered sufficient. ($\neg C$)

   The label $\neg C$ is assigned to the pair "thought-put" because the event e_{think} in the current context is being used to explain something about event e_{boost} and is not directly relevant to the event e_{put}.

A.2 Verb-Noun Pairs

In this section, we provide details of annotations for the following examples of verb-noun_phrase pairs:

3. The most significant number of **deaths occurred** in New Orleans , Louisiana , which flooded as the levee system catastrophically failed , in many cases hours after the storm had moved inland. ($\neg C$)

   The label $\neg C$ is assigned to the pair "deaths-occurred" because both represent the same state of affairs i.e., "occurring of deaths".

4. Prior to the war, the governments of the United States and the United Kingdom claimed that Iraq's **alleged possession** of weapons of mass destruction (WMD) **posed** a threat to their security and that of their coalition
   regional allies. (C)

   The label C is assigned to the pair "alleged possession-posted" because this pair satisfies the conditions of manipulation theory of causality i.e., (1) "alleged possession" is an event that happened before the event represented by the verb "posed" and (2) while keeping as many state of current affairs constant if we can modify/control the event "alleged possession" then we can predictably modify/control the event represented by the verb "posed".

Appendix B

Frame Elements for the Cause and Non-cause Relations

This appendix shows assignments of the frame elements of FrameNet corpus to the labels of Cause (C) and Non-cause ($\neg C$) relations. We also present the list of unassigned frame elements (see section 5.1.1 for details). It took us around 3 hours to assign the following 729 frame elements to the C and $\neg C$. These assignments are as follows:

- C**:** Cause, Purpose, Reason, Explanation, Required situation, Purpose of event, Negative consequences, Resulting action, Internal cause, Result, External cause, Effect, Cause of shine, Purpose of goods, Response action, Enabled situation, Grinding cause, Trigger, Resulting action, Stimulus, Preventing cause, Purpose of created entity, Response, Purpose of recipient, Imposed purpose, Resultant situation, Purpose of theme, Enablement, Enabled action.
- $\neg C$: Place, Speed, Driver, Attribute, Time, Path, Manner, Duration, Means, Activity, Group, Protagonist, Difference, Process, Content, Executioner, Amount of progress, Treatment, Sender, Holding location, Food, Produced food, Copy, Source, Original, Creator, Iteration, Frequency, Characterization, Agent, Body part, Depictive, Theme, Subregion, Angle, Fixed location, Path shape, Direction, Area, Degree, Sub-region, Addressee, Coordinated event, Entity, Road, Distance, Speaker, Information, Medium, Topic, Clothing, Wearer, Bodypart of agent, Locus, Cognizer, Mental content, Salient entity, Action, Experience, Message, Name, Ground, Inspector, Unwanted entity, Location of inspector, Researcher, Question, Population, Searcher, Instrument, Created entity, Components, Forgoer, Bad entity, Dodger, Vehicle, Self mover, Containing event, Circumstances, Re encoding, Cotheme, Individuals, Complainer, Complaint, Communicator, Final value, Item, Initial value, Value range, Co participant, Phenomenon, Target symbol, Location of perceiver, Perceiver agentive, State, Location, Expected entity, Forgery, Experiencer, Focal participant, Event, Time of event, Variable, Limit1, Limit2, Limits, Point of contact, Goods, Lessee, Lessor, Money, Rate, Unit, Perceiver passive, Appraisal, Inference, Sound, Location of source, Dependent state, Noisy event, Official, Selector, Role, Function, Fidelity, Evidence, New leader, Body, Standard, Old leader, Old order, Leader, Governed, Result size, Size change, Dimension, Elapsed time, Paradigm, Focal occasion, Landmark occasion, Interval, Category, Criteria, Text, Correlate, Final correlate, Initial correlate, Sides, Side 1, Side 2, Issue, Perpetrator, Value 2, Value 1, Actor, Partner 2, Partner 1, State of affairs, Figure, Resident, Co resident, Partners, Subject, Institution, Level, Qualification, Student, Teacher, Undesirable event, Undergoer, Course, Subregion bodypart, Norm, Act, Phenomenon 2, Phenomenon 1, Quality, Phenomena, Owner, Possession, Support, Proposition, Domain of relevance, Charges, Defendant, Judge, Idea, Location of appearance, Material, Accused, Arraign authority, Hair, Configuration, Emitter, Beam, Initial subevent, Hypothetical event, Evaluee, Seller, Buyer, Recipient, Relative location, Connector, Items, Part 1, Part 2, Parts, Whole, Motivation, Fine, Payer, Executed, Interlocutor 2, Interlocutor 1, Affliction, Medication, Healer, Container, Cook, Temperature setting, Resource, Resource controller, Heating instrument, Donor, Last subevent, Constant location, Carrier, Transport means, Co theme, Rope,

Knot, Handle, Containing object, Enclosed region, Container portal, Fastener, Aggregate, Suspect, Authorities, Offense, Source of legal authority, Patient, Ingestor, Ingestibles, Sleeper, Pieces, Goal area, Mode of transportation, Ingredients, Cognizer agent, Excreter, Excreta, Air, Perceptual source, Interlocutors, Undesirable situation, Undesirable location, Escapee, Capture, Pursuer, Evader, Periodicity, Author, Honoree, Reader, Child, Father, Mother, Egg, Flammables, Flame, Kindler, Mass theme, Address, Intermediary, Communication, Location of communicator, Firearm, Sleep state, Indicated entity, Hearer, Sub region, Member, Object, Organization, New status, Guardian, Arguer, Reversive, Liquid, Force, Legal basis, Voice, Precipitation, Duration of endstate, Period of iterations, Employer, Employee, Position, Field, Place of employment, Amount of work, Task, Contract basis, Criterion, Recipients, Temperature goal, Temperature change, Dryee, Initial state, Traveler, Iterations, Deformer, Resistant surface, Fluid, Injured party, Avenger, Injury, Offender, Grinder, Profiled item, Standard item, Profiled attribute, Standard attribute, Finding, Case, Emission, Item 1, Item 2, Form, Chosen, Role of focal participant, Injuring entity, Severity, Substance, Delivery device, World state, Wrong, Amends, Entry path, Emotion, Emotional state, Grounds, Expressor, New idea, Basis, Manufacturer, Product, Factory, Consumer, Outcome, Interested party, Production, Artist, Studio, Performer, Distributor, Scene, Performance, Performer1, Performer2, Whole patient, Undertaking, Exporting area, Importing area, Eventuality, Time of eventuality, Accuracy, Indicator, Audience, Valued entity, Journey, Parameter, Destination time, Landmark time, Arguers, Arguer2, Arguer1, Company, Asset, Intended event, Origin, Initial size, Sound maker, Static object, Themes, Following distance, Intended perceiver, Cognate event, Location of expressor, Path of gaze, Relatives, Final temperature, Language, Means of communication, Benefit, Occasion, Time length, Reference point, Completeness, Faculty, Skill, Determinant, Feature, Element, Believer, Remainder, Original context, Final state, Final category, Transitional period, Sign, Part of form, Formal realization, Individual 2, Context, Individual 1, Particular iteration, Executive authority, Counter actor, Misdeed, Wrongdoer, Prisoner, Prison, Precept, Old, New, Pattern, Entity 1, Entity 2, Differentiating fact, Entities, Participants, Participant 1, Degree of involvement, Participant 2, Agreement, Signatory, Beneficiary, Total, Part, Contrast set, Exchangers, Exchanger 2, Theme 1, Theme 2, Exchanger 1, Exchange service, Exchange rate, Money owner, Sum 1, Concessive, Old tool, New tool, Body location, Unconfirmed content, Condition, Circumstance, Authority, Label, Duration of final state, Term, Location of protagonist, Re-encoding, Subevent, Artifact, Tester, Tested property, Unwanted characteristics, Side, Defender, Assailant, Post state, Party 1, Obligation, Parties, Party 2, Deliverer, Supplier, Helper, Focal entity, Project, Prior state, Grantor, Ratifier, Proposal, Dependent, Situation, Comparison set, Rank, Re enconding, Assessor, Value, Figures, Method, Astronomical entity, Patricular iteration, Fact, Obligator, Principle, System, Operator, Capitulator, Surrenderer, Cognizers, Cognizer 1, Cognizer 2, Opinion, Estimation, Reported fact, Participant, Documents, Submittor, Interceptor, Problem, Original path, Component, Expected event, Commitment, Exporter, Importer, Options, Initial category, Event description, Set, Initial number, Final number, Specified entity, Collection, Facility, Interlocutor, Proposed action, Contrast, Possibilities, Danger, Protection, Traveller, Source symbol, Represented, Depictive of represented, Representation, Location of representation, Characteristic, Hidden object, Hiding place, Obstruction, Potential observer, Device, Plan, New member, Contents, Front, Illicit organization, Dispute, Decision maker, Bad outcome, Abuser, Time span, New duration, Attack, Survivor, Dangerous situation, Colonists, Descriptor, New area, Homeland, Controlling entity, Dependent entity, Dependent situation, Rule, Weapon, Earnings, Earner, Shopper, Host, Co-guest, Guest, Events, Targeted, Target location, Tourist, Attraction, Final element, Topical entity, Supported, Supporter, Hunter, Crop, Gatherer, Agriculturist, Grower, Co-participant, Conqueror, Invader, Land, Enemy, Invasion act, Grantee, Other, Concept 1, Concept 2, Concepts, Epistemic stance, Competitor, Competition, Margin, Score, Opponent, Venue, Prize, Loser, Winner, Specified content, Unresolved referent, Perceiver, Sufferer, De-

ceased, Game, Speech, Existing member, Potential recipient, Gambler, Uncertain situation, Sub part, Sensory attribute, Projectile, Court, Sentence, Term of sentence, Location of confinement, Decision, Referent, Mode of transfer, Countertransfer, Circumstnaces, Informer, Target, Investigator, Incident, Fugitive, Representative, Jury, Possible sentence, Witness, Questioner, Domain, Claimant, Property, Numbers, Endangering act, Alterant, Eclipsed, Vantage point, Point of view, Crime, Governing authority, Location of sound source, Competitors.

- **Unassigned frame elements:** Goal, Impactee, Impactor, Sought entity, Desirable, Desired goal, Intended goal, Sound source, Manipulator, Victim, Name source, Impactors, Compensation, Hot cold source, Hot/cold source, Punishment, Source emitter, Sub source, Change agent, Destroyer, Killer, Destination event, Goal conditions, Heat source, Original punishment, New punishment, Conviction, Target currency, Source currency, Source material, Desired state, Responding entity, Benefited party, Affected, Current jurisdiction, Crime jurisdiction, Requirement, Required entity, Simulated entity, Source of information, Convict, Intended purpose, Hindrance, Affected party

Appendix C

Frame Elements for the Semantic Classes of Nouns

This appendix shows assignments of the frame elements of FrameNet corpus to the labels of semantic classes of nouns C_{np} and $\neg C_{np}$. We also present the list of unassigned frame elements (see section 5.2.1 for details). It took us around 3 to 4 hours to assign the following 936 frame elements to the C_{np} and $\neg C_{np}$. These assignments are as follows:

- C_{np}**:** Characterization, Event, Goal, Purpose, Cause, Internal Cause, External cause, Result, Means, Reason, Phenomena, Coordinated Event, Action, Activity, Circumstances, Desired Goal, Explanation, Persistent characteristic, Ethnicity, Context of acquaintance, Contract basis, Information, Topic, Completeness, Theme, Experience, Context, Use, Source, Manner, Desirable, Cotheme, Opinion, Complaint, Intended goal, Phenomenon, State, Simulated entity, Forgery, Idea, Content, Evidence, Perceiver passive, Percept, Old order, Domain, Contents, Criteria, Attribute, Difference, Correlate, Issue, Victim, Relationship, Ego, Examination, Knowledge, Qualification, Measurement, Power source, Hypothetical event, Undesirable event, Norm, Act, State of affairs, Proposition, Support, Possession, Social event, Refreshment, Occasion, Configuration, Skill, Process, Occupant, Expression, Intended perceiver, Conveyed emotion, Effect, Offense, Accoutrement, Cause of shine, Formational cause, Charges, Behavior, Characteristic, Practice, Response action, Treatment, Affliction, Medication, Resource controller, Resource, Competition, Perceptual source, Undesirable situation, Communication, Situation, Desired state, New status, Decoration, Impactors, Impactee, Impactor, Suspect, Specification, Precipitation, Task, Compensation, Fidelity, Orientation, Travel means, Mode of transportation, Enabled situation, Injury, Punishment, Dispute, Finding, Possibilities, Trigger, Stimulus, Wrong, Amends, Intoxicant, Evaluation, Empathy target, Undertaking, Factor, Phenomenon 2, Phenomenon 1, Nationality, Performance, Score, Indicated, Sound source, Noisy event, Themes, Perceiver, Means of communication, Law, Required, Forbidden, Jurisdiction, Religion, Influencing situation, Dependent situation, Influencing variable, Misdeed, Severity, Pattern, Feature, Signatory, Agreement, Question, Relevant feature, Requirement, Project, Differentiating fact, Epistemic stance, Unconfirmed content, Response, Obligation, Crime jurisdiction, Current jurisdiction, Landmark feature, Focal feature, Supply, Force, Estimation, Requirements, Desired state of affairs, Consideration, Competing consideration, Artifact, Ammunition, Economy, Alliance, Employment, Connections, Proposed action, Represented, State of represented, Dep, Eventuality, Asset, Harmful event, Events, Defining event, Preparatory act, Event description, Work, Ailment, Symptom, Salient event, Presentation, Impression, Opportunity, Desirable situation, Final correlate, Initial correlate, Target, Status, Rel, Idiom, Connection, Specialty, Containing event, Convict, Sentence, Purpose of theme, Security, Benefiting action, New idea, Meaning, Form, Source of relationship, Comparand, Assailant, Incident, Decision, Representative, Affected party, Witness, Restrictor, Assessor, Claimant, Fight, Salient activity, Prosecution, Jury, Plea, Mental content, Sought entity, Re encoding, Required situation, Purpose of event, Appraisal, Inference, Dependent state, Official, Function, Paradigm, Focal occasion,

Landmark occasion, Manipulator, Resulting action, Domain of relevance, Arraign authority, Initial subevent, Purpose of goods, Motivation, Name source, Last subevent, Transport means, Co theme, Source of legal authority, Cognizer agent, Excreta, Capture, Mass theme, Sleep state, Criterion, Hot cold source, Hot/cold source, Initial state, Grinding cause, Profiled attribute, Standard attribute, Emission, Source emitter, Sub source, Change agent, Injuring entity, World state, Emotion, Emotional state, Accuracy, Journey, Destination event, Intended event, Goal conditions, Heat source, Cognate event, Benefit, Remainder, Final state, Formal realization, Original punishment, New punishment, Conviction, Prison, Precept, Theme 1, Theme 2, Exchange service, Condition, Circumstance, Re-encoding, Source material, Preventing cause, Purpose of created entity, Subevent, Unwanted characteristics, Post state, Purpose of recipient, Affected, Prior state, Proposal, Imposed purpose, Re enconding, Method, Fact, Required entity, Obligator, Principle, Reported fact, Problem, Resultant situation, Expected event, Commitment, Options, Collection, Facility, Contrast, Protection, Depictive of represented, Representation, Obstruction, Plan, Attack, Dangerous situation, Controlling entity, Rule, Earnings, Attraction, Source of information, Enablement, Enabled action, Invasion act, Concept 1, Concept 2, Concepts, Sufferer, Uncertain situation, Sensory attribute, Intended purpose, Mode of transfer, Countertransfer, Hindrance, Circumstnaces, Fugitive, Possible sentence, Endangering act, Point of view, Crime.

- $\neg C_{np}$**:** Artist, Performer, Duration, Time, Place, Distributor, Area, Path, Direction, Sub-region Frequency, Body part, Degree, Angle, Fixed location, Path shape, Addressee, Interval, Person, Age, Place of employment, Rank, Type, Employer, Speaker, Medium, Addressee, Original, Copy, Cognizer, Salient entity, Garment, Anchor, Entity, Inspector, Ground, Vehicle, Forgoer, Distance, Goods, Importing area, Complainer, Location of perceiver, Selector, New leader, Old leader, Leader, Protagonist, Item, Result size, Size change, Dimension, Instrument, Rate, Elapsed time, Initial size, Group, Container, Relative location, Material, Part, Owner, Category, Initial value, Final value, Side 2, Sides, Side 1, Perpetrator, Partner 2, Partner 1, Partners, Individuals, Aggregateproperty, Alter, Relatives, Gizmo, User, Examiner, Examinee, Object, Location, Resident, Student, Institution, Level, Teacher, Subject, Undergoer, Quality, Co resident, Possessor, Orientational location, Attachment, Subregion, Defendant, Attendee, Host, Honoree, Subtype, Type property, Item property, Hair, Hair property, Accessory, Wearer, Focal participant, Body location, Expressor, Piece, Whole, Evaluee, Unit, Stuff, Count, Size, Seller, Buyer, Money, Figure, Light, Mass, Quantity, Locale, Name, Constituent parts, Container possessor, Accused, Subpart, Clothing, Roadway, Length, Endpoints, Frequency of use, Abundant entities, Addictant, Addict, Relative time, Interlocutors, Interlocutor 2, Interlocutor 1, Patient, Communicator, Experiencer, Participants, Participant 2, Participant 1, Components, Created entity, Ingestibles, Air, Evader, Pursuer, Road, Speed, Child, Father, Mother, Location of communicator, Sleeper, Author, Title, Time of creation, Genre, Member, Organization, Arguer, Picture, Decorated individual, Decoration desriptor, Individual, Holding location, Temperature, Employee, Position, Field, Creator, Shape, Traveler, Period of iterations, Co participant, Dryee, Avenger, Offender, Injured party, Decision maker, Judge, Legal basis, Items, Item 2, Item 1, Parameter, Chosen, Business, Proprietor, Product, Business name, Time of event, Location of event, Basis, Society, Ingestor, Penal institution, Operator, Inmates, Country of origin, Body mark, Extent of acclaim, Signs, Sign 1, Sign 2, Factory, Interested party, Manufacturer, Production, Studio, Performer2, Performer1, Connector, Connected item, Exporting area, Weapon, Wielder, Indicator, Itinerary, Audience, Emitter, Beam, Component sound, Location of sound source, Margin, Venue, Language, Iteration, Reference point, Faculty, Determinant, Rest, Noise maker, Building, X, Executive authority, Wrongdoer, Degree of involvement, Alterant, Researcher, Old, Transferors, Donor, Recipient, Entity 2, Entity 1, Entities, Numbers, Term, Indicated resident, Location of inspector, Consumer, Tester, Tested property, Side, Helper, Benefited party, Focal entity, Party 1, Party 2, Parties, Deliverer, Participant, Gov, Infrastructure, Value, Interceptor, Electricity, Political region, Member 2, Member

1, Members, Landmark, Period, Fastener, Container portal, Set, Judicial body, Judges, Sent item, Complex, Network, Nodes, Interlocutor, Population, Period of existence, Sub part, Time period, Current country, Professional, Terrorist, Guest, Start time, End time, Time span, Host location, Food, Conqueror, Invader, Land, Practitioner, Loser, Winner, Relation type, Value range, Ant, Building part, Individual1, Term of sentence, Court, Keeps, Entity2, Entity1, Color, Investigator, Questioner, Grantor, Number, Informer, Subordinate, Superior, Combatants, Issuer, Document, Bearer, Authority, Bodypart of agent, Locus, Searcher, Bad entity, Dodger, Driver, Target symbol, Expected entity, Variable, Limit1, Limit2, Limits, Point of contact, Lessee, Lessor, Sound, Location of source, Body, Value 2, Value 1, Actor, Course, Subregion bodypart, Location of appearance, Amount of progress, Part 1, Part 2, Parts, Fine, Payer, Cook, Heating instrument, Temperature setting, Constant location, Carrier, Sender, Rope, Knot, Handle, Containing object, Enclosed region, Pieces, Produced food, Ingredients, Excreter, Undesirable location, Escapee, Periodicity, Reader, Egg, Kindler, Address, Intermediary, Firearm, Indicated entity, Hearer, Sub region, Guardian, Reversive, Liquid, Coparticipant, Voice, Duration of endstate, Amount of work, Recipients, Temperature goal, Iterations, Deformer, Resistant surface, Fluid, Grinder, Profiled item, Standard item, Role of focal participant, Delivery device, Entry path, Grounds, Scene, Whole patient, Destroyer, Time of eventuality, Valued entity, Destination time, Landmark time, Company, Sound maker, Static object, Following distance, Location of expressor, Path of gaze, Final temperature, Time length, Element, Believer, Final category, Transitional period, Sign, Part of form, Individual 2, Individual 1, Particular iteration, Counter actor, Prisoner, New, Beneficiary, Total, Contrast set, Exchangers, Exchanger 2, Exchanger 1, Target currency, Source currency, Exchange rate, Money owner, Sum 1, Concessive, Old tool, New tool, Label, Duration of final state, Location of protagonist, Defender, Supplier, Ratifier, Comparison set, Figures, Astronomical entity, Patricular iteration, System, Capitulator, Surrenderer, Cognizers, Cognizer 1, Cognizer 2, Documents, Submittor, Original path, Component, Exporter, Importer, Initial category, Initial number, Final number, Specified entity, Danger, Traveller, Source symbol, Location of representation, Hidden object, Hiding place, Potential observer, Device, New member, Front, Illicit organization, Abuser, New duration, Initial duration, Survivor, Colonists, New area, Homeland, Earner, Shopper, Coguest, Targeted, Target location, Tourist, Final element, Topical entity, Supported, Supporter, Hunter, Crop, Gatherer, Agriculturist, Grower, Co-participant, Enemy, Other, Specified content, Unresolved referent, Deceased, Game, Speech, Existing member, Potential recipient, Gambler, Projectile, Location of confinement, Referent, Vantage point, Governing authority.

- **Unassigned frame elements:** Origin, Case, Service provider, Signature, Responding entity, Grantee, Property, Goal area, Flammables, Flame, Temperature change, Descriptor, Namesake, Agent, Unwanted entity, Self mover, Depictive, Message, Perceiver agentive, Standard, Governed, Role, Aggregate, Style, Instance, Behavior product, Substance, Piece prop, Part prop, Compeller, Executioner, Executed, Healer, Prize, Text, Containing text, Authorities, Shape prop, Outcome, Killer, Arguer1, Arguer2, Arguers, Opponent, Competitor, Competitors, Standing, Original context, Dependent variable, Dependent entity, Influencing entity, Dependent, Instance prop, Output, Odds, Bad outcome, Eclipsed.[Radinsky and Horvitz 2013, Girju 2003]

References

[Ando and Zhang 2005a] Ando, R. K. and Zhang. T. 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research.*

[Ando and Zhang 2005b] Ando, R. K. and Zhang. T. 2005. A High-Performance Semi-Supervised Learning Method for Text Chunking. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics..*

[Asher and Lascarides 2003] Asher, N. and Lascarides, A. 2003. Logics of Conversation. *Cambridge University Press.*

[Baker et al. 1998] Baker, C. F., Fillmore, C. J., and Lowe, J. B. 1998. The Berkeley FrameNet project. *In proceedings of the Association for Computational Linguistics and International Conference on Computational Linguistics (COLING-ACL).*

[Barzilay et al. 2002] Barzilay, R., Elhadad, N., and McKeown, K. 2002. Inferring strategies for sentence ordering in multidocument summarization. *Journal of Artificial Intelligence Research (JAIR).*

[Beamer and Girju 2009] 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 (CICLing).*

[Bethard 2007] Bethard, S. 2007. Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach. *Ph.D. Dissertation. University of Colorado at Boulder, Boulder, CO, USA.*

[Bethard and Martin 2006] Bethard, S. and Martin, J. H. 2006. Identification of Event Mentions and their Semantic Class. *In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06).*

[Bethard and Martin 2008] Bethard, S. and Martin, J. H. 2008. Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (HLT-Short '08).*

[Bethard et al. 2007] Bethard, S., Martin, J. H., and Klingenstein, S. 2007. Finding Temporal Structure in Text: Machine Learning of Syntactic Temporal Relations. *In International Journal of Semantic Computing (IJSC) 1(4), pp. 441458.*

[Carlson et al. 2002] Carlson, L., Marcu, D., and Okurowski, M. E. 2002. RST Discourse Treebank. *Linguistic Data Consortium.*

[Chambers and Jurafsky 2008] Chambers, N. and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. *In Proceedings of the Joint Conference of Association for Computational Linguistics and Human Language Technologies (ACL-HLT).*

[Chambers and Jurafsky 2009] Chambers, N. and Jurafsky, D. 2009. Unsupervised learning of narrative schemas and their participants. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*

[Chambers et al. 2007] Chambers, N., Wang, S., and Jurafsky, D. 2007. Classifying Temporal Relations Between Events. *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL).*

[Chan and Roth 2010] Chan, Y. and Roth, D. 2010. Exploiting background knowledge for relation extraction. *In Proceedings the International Conference on Computational Linguistics (COLING).*

[Chang and Choi 2006] Chang, D. and Choi, K. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information Processing and Management, 2006, volume 42 issue 3, 662678.*

[Chklovski and Pantel 2004] Chklovski, T and Pantel, P. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. *In proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP '04).*

[Clarke 2008] Clarke, J. 2008. Global Inference for Sentence Compression: An Integer Linear Programming Approach. *Ph.D. Dissertation. School of Informatics, University of Edinburgh, UK.*

[Cooper 1997] Cooper, G. F. 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery, 1997, volume 1, 203-224.*

[Do 2012] Do, Q. X. 2012. Background Knowledge in Learning-Based Relation Extraction. *Ph.D. Dissertation. University of Illinois at Urbana-Champaign, IL, USA.*

[Do and Roth 2012] Do, Q. X. and Roth, D. 2012. Joint inference for event timeline construction. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning (EMNLP-CoNLL)..*

[Do et al. 2011] Do, Q., Chen, Y. S., and Roth, D. 2011. Minimally Supervised Event Causality Identification. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*

[Dowty 1979] Dowty, D. R. 1979. Word Meaning and Montague Grammar. *D. Reidel Publishing Company, Dordrecht, 1979.*

[Finkel et al. 2005] Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *In Proceedings of the Association for Computational Linguistics (ACL).*

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

[Girju and Moldovan 2002] Girju, R. and Moldovan, D. 2002. Mining Answers for Causation Questions. *In American Associations of Artificial Intelligence (AAAI '02) symposium.*

[Girju et al. 2009] Girju, R., Beamer, B., Rozovskaya, A., Fister, A., and Bhat, S. 2009. A knowledge-rich approach to identifying semantic relations between nominals. *Journal of Information Processing and Management, volume 46, issue 5. 2009.*

[Granger 1969] Granger, C. W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometricia, 1969, Volume 37 Issue 3, 424-438.*

[Kamp and Reyle 1993] Kamp, H. and Reyle, U. 1993. From Discourse To Logic. Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. *Dordrecht, The Netherlands: Kluwer Academic Publishers.*

[Khoo et al. 2000] Khoo, C. S. G., and Chan, S., and Niu, Y. 2000. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. *In Proceedings of 38th Annual Meeting of the ACL, 2000.*

[Kim 1993] Kim, J. 1993. Causes and Events. Mackie on Causation. *In Cansation, Oxford Readings in Philosophy, ed. Ernest Sosa, and Michael Tooley, Oxford University Press, 1993.*

[Kipper et al. 2000] Kipper, K., Dang, H. T., and Palmer, M. 2000. Class-based construction of a verb lexicon. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.*

[Klein and Manning 2003] Klein, D. and Manning, C. D. 2003. Accurate Unlexicalized Parsing. *In Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.*.

[Levin 1993] Levin, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. *University of Chicago Press, Chicago, IL.*

[Lin 1998] Lin, D. 1998. An Information-Theoretic Definition of Similarity. *In proceedings of the 15th International Conference on Machine Learning (ICML).*

[Macleod et al. 1998] Macleod, C., Grishman, R., Meyers, A., Barrett, L., Reeves, R. 1998. NOMLEX: A Lexicon of Nominalizations. *In proceedings of EURALEX.*

[Mani et al. 2006] Mani, I., Verhagen, M., Wellner, B., Lee, C. M., and Pustejovsky, J. 2006. Machine learning of temporal relations. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.*

[Mann and Thompson 1987] Mann, W. C. and Thompson, S. A. 1987. Rhetorical structure theory: A theory of text organization. *Technical Report ISI/RS-87-190, ISI, Los Angeles, CA.*

[Marcu and Echihabi 2002] Marcu, D. and Echihabi, E. 2002. An unsupervised approach to recognizing discourse relations. *In proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02).*

[Markert and Nissim 2009] Markert, K. and Nissim, M. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation Volume 43 Issue 2, Pages 123−138.*

[Marneffe et al. 2006] Marneffe. M. D., MacCartney, B., and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC).*

[McCallum 2002] McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

[Menzies 2008] Menzies, P. 2008. Counterfactual theories of causation. *Online Encyclopedia of Philosophy, 2008.*

[Miller 1990] Miller, G. A. 1990. WordNet: An online lexical database. *International Journal of Lexicography, 1990, volume 3 issue 4, 235-244.*

[Pearl 2000] Pearl, J. 2000. Causality. *Cambridge University Press.*

[Pitler and Nenkova 2009] Pitler, E. and Nenkova, A. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. *In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort '09). Association for Computational Linguistics.*

[Pitler et al. 2009] Pitler, E., Louis, A., and Nenkova, A. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*

[Prasad et al. 2008] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, J., and Webber., B. 2008. The penn discourse treebank 2.0. *In proceedings of LREC 2008.*

[Pustejovsky et al. 2003] Pustejovsky, J., Hanks, P., Saur, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo., M. 2003. The TIMEBANK Corpus. *In Proceedings of Corpus Linguistics 2003, 647–656.*

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

[Riaz and Girju 2010] Riaz, M and Girju, R. 2010. Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision. *In proceedings of the IEEE 4th International Conference on Semantic Computing (ICSC '10).*

[Riaz and Girju 2013] Riaz, M and Girju, R. 2013. Toward a Better Understanding of Causality between Verbal Events: Extraction and Analysis of the Causal Power of Verb-Verb Associations. *In proceedings of the 14th annual meeting of Special Interest Group in discourse and dialogue (SIGdial '13).*

[Riaz and Girju 2014a] Riaz, M and Girju, R. 2014. Recognizing Causality in Verb-Noun Pairs via Noun and Verb Semantics. *In proceedings of the Workshop on Computational Approaches to Causality in Language, 14th Conference of the European Chapter of the Association for Computational Linguistics.*

[Riaz and Girju 2014b] Riaz, M and Girju, R. 2014. Hit the Nail on the Head: Employing Right Types of Knowledge for Identifying Causality between Events. *Journal of Dialogue and Discourse (D&D). In review.*

[Riaz and Girju 2014c] Riaz, M and Girju, R. 2014. In-depth Exploitation of Noun and Verb Semantics to Identify Causation in Verb-Noun Pairs. *The 15th annual meeting of Special Interest Group in discourse and dialogue (SIGdial '14).*

[Roth and Yih 2004] Roth, D. and Yih, W. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. *In Proceedings of the Annual Conference on Computational Natural Language Learning (CoNLL).*

[Roth and Yih 2007] Roth, D. and Yih, W. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to Statistical Relational Learning. MIT Press.*

[Saeed 1997] Saeed, J. L. 1997. Semantics. *Blackwell Publishers, 1997..*

[Sanders et al. 1992] Sanders, T. J. M., Spooren, W. P. M. S., and Noordman, L. G. M. 1992. Toward a taxonomy of coherence relations. *Discourse Processes, volume 15 issue 1..*

[Sauri et al. 2005] Sauri, R., Knippen, R., Verhagen, M. and Pustejovsky, J. 2005. Evita: A robust event recognizer for QA systems. *In Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005.*

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

[Soricut and Marcu 2003] Soricut, R. and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. *In proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003.*

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

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

[Toutanova et al. 2003] Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of Human Language Technology and North American Chapter of the Association for Computational Linguistics (HLT-NAACL).*

[Vendler 1957] Vendler, Z. 1957. Verbs and times. *Philosophical Review, 56:143160, 1957.*

[Verkuyl 1972] Verkuyl, H. J. 1972. On the Compositional Nature of the Aspects. *D. Reidel Publishing Company, Dordrecht, 1972.*

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