

Dissertation

Extracting Causal Relations between News Topics from Distributed Sources

submitted in partial satisfaction of the requirements
for a degree of Doktoringenieur (Dr.-Ing.)

at

Technische Universität Dresden,
Faculty of Computer Science

by

M.Sc. Eduardo Jacobo Miranda Ackerman

born on June 15, 1977 in Tijuana, Baja California, México.

Advisers:

Prof. Dr. rer. nat. habil. Dr. h. c. Alexander Schill

Prof. Dr. Ing. Michael Schroeder

Dr. Manuel Montes y Gómez

Date of Defense: November 8, 2013

Dresden, December 5, 2013

Statement of Authorship

This dissertation has been conducted and presented solely by myself. I have not made use of other peoples work (published or otherwise) or presented it here without acknowledging the source of all such work.

Dresden December 5, 2013

Eduardo Jacobo Miranda Ackerman

Abstract

The overwhelming amount of online news presents a challenge called news information overload. To mitigate this challenge we propose a system to generate a causal network of news topics. To extract this information from distributed news sources, a system called *Forest* was developed. *Forest* retrieves documents that potentially contain causal information regarding a news topic. The documents are processed at a sentence level to extract causal relations and news topic references, these are the phrases used to refer to a news topic. *Forest* uses a machine learning approach to classify causal sentences, and then renders the potential cause and effect of the sentences. The potential cause and effect are then classified as news topic references, these are the phrases used to refer to a news topics, such as “The World Cup” or “The Financial Meltdown”. Both classifiers use an algorithm developed within our working group, the algorithm performs better than several well known classification algorithms for the aforementioned tasks.

In our evaluations we found that participants consider causal information useful to understand the news, and that while we can not extract causal information for all news topics, it is highly likely that we can extract causal relation for the most popular news topics. To evaluate the accuracy of the extractions made by *Forest*, we completed a user survey. We found that by providing the top ranked results, we obtained a high accuracy in extracting causal relations between news topics.

Acknowledgements

In memory of my father Carlos

for my children Jacobo and Sebastián

to my wife Karina

Thanks to my professor Alexander Schill,
my colleagues David Urbansky, Philipp Katz and Klemens Muthmann
and the support of Conacyt-DAAD

Contents

Contents	ix
1 Introduction	1
1.1 Use Cases	3
1.2 Intended Results	5
1.3 Hypotheses	5
1.4 Scope	7
1.5 Conventions	8
2 Related Works	11
2.1 Supporting Works	11
2.1.1 Topic Detection and Tracking	11
2.1.2 Topic Representations	13
2.2 News Topic References	16
2.2.1 NTR Validation	17
2.2.2 NTR Aliases	18
2.3 Topic Relations	19
2.3.1 Topic Similarity	19
2.3.2 Temporal Relations	20
2.3.3 Information Coherence	22
2.3.4 Causal Relations	23
2.4 News Text Processing	27
2.4.1 Document Summarization	27
2.4.2 Keyword Extraction	28
2.4.3 Named Entity Recognition	29
2.4.4 Event Extraction	29
2.5 News Understanding	31
3 Conceptual Architecture	35
3.1 General Overview	35
3.2 Processing Components	37
3.2.1 Processing Steps and Intermediate Results	37
3.3 Additional Modules	41
3.3.1 Presentation	41

3.3.2	Control Module	42
3.3.3	Task Manager	42
3.3.4	Processing Components	43
3.3.5	Processing Nodes	43
3.3.6	Data Sources	43
3.3.7	Information Repository	43
3.3.8	Configuration	44
3.4	Conclusion	44
4	News Topic References	45
4.1	Introduction	45
4.1.1	News Topic Reference Definition	46
4.1.2	News Topic Reference Types	46
4.2	Architecture Interface	47
4.2.1	Semantic Relations	47
4.3	News Topic Reference Validation	49
4.3.1	Input Text	49
4.3.2	Phrase Location and Frequency	49
4.3.3	Knowledge Based	52
4.3.4	Machine Learning	60
4.4	Approach Combination	61
4.4.1	Additional Approaches	62
4.4.2	Complimentary Approaches	63
4.5	News Topic Reference Alias	63
4.5.1	Wikipedia Redirect	63
4.5.2	Additional Approaches	65
4.5.3	Outlook	66
4.6	Conclusion	66
5	Causal Relation Extraction	67
5.1	Introduction	67
5.1.1	General Definition	68
5.1.2	Definition in literature	68
5.1.3	Interpretation	68
5.2	Additional Semantic relations	69
5.3	Causal Markers Approach	69
5.4	Causal Dataset	71
5.4.1	Generated Dataset	72
5.5	Machine Learning Approach	73
5.5.1	Proposed Features	73
5.5.2	Feature Extraction	76
5.5.3	Feature Evaluation	76
5.5.4	Classification Technique Evaluation	79
5.5.5	Cross Validation	80
5.6	Conclusions	80

6	Evaluation	81
6.1	Hypothesis I Evaluation	82
6.1.1	Survey Evaluation Units	83
6.1.2	User Survey	88
6.1.3	Survey Results	90
6.1.4	Conclusion	95
6.2	Hypothesis II Evaluation	95
6.2.1	News Topic References in Corpus	95
6.2.2	Causally Related News Topic References in Corpus	97
6.2.3	Conclusion	99
6.3	Hypothesis III Evaluation	100
6.3.1	Theoretical Evaluation	101
6.3.2	Practical Evaluation	102
6.3.3	Dataset Generation	102
6.3.4	User Survey	104
6.3.5	Result Analysis	106
6.3.6	Conclusion	108
6.4	Conclusions	109
7	Conclusions	111
7.1	Outlook	115
	Bibliography	117

Chapter 1

Introduction

The development of technologies for the Internet is making increasing amounts of information available to users, this increase brings with it a challenge in processing vast amounts of information. Because the same information can come from different sources, once a user selects one source the other sources become a distraction. There may also be inconsistencies between sources, this may confuse the user and generate a heavy cognitive load for the user to understand the information. The problem stated above is called “*information overload*”. This problem is particularly present for online news, because there are many sources generating the same information, and there are many media to present this information to the user, including blogs and email.

In this work we propose a system to reduce information overload of online news, the system is called *Forest*¹ it does this by generating a network of interconnected news topics. By interconnecting the news topics a context is provided, this makes it easier for the user to select news topics and understand them. Because of the overabundance of sources it is possible to use these distributed online news providers to generate the network automatically. The multiple sources are analyzed to generate a summary of a news topic and to find causal relations between news topics. This information is used to provide a summary overview to the user.

Because the information is extracted and not generated, the system relies on the news generators for the validity of the information, for this reason it is important for *Forest* to be able to reference back to the information source.

Generating a causal chain is a task not only useful in *Forest*, it can be useful within companies to track bugs, in finance to review influences and in biotechnology to find chemical interactions. The development of the system focused on news articles, as these provide a very broad scope in language use and source availability. The broad scope in language use, from formally to informally structured language, requires a capable language interpretation. It

¹“Die Herren dieser Art blendt oft zu vieles Licht; Sie sehn den Wald vor lauter Bäumen nicht” Musarion, Christoph Martin Wieland

is very challenging to achieve a robust interpretation without a large corpus, therefore we use online news as this is well suited for the task.

Several methods have previously been developed to provide some context for news articles or news topics. These include clustering news articles into domains or topics, another method is to present only a small selection of articles as not to overwhelm the user. To illustrate some of these methods we present examples found in prominent online news portals such as Google News². The domains shown in Google News include finance and politics, articles are clustered into these domains, and the context in this case is the domain. This method reduces the cognitive load by reducing the available articles thus minimizing the complexity of selecting a new topic. It is notable that a single news topic may be found in several domains with news articles that focus on different aspects of the topic, for example if a news topic is about a new law, one story may present the socio-political aspects and another story the economic effects, it follows that each article will be placed in a different domain. This reduces news information overload by presenting the topic in a context that may be familiar to the user. Another method is to cluster the articles about a single news topic and present only one article as representative of the news topic. A single news topic often has many stories because different news sources generate their own content, in most cases these stories contain much of the same information, therefore selecting one news article to represent the news topic is very useful for minimizing information overload. Many of the aforementioned techniques have been developed under the patronage of government agencies for news analysts.

These methods also provide an insight into the format in which stories are presented, namely in a list or column layout ordered by domain. This layout mimics the traditional news print layout of columns and folds, it also allows the user to move from one article to the next quickly and presents, what the editor considers, most relevant content upfront. One problem is that this layout has almost no considerations for inter-topic relations, therefore news topic context is not well presented in this layout.

Certain visual layouts do consider inter topic relations, for example temporal order or spatial proximity. These layouts intend to provide a time-line or facilitate discovery of novel interesting stories. To illustrate consider the now defunct Google Timeline, it presented stories in chronological order, this aided in the review of story development over time. Another layout is Google Fastflip, it placed unrelated stories in close spatial proximity, similar to a magazine layout, to aid the reader in discovering novel interesting content.

Some news portals use dynamic visual layouts to provide a more attractive environment for the user. For example in *The Economist* opinion cloud³ an animated network of interrelated entities is presented, the entities are extracted from user comments and the links are based on entities mentioned in the same comment. There are other, so called, visual news readers⁴ that use visualization aspects, like color and transparency, to convey information thus

²<http://news.google.com> accessed 1.3.2012

³<http://www.economist.com/conversation-cloud?days=1> accessed 1.3.2012

⁴<http://msnbcmedia.msn.com/i/msnbc/components/spectra/index.html> accessed 1.3.1012

aiding in understanding and navigating between news topics.

Some visual layouts present interconnected topics, a good example of this is Wikipedia, where some articles contain links to related articles and the articles have sections that provide a semantic relation type for the link. In *Forest* the semantic relations to link topics are causal. Because causal relations often provide a better context for a topic compared to other semantic relations such as temporal order or spatial proximity. Causal relations also make it possible for a user to find novel interesting content, based on the assumption that interest in a topic indicates interest in its causes or effects, and the assumption that the user does not know all the stated causes and effects of a topic.

1.1 Use Cases

In this section we present three use cases for *Forest*. First, we present a use case where the functionality of *Forest* is desired yet it is not available. Therefore the user mimics the functionality of *Forest*, afterwards we present a more specific use case of *Forest*, where a casual user reads the news online with the aid of *Forest*. At last we present an expert user utilizing the system to enhance articles where content is incomplete.

Use Case 1: User Generated Causal Network The user is browsing online news and discovers an interesting topic called “*The US financial meltdown*” of 2008. The user is aware that this topic is related to “*The mortgage meltdown*” topic, but it is not clear how they are related. While reading about “*The US financial meltdown*” the user notices that multiple sources provide the same or very similar information. This is redundant and confusing, because some information is slightly different between sources. The user is mainly interested in two aspects of the topic, namely how the user contributed to the topic and how the user is affected by the topic. To find out if the user has contributed to the topic, he searches for the causes of “*the US financial meltdown*”, after reviewing many articles he finds several articles with causal information. The causes are often stated in a single sentence or one paragraph and thereafter further explanation is given. The user notices different causes are stated in different stories, and some causes appear more frequently than others, it is difficult to keep track of the causes because some sources use different words to refer to the same cause. The main causes are deregulation and mortgage defaults, the user is confident he did not contribute to these causes. Afterwards the user searches for the effects of the topic, similar to finding the causes, he finds several effects. One of the effects stated in the stories is a reduction of liquidity in the users’ bank. The user understands what a reduction in liquidity means and takes action, he decides to transfer his account to another bank. Later on, the user discovers there has been a run on the bank and he has avoided it.

The user was able to better understand a news topic because he knows about the causes and effects of the topic. The main challenge in understanding the causes and effects of the topic is sorting through the information. The user has

taken steps similar to what we propose to do in *Forest* and has obtained results similar to what the system would provide.

Use Case 2: Novice User In this use case we illustrate how *Forest* could be used by novice user. The user sometime reads online news and has recently discovered *Forest*, he has found an opportunity to use it to correlate two known topics. In this case we will call the user a reader, as he is a reader of online news.

The reader is interested in traveling to Indonesia from Europe but can not find flights so he investigates the reason for this using *Forest*. He finds that currently there is a ban on Indonesian flights in Europe due to a flight accident. Deeper causes show that the flight accident was due to bad maintenance of the airplane. This is new interesting information to the reader. Because of reading the news earlier, the user is aware that there are some troubles with the ruler of Indonesia. The people of Indonesia protest that the ruler embezzles money from the government. The reader decided to use *Forest* to find a causal relation between news topics and starts with the Indonesian flight ban, there he finds that it was due to an airplane accident and that that accident was due to bad maintenance. The reader also notices that the maintenance company was given to one of the rulers' children and that it is a monopoly by law, this is also a cause for the protests. The reader noticed this because he reviewed one of the sources of the extracted information. The source article also states that the airline accused the maintenance company of providing substandard service possibly because of the monopoly.

The reader is skeptical, he thinks it is far fetched that the airlines accuse the maintenance company, so he reviews other sources in *Forest*. The reader notices that there are multiple articles that state something similar including one from the European airplane commission. The reader finds the exact paragraph that states that the company was able to provide abysmal quality because it had a government monopoly and the government did not require improvements.

The reader was able to get the information in context, the relationship between the protests, the maintenance company and the travel ban was clear. The reader was more interested in the information because he was able to link it to something he was initially interested in. It was easy for him to discover novel topics and correlate them to known topics, it was also simple to move from one topic to another because *Forest* provided relations between these topics. The reader was able to find the source of the information, to verify why the system provided these results, and was able to compare the results to other sources.

Use Case 3: Expert User The last use case presents a seasoned user of *Forest*. The user is an avid Wikipedia editor therefore we call him the editor. He is interested in providing references to an event article in Wikipedia, namely "*the CDS black market*". The editor found comments that may reflect reality but seem to contradict each other, also there are no citations for the contradicting information. The editor uses the title of the article as a query for *Forest* to generate a causal network. In *Forest* he finds that there are several causes for the

event in the article including *“the CDS oversight”* and *“the lack of oversight for the CDS”*, these contradict each other, therefore the editor reviews the phrases that are the source of the information, one phrase is *“oversight led to the creation of a black market for CDS”* and the other phrase is *“the CDS black market was created because of a lack of oversight”*. The editor can also see that there are many more sources for the former phrase than for the latter. The editor can now see why the comments are unclear and can provide a source reference for each of the contradictory comments. The references he chooses are the ones he considers most reliable and clearly explained. He can also comment that most sources for the first phrase are of political domain and most sources of the second phrase are of an academic domain.

The editor is able to access a graphical summary of the event and extract source information from that summary graph. And he is able to make a distinction between the different opinions about the causes by reviewing the sources.

1.2 Intended Results

To provide an illustration of the intended results, Figure 1.1 is provided. In the figure we see three panels: the left panel is a navigation control, to navigate between topics or to review the news articles that make the news topic. The navigation can also be used to highlight certain aspects of the topic extraction, for example news articles that provide similar information. In the middle panel we see the graphic visualization of the news topics and causal relation, each large circle represents a news topic, the smaller points around the circles represent the news articles that generate the topic. A line between topics represents a causal relation. The topics and relations have boxes that link to the source information. In the right panel the source information is provided, here the source news article is presented, with a highlight of the sentence where the information was extracted. The results of the system can also be provided in a compressed format, such as XML, to facilitate the transfer or for streaming update to the user.

1.3 Hypotheses

In this section the hypotheses are presented with a real world problem as a background for the hypotheses. The problem is the lack of context provided by online news portals for news articles and more generally news topics. Although news portals provide some context by linking to related articles, these links often do not supply an overview of how topics are interrelated. Links are predominantly about the same topic. The reason for the link may not always be clear and this may add to information overload.

Problem Currently online news portals do very little to provide context for news topics.

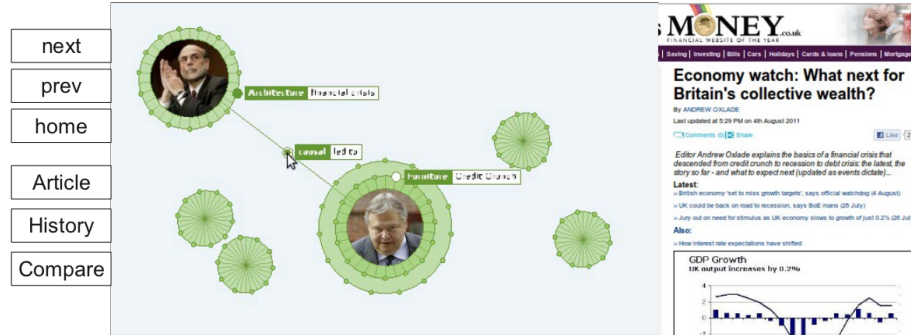


Figure 1.1: Sample Intended results

Question What type of link can be established between news topics to provide a context, and how can this link be established?

Hypothesis 1 Causal relations between news topics provide a context that aids the user to understand the news.

There are several semantic relation types that may be suitable for providing a context for a news topic, these include temporal order and part-of relations. Temporal order, for example, provides a semantic relation to review the progression of a topic, this is also called a time-line. A part-of relation or subsumes relation is where a news topic is part of another topic. This relation type may also be valid for context. For this work causal semantic relations were selected because this type of relation is often used to provide a deep understanding of a topic. This type of relation also allows users to relate unknown topics to known topics, thus supporting the understanding of the unknown topic.

Because this hypothesis deals with user understanding, it is subjective to each user opinion, despite this it can be objectively analyzed to show an improvement as compared to no context. A hindrance to understanding the news may develop if there is an excess of causal relations or inconsistencies between causal relations. This hypothesis deals with the relation type and not the amount or detail of the extracted relations.

Hypothesis 2 At least 95% of news topics can be found in a causal relation to another news topic.

It may not be possible to extract causal relations for all news topics, therefore it is necessary to define a scope for which news topics it is possible. We assume that most if not all news topics are causally related to another news topic because of the abundance of information. Often when a news topic is of high relevance it is reported by several sources, it follows that the more is written about the topic the more probable it is to

find a causal relation for the topic. The scope defined in the hypothesis also discards ephemeral topics because we assume that these often do not provide enough sources of information and there are very few of these type of topics.

Hypothesis 3 By analyzing news articles at a sentence level is it possible to extract explicit causal relations between news topic references with a 95% accuracy.

In contrast to causal relations found spread throughout a news article, causal relations at a sentence level are well suited to provide a summary overview of the related elements. The scope of this system defines the related elements as news topic references, because these are suitable to extend the causal network. There may be causal relations found between news topics and other elements, nonetheless these are not the focus of the system. The causal relations are defined as explicit to minimize the cognitive load and avoid confusion for the user, this is to say that it is easy for the user to understand and focus on the information. The analysis of the news article does not include the verification of veracity of the information. Because *Forest* does not generate the information, the extracted information may present the same inaccuracies as the source. Clearly defining the causally related news topics is at the core of the functionality of *Forest*, therefore this task should perform with a high accuracy in order to provide the best results for the user.

In the process of validating these hypotheses several other theories will be tested, for example: causal relations make past news articles currently relevant, and users that are interested in a news topic are also interested in the causes and effects of that topic. These theories are mostly objective in nature, therefore they are not the focus of this work, but they are still considered. The expected results of the evaluations are further refined with technical and journalistic requirements. These requirements are in line with reasonable expectations of a user reading online news, for example the ability to reference the source and author of the information and that the information be presented in complete semantic statements. These requirements are intended to improve the usability of *Forest*.

1.4 Scope

Because *Forest* is a system to extract information, it is limited by the sources of information. In this section we define this and other limitations. Regarding the quality of the information, the system is not intended to differentiate between truth and opinion. The system will provide tools to mitigate this problem, for example the sources can be selectively filtered to utilize only user trusted sources, also a link to the source is provided for the user to verify the validity of the extracted information. The intended

sources of information are online news producers, therefore causally related news topics are limited to those that are reported and available online. To evaluate the system and for focused applications, a news article corpus can and will be used as the source of information instead of online sources. To facilitate processing a single language is selected, namely English, as there is an abundance of online content generated in this language. Although the system is intended to work with low latency, the timeliness of the results is dependent on the speed at which the sources can be accessed. We intend to achieve a performance that is suitable for real world users. Because the system is not generative it is required to be initiated by the user, this is to say the user must provide the initial news topic.

improved performance. This is to say the system should improve performance when given more processing resources.

1.5 Conventions

To maintain nomenclature defined in previous work we define some conventions here. A news article including metadata such as author and date of production is called a *story*, these are the sources of the information. The human readable text from a story obtained from Internet is called *text*, this is in contrast to the non human readable content formatting such as HTML. The first paragraph in the story is called the *introduction*, many stories provide a summary in the introduction, and therefore it is often highly relevant. The *source* or corpus of information is a selection of documents extracted from online sources. the term *news topic* is defined based on *Topic Detection and Tracking*[Allan et al., 1998], it is the subject of discourse regarding the seminal events or activities that have been presented in the news. A *causal relation* is the semantic relation between two elements, where one is the cause and the other is the effect, and this relation is explicitly stated in a sentence or paragraph.

Document Structure

The dissertation is structured as follows: after this introduction we present some of the background material regarding news processing and seminal works regarding topic detection and tracking. These works include state-of-the-art systems that are comparable to *Forest* by having similar goals, results, or methods that are comparable to key components in *Forest*. To give an overview of the system, we present a conceptual architecture. The architecture highlights two main components news topic reference extraction and causal relation extraction. The following chapters are regarding these main components and how we developed a state-of-the-art approaches to deal with these tasks. Then we present an evaluation chapter

where the hypotheses are tested. To conclude we provide a conclusions and outlook chapter.

Chapter 2

Related Works

In this chapter we present related works. The chapter is divided into two main sections, first an introduction to the fundamental concepts used in *Forest*, these include works that define terms used in the domain, works that extract and define topics, works that deal with causal relations, and finally work that are valuable from the domain of text processing. In the second section we present works that are comparable to *Forest*, this is to say they have similar goals or methods. In the first section we highlight the requirements of *Forest* which motivates the scientific contributions developed in *Forest*. In the second section we highlight and contrast the key differences between *Forest* and other works that aid users in understanding the news.

2.1 Supporting Works

In this section we present the works that support the main concepts found in *Forest* and works in the text processing domain that are influential in *Forest*. To help us define terms used throughout the documents we begin with the seminal work of Topic Detection and Tracking that not only aids us in defining fundamental concepts such as topics and stories, but also introduces us to motivates the need for multiple text processing tasks including topic detection and tracking, document summarization, keyword extraction and named entity recognition.

2.1.1 Topic Detection and Tracking

Topic Detection and Tracking (TDT)[Allan et al., 1998] was a research initiative started in 1996 by the U.S. military research agency DARPA. The purpose of TDT is to develop technologies for the retrieval and automatic

organization of news and to evaluate the performance of these technologies. Some of these technologies are currently used in online news portals, such as Google News, to cluster stories and define the stories that are used to update a topic.

To compare the performance of different systems TDT provided an infrastructure that included an annotated corpus, evaluation software and task definitions. The key research applications defined in the tasks are the following:

Story Segmentation Given a continuous stream of news detect the boundaries between stories.

Topic Tracking Given a set of example stories define if a new story is about the same topic.

Topic Detection Given a collection of stories build clusters of stories about the same topic.

First Story Detection Detect if a story is the first story of a new unknown topic

Link Detection Given two stories define if they are about the same topic or if they are not.

Some research applications which were not the focus of TDT but aid in message understanding research area include document summarization, entity extraction and keyword extraction. Some of these works are reviewed in this chapter. Additional contributions of TDT are the definitions for several terms in the field of news analysis, some of these terms are stated below [Cieri et al., 2000].

Topic A seminal event or activities and all related events and activities.

Event A happening at a time and place, and all the unavoidable consequences.

Activity A connected set of actions that have a common focus or purpose.

Story A news article or segment of a news broadcast with a coherent news focus, including metadata such as source, publication time and additional information.

Because the distinction between topic and event is not always agreed upon, the Linguistic Data Consortium(LDC) has published a guideline to improve agreement and consistency of topic labeling.

In TDT a term that was also defined but later discarded was a *brief*, this was a mention about one topic within a story about another topic. This term was no longer used because of simplicity assumptions, made by TDT, to facilitate the tasks defined above. The simplicity assumptions provide the basis for future works that focus on addressing these faulty assumptions.

A faulty assumption is that each story refers to only one topic [Nallapati et al., 2004]. We will see in related works, how this assumption is addressed by segmenting a story into phrases and assuming each phrase is, at most, about one topic.

Topic Detection and Tracking gives us a theoretical definition of a topic that is used as the basis of the topic model that we developed, namely News Topic References.

2.1.2 Topic Representations

Topics can be represented by different types of models [Zhao and Wang, 2010]. In this section we will review the development of topic models with a focus on the requirement to semantically relate them. We begin by giving the concept of a topic, we consider a single document to be about a single topic, and therefore a cluster of similar documents are considered to be about a single topic. Thereafter we consider a collection of words represent the topic, therefore a document may contain one or more topics. Finally we review how certain topic models are better suited than others for semantic relation extraction. Given the presentation of the development and types of topics models, we show how these topic models are used in related works, or works that have similar goals.

Salton Vector Space Model

We begin with a document as a topic, this topic model works under the premise that each story contains one main topic, and therefore similar stories are about the same topic. In this case a collection of one or more documents represents a topic. The vector space model by Salton [Salton et al., 1975] is used to cluster similar documents from a given corpus therefore generating cluster of documents about a topic. In this model, and others based on it, the bag of words simplicity assumption is used, this is to say that the order of the terms in a document has a negligible affect on the analysis of the document. To find similar documents the method below is followed.

1. A term-document matrix is generated, where one dimension is a list of all the terms in the entire corpus, and another dimension is a list of all the documents in the entire corpus, the values in the matrix are the term frequencies in each document, or possibly another value such as TF-IDF. To simplify the matrix, words that occur in almost every document, known as stop words, and words that are very sparse are removed, these types of words adversely affect the similarity metrics.
2. A word-weight vector is generated to represents each document
3. To find similar documents a cosine distance between vectors is calculated, similar documents have a small angle between them.

One of the challenges of this model is handling synonyms and polysemy. Synonymy leads to lower recall because different words are used in the same topic. Polysemy leads to lower precision because a single word will be used in different topics, for example bank as in money and as in river.

Latent Semantic Indexing

Latent Semantic Indexing (LSI)[Deerwester et al., 1990] has been developed to mitigate the challenges in the vector space model. LSI works under the principle that terms with similar semantics are present in similar contexts, therefore the semantics of a term is extracted from its context and not only the term itself. Singular Value Decomposition is performed on the term-document matrix of the corpus to reveal the latent semantic information, the resulting matrix is called the “semantic space” matrix. In this model topics are represented as a vector in the “semantic space” matrix, therefore given a sample document it is possible to find similar documents although they may not have keywords in common. LSI is not limited to documents; it can be used for image recognition, classification and many other tasks. The implementation has three basic steps:

1. Generating the term-document matrix using a selected weight function, empirical studies show that Log Entropy weighting functions work well with many datasets
2. Rank-Reduced Singular Value Decomposition(SVD) of the term-document matrix. In this step the “semantic space” matrix is generated, to provide an intuition of SVD consider the following simplified example: take a matrix of the height vs. weight of a group of people, the data points will show a diagonal trend, a single vector from the shortest to the tallest may model the distribution, a second orthogonal vector may model the variance from the values to the original vector, now the original matrix is modeled in a reduced dimension matrix.
3. Use similarity measure such as cosine distance to find similar entities within the “semantic space” matrix.

Some of the issues in LSA is finding the optimal number of dimensions needed for the “semantic space” matrix to perform well. Another issue is that the resulting topic vectors may be difficult to interpret because the vector is based on the “semantic space” and not the complete term-document matrix. To make causal relations between news topics clear it is necessary to have a clear interpretation of a topic.

Probabilistic Topic Models

Probabilistic Latent Semantic Indexing (PLSA)[Hofmann, 1999] is a topic model that deals with some of the challenges in LSA, namely clarity in

representing a topic. Similarities between LSA and PLSA are that both find a latent semantic structure, LSA uses SVD to generate a “semantic space” matrix to find semantically similar concepts, while in PLSA uses latent variables (K) that can be considered topics. Each topic is a probability distribution over a word vector, so a topic can be represented by a sorted list of words. To fit the corpus to the topic model several methods can be used, a commonly used method is Expectation Maximization [Dempster and Laird, 1977]. One of the improvements on PLSA is Latent Dirichlet Allocation (LDA) [Blei and Ng, 2003], where the probability distribution of a topic in a document, is a Dirichlet distribution, in other words there is a distribution over the distribution of topics in a document. Several methods can be used to fit this model, a commonly used method is collapsed Gibbs sampling [Porteous et al., 2008]. There are several variations to LDA for example clustering into hierarchies instead of a given number of topics, called (HLDA) [Blei, D.M. and Griffiths, T.L. and Jordan, M.I. and Tenenbaum, 2004].

One of the challenges with these probabilistic topic models is the bag of words assumption, in other words that no semantics are assigned to word order. Another challenge is that topics are not specific, they are interpreted by a user with foreknowledge of the topic. In establishing semantic relations specific items are required, therefore the word vector representation of a topic needs to be refined for this task.

Topic Themes and Signatures

Topic models developed for multi-document summarization (MDS) [Harabagiu and Lacatusu, 2005] present an approach that uses comprehensive semantic units, namely sentences, to define topics and modifiers for the topics. To follow the established nomenclature, the topics within sentences are called topic themes and the modifiers to the topics are called topic signatures. In this approach a seed is generated using the top results from LDA, then using part of speech patterns certain phrases are selected, these are used to generate the topic signatures. Further information about a topic is extracted based on phrases that contain topic signatures, this information is called topic theme. A summary example of topic signature and topic themes is presented below:

1. Given a document corpus, generate the topic models using LDA. Then, For each topic select the highest ranking verb and noun, e.g. *Pinochet-arrest*, these are the topic signatures.
2. Topic signatures can be enhanced with additional bigrams, by using part of speech tags on phrases that contain the topic signature, syntactic relations, such as Verb-Subject, Verb-Object, and Verb-Preposition can provide enhanced signatures, for example *charge-murder* and *react-Chile*

3. Topic themes are extracted by filtering information that is repeated in multiple phrases that contain the topic signature. To illustrate, a theme can be *in London* or *on Friday*, because this information is repeated for the topic signature.

This model provides a fine grained topic representation and it is possible to directly reference a topic in the corpus. One of the challenges of this it does not consider the temporal aspects of news topics. To deal with these challenge we propose a novel topic model called News Topic Reference (NTR) that will be presented in the following section.

2.2 News Topic References

The topic model developed for *Forest* is called News Topic References, these are the phrases used to refer to a news topic, for example in a Wikipedia article about a news topic such as “*The Arab Spring*”, the title of that article is a NTR. News topic references are often found in the title of news articles or in the initial part of the body where a summary is provided, for example the title “*After the Arab spring*” or in the body “...*As the Arab Spring remakes the...*”. We can also find multiple NTR that refer to the same news topic. For example the article titled “*Tunisia Effect*” in Wikipedia is a redirect page to the “*Arab Spring*”, we can assume that both “*Tunisia Effect*” and “*Arab Spring*” refer to the same news topic. A feature of NTRs is that they are coherent in a causal relation, for example in the title “*Arab Spring Causes Bankruptcy of Russia’s Top Fruit Importer*”, two NTRs are causally related “*Arab Spring*” and “*Bankruptcy of Russia’s Top Fruit Importer*”, this example shows how different news topics can be causally related. Because most of the information is extracted from online sources, there are limitations to the information the system can provide, namely information for poorly documented news topics and distinctions between fact and opinion.

A NTR can be defined as: The collection of surface forms, namely the phrases, used to define a seminal event or activities delimited by multiple features including creation date and referenced entities.

A work that supports the concept of word base or phrase based topic modeling include Story Link Detection Based on Event Words[Wang and Li, 2011] events are defined based on named entities and temporal data, they use this information to link different stories to one event. Their results indicate that they can achieve a better precision than a word vector representation of the topic, yet the recall is lower by comparison because stories may not present the same entities for a single event.

Another work that focuses on specific words to define event is Term Committee Based Event Identification within News Topics[Zhang et al., 2008], they use a word vector representation of the topic based on LDA,

but also include temporal data to define the events. Their approach is an iterative clustering of stories where there is a reevaluation the weights of the word vector in each iteration. By discarding the terms with a weight below a certain threshold they find a list of terms that describe events within a topic. The process can be described as clustering stories into topics and topics into events, then removing the event words to find different events. The approach provide results in the form of a word vector, this is not well suited for discovering causal relations as required in *Forest*.

Explicitly stated the originality of the proposed approach in contrast to similar works is that news topic references can represent more than one topic within a sentence based on a language model not derived from a document collection but from a set of known topic surface forms, namely known news topic references. This allows us to find semantic relations between topics at a sentence level. In following sections we will demonstrate how well news topic references perform for this narrowly defined task.

2.2.1 NTR Validation

To validate if a phase is an NTR three heuristics are proposed:

Heuristics 1 The seed phrase is the title for a Wikipedia article that is classified as an event. We classify a Wikipedia article as an event type article when it present several markers such as a time and place of happening, or the article contains section titles including timeline, aftermath, impacts, reaction, reponse; negative markers are also used such as title including the phrases “list of”, “disambiguation”, “band”, “film” or “country”. This heuristic is similar to topic indexing in [Medelyan and Witten, 2008] with the distinction that Wikipedia articles are selected to be regarding news topics or events. The work of [Medelyan and Witten, 2008] is discussed below.

Heuristics 2 The seed phrase is found with semantic markers, such as “The”, in the title of multiple news articles.

Heuristics 3 A collection of similar news articles is generated by using the seed phrase as a query. The search engine used may semantically enhance the query by normalizing the phrase, for example with stop word removal, word stemming and synonyms addition. The similarity metric used can be term vector based such as Jaccard’s Coefficient or KL divergence.

In Human-competitive automatic topic indexing (Maui), [Medelyan et al., 2009] Wikipedia is used to aid in naming topics. Several syntactic and semantic features are evaluated to select the phrase naming a topic, the best performing features are based on phrases used to links Wikipedia article, the evaluation shows that a combination of features in a bagged decision tree classifier performs at a level comparable to a human annotator. This approach is a source of motivation for *Forest*, yet the definition of a topic is more narrow in *Forest*, therefore the approach is modified.

A works worth mentioning because it aims to find phrases to represent a document cluster is Efficient Phrase-Based Document Similarity for Clustering [Chim et al., 2008]. This work is based on the word vector model, where the words are replaced with the leaves of a suffix tree, the suffix tree is built from phrases found in the document collection. Using this model phrases are used to calculate document similarity and thus name topics from the document cluster. The approach perform well for document clustering and selecting key phrases, yet results indicate that some modifications can be used to improve the selection process, for example phrase overlap that occurs in at least two different documents can be used to reduce the computational complexity with minimal affect on performance, we believe that this type of approach can be used to validate news topic references yet for the extraction process we believe that using semantic information from Wikipedia to train a machine learning approach, we can select phrases that are more informative while maintaining syntactic flexibility of the phrase.

2.2.2 NTR Aliases

There may be several phrases that refer to the same news topic, these NTRs are called aliases. To illustrate consider the phrase *"housing market crash"* this refers to a news topic, the phrase *"mortgage meltdown"* also refers to the same news topic, both phrases are NTRs. The phrases are considered aliases because they refer to the same news topic.

We propose two methods to discover NTR aliases: First, when the seed NTR is the title of a Wikipedia article classified as an event, then the titles of the redirect pages are considered NTR aliases. Second, an experimental approach is based on the assumption that alias NTRs are used as search terms with correlated patterns, this is to say an NTR will be used as a search term with a similar popularity distribution to an alias of that NTR. Thus, by analyzing query patterns, like in Google Correlate¹, potential NTR aliases can be discovered.

We considered using an approach similar to Measuring Semantic Similarity between Words Using Web Search Engines [Bollegala, 2007] to find NTR aliases, where word similarity is calculated using support vector machines(SVM). The features for the SVM are based on results from web search engines and pointwise mutual information, namely synonym word pairs are used to extract syntactic patterns surrounding the word pairs, these patterns are then evaluated using non-synonymous word pairs. The patterns are then given a probability of being indicators for synonyms and semantic word similarity is calculated based on this information. We found that semantically similar words are not always consistent with news topic

¹<http://www.google.com/trends/correlate> accessed 20.06.2012

reference aliases, this is to say that similar words do not indicate similar event references.

2.3 Topic Relations

In this section we review the works that organize the news by linking news items, for example in Google News stories are linked by having an entity in common. We classify works by the type of link that they generate between the news items, including term overlap and story coherence. We also classify the works by the granularity of the news item that is linked. A news item is an element of a news, for example a topic, and article or an event in the article. The news item granularity is the level at which a topic is defined for example in a story, in a passage or in a word vector.

In the previous section we presented how topics are represented and how multiple stories build topics. Therefore we will begin with story relation based on topic, this is how documents are clustered into topics. We assume that each story contains at least one topic and this topic is then used to related several stories based on topic, practically this is very similar to clustering based on document similarity, yet the similarity is based on the fact that the documents contain the same topic. thereafter we review topics that are temporally ordered, also known as a timeline, this is when a topic occurred before another. This type of temporal relations help user view the evolution of a topic, if there is one, yet topics may have temporal order without being related. We also review coherence relations, this is an objective observation over topics, it is how users perceive to understand the different relations between topics, for example how topics evolve over time or how topics affect one another. Finally, we focus on causal relations by giving a conceptual definition and providing an overview of approaches for extracting causal relations from natural language text.

2.3.1 Topic Similarity

When stories are related because they are about the same topic, we call this a generic link. For example the links generated in [Feng and Allan, 2009] are generic, the links are based on a term overlap, and semantic relations between the linked items are not established, we will give a review of [Feng and Allan, 2009] in the following sections.

A work that links similar topics at a paragraph level is Automatic generation of inter-passage links based on semantic similarity [Knoth et al., 2008], their approach first clusters at a story level using a term document matrix with TF-IDF weights and cosine distance as a similarity measure between stories. Then links are generated between the paragraphs of the stories in each cluster, where a maximum and minimum similarity is required as not

to link near identical paragraphs. The approach presents better results from the paragraph to paragraph links than the document to document links, and indicates that the increased granularity, from documents to paragraphs, provides a better insight into topical similarity.

In Mining Cross-document Relationships from Text[Knoth, 2011], the degree of similarity between texts is used to define the link type, the four types of similarities are tangent, expansion, equivalence or aggregate. In their implementation they focused on the document level, yet they noted the importance of using a finer granularity, such as paragraph or sentence level linking.

2.3.2 Temporal Relations

When topics or events are given in a chronological order we call the relation between items a temporal relation. In this section we review a recent work that uses a word vector to represent a topic, topics are clustered by similarity and linked to similar clusters from different time frames, thus generating a topic chain. In topic chains for understanding a news corpus[Kim and Oh, 2011] nine months of Korean online news are analyzed to find a latent structure and trends within the news. By analyzing the changes in a topic over different time periods the authors generate a topic chain that changes over time and often splits or joins different topics.

To explain the parts of the system we present an overview of to the steps below.

1. **Selecting the Stories:** A collection of 130K stories were selected from three Korean news sources with a wide range of topics and domains, such as politics and sports.
2. **Partitioning and Selecting:** To improve the performance of Latent Dirichlet Allocation (LDA) [Blei and Ng, 2003] for generating topic word vectors the stories are preprocessed. The words in the stories are stemmed and stop words are removed. The partitions for generating topics is based on a time window. To elaborate, the collection is first partitioned into time windows then the clusters are generated from within each partition, so if the time window is three days the difference between creation dates for each story in a cluster will be at most three days.
3. **Clustering Similar Incidents:** Six similarity metrics were used to cluster similar topics, including Jaccard's coefficient[Lee, 1999] and Kullback-Leibler divergence[Lee, 1999].
4. **Linking Clusters:** A link between clusters is generated between clusters from different time windows when the similarity between clusters is above a preset threshold. If there are no similar clusters in

consecutive time windows the following time windows are iteratively evaluated up to a set limit or when a link is generated.

The above algorithm shows three important variables, the similarity algorithm used, the linking threshold and the time window distance limit, also known as sliding window size. The six similarity metrics are compared in the evaluation and results show that a commonly used metrics, namely Cosine Distance, does not perform as well as others, and that Discounted Cumulative Gain (DCG)[Järvelin and Kekäläinen, 2000], which performs well for ranked results in information retrieval, does not perform well with LDA type results.

An empirical evaluation for variations in linking threshold and the time window distance was also presented. It shows how some of sporadic events are coherent, such as a famous actor’s romance, but others are useless because LDA generates incoherent topics. Results also show that longer event chains tend to be coherent and the evolution of a topic is apparent in these Chains. The longest topic chains often resembled domains such as the sections of a newspaper.

In *Forest* the issue of incoherent topics is addressed by using a more comprehensive topic model. One of the features that would be useful in this system, and is found in *Forest*, is the ability to directly reference the source story, to be able to drill into a single topic.

Another work that tracks changes in a topic is Discovering event evolution graphs from news corpora [Yang et al., 2009], this approach uses event time stamps and content similarity to detect changes within a topic. In their approach each story is manually annotated to a topic and a word vector representation of the document is generated, this vector and the temporal annotations are used to link topics where the vector similarity is above a predefined threshold and temporal rules are followed, results show how topics split and join over time. Though the approach is interesting the results do not compare well to human annotators in the evaluation.

In a more refined approach to track topic changes A sentence level probabilistic model for evolutionary theme pattern mining from news corpora [Liu et al., 2009], uses not only word vector similarity and temporal annotation, but also changes the granularity of a topic to a sentence level, where a topic sentence is composed of a named and entity and sentence keywords. Using sentence level tracking of a topic a summary is generated for the topic and given a set of stories clustered on a topic the resulting summary was shown to aid in topic understanding. These results indicates that sentence level analysis is valuable for understanding and that sentence selecting based on topic information is also useful, in their approach the topic models are similar to topic patterns where a topic is represented by a set of elements for example: entity, keyword and time. We believe news topic references can perform better than this topic representation.

In Discovering evolutionary theme patterns from text: an exploration of temporal text mining [Mei and Zhai, 2005], a stream of stories is partitioned into temporal intervals, where intervals can overlap, topics are extracted from these partitions using a probabilistic approach similar to LDA, using similarity measures such as Kullback-Leibler (KL) divergence, topics from different intervals are compared and if they are above a predefined threshold they are linked. They further propose to evaluate the topic strength based on a Hidden Markov Model, where the topic-word-vectors are the observed states, including one background state for the complete document collection. The output probabilities are given by the weights of the word vector for each topic, then the Viterbi algorithm is used to calculate the most probable state transitions, this probability is also used to calculate the topic strength. The linking of temporally segregated topics is very comparable to aforementioned approaches. Yet, the topic strength calculation provides an insight into the topics evolution, this type of information is not within the focus of *Forest* though.

2.3.3 Information Coherence

When the node type is a full story, as in [Shahaf and Guestrin, 2010], the selection and processing of a story is similar to a phrase level selection and processing, but because several stories are brought together there are more stories for the user to use to understand and the complexity is increased for the user. This is not a desired effect for a system to facilitate understanding.

In Connecting the dots between news articles [Shahaf and Guestrin, 2010], the goal of the approach is to help readers understand the news, to do this they propose to select a series of stories that show how different topics are connected. Based on a collection of stories the user selects the initial and final stories, the approach also required a predetermined number of stories the system will return. The initial and final stories represent topics that are to be linked, the topics are modeled as word vectors using a keyword extraction process, keywords are also extracted for all the stories in the corpus, stories are then selected and ordered based on the weight of the keywords and the sustained presence of the keyword in the sequence of stories. One valuable aspect of this work is the approach to evaluate the subjective information, their approach was to ask users to compare the results with all the predetermined number of stories and with omitted results, to find out if the results are valuable to the user. Several subjective features were evaluated including relevance, coherence and non-redundancy, where the approach performed comparable to commercially available systems.

2.3.4 Causal Relations

In this section we review the concept of causal relations for clarity, and we will review some of the methods used to extract causal relations. We classify the methods into three categories, methods that use the causal relations to extract the causal actors, methods that use the causal actors to extract the causal relations and hybrid methods that use both previous methods to contribute to each other.

Causal Semantic Relations

Causation is an important semantic relation, the notion of causation is very broad and used in many fields including linguistics, statistics and philosophy. The precise definition of causation can vary between these fields, there may also be confusion in interpreting causation due to correlation or conditional situations. While causation does imply correlation the reverse is not true, coincidence is an example of correlation without causation. Conditional situations can be distinguished from causation based on the temporal order of events, for example in the sentence "*if I travel then I move*" the movement can be perceived to be caused by traveling, although the movement can happen before the traveling.

There are different types of causes between causal actors. Three main types are listed below:

Necessary Causes The cause is necessary for the effect to exist, therefore the existence of the effect implies the cause, consider clouds and rain, clouds are necessary for rain to occur, although the cause does not imply the effect will occur, the presence of clouds does not mean it will rain, but that it may rain.

Sufficient Causes The cause may generate the effect but there may be other causes, therefore the cause implies the effect, consider rain and wet grass, rain will cause wet grass, but there may be several other reasons for wet grass.

Contributory Causes This type of cause may not be necessary or sufficient, but it may facilitate the effect therefore altering the cause will alter the effect, this type of cause does not imply the effect, it will only generate the effect in conjunction with other causes. To illustrate consider wet grass and slipping on the grass, wet grass is not a necessary or sufficient cause for slipping on the grass, yet wet grass in conjunction with other causes, for example unbalanced walking may generate the effect.

Contributory causes may be ambiguous, therefore they may not be well suited for facilitating comprehension in natural language. Causal phrases in natural language may be implicitly or explicitly stated. Phrases that

contain unambiguous causal patterns are considered explicitly stated, for example “Alice was running because Bob was chasing” uses the pattern [Noun phrase 1] because [Noun phrase 2], it follows that it is an explicit causal phrase. Phrases may also contain implicit causal statements, where entailment is required to deduce the causal relation, for example “Bob was chasing so Alice ran away” in this case the phrase can be classified as causal although it is not explicitly stated. To maintain clarity in the results and facilitate processing *Forest* focuses on extracting explicit causal statements and implicit causal statements may be used as supporting material.

Causality to Extract Causal Actors

A direct approach to detect and extract causal phrases is by using patterns and causal keywords such as “because”, several causal taxonomies [Wolff et al., 2002, Joshi et al., 2010] can be used to define the causal keywords and the patterns, these patterns can also be hand-crafted or extracted from different sources of information. Some implementations [Radinsky, 2011] use part of speech taggers to add features to the patterns. One of the main challenges of this type of approach is domain over-fitting. The patterns and causal keywords often vary between domains therefore the performance fluctuates greatly.

Causal Actors to Extract Causality

Some causal phrases do not contain causal keywords, such as implicit causal phrases, therefore a different method needs to be used to extract causal relations. Based on the premise that a phrase that contains a cause and its effect is a causal phrase, causal phrases can be detected based on the causal actors [Saito et al., 2007, Perrin et al., 2008]. The causal actors can be noun or verb phrases that have been extracted from the corpus in a preprocessing step. To illustrate consider the verb pair “run-chase”, a phrase that contains this verb pair will likely be a causal phrase [Riaz and Girju, 2010], another example may be the noun pair “HIV-AIDS”, a phrase containing these nouns in separate clauses will likely be a causal phrase [Beamer and Girju, 2009]. One of the challenges of this approach is that the results are seldom explicitly causal, therefore they may not be unambiguously causal to the user.

In Using a bigram event model to predict causal potential [Beamer and Girju, 2009], the authors use temporal order in screenplays to extract causal word pairs. Screenplays are chosen due to the nature of sequential events in screenplay writing. The texts are annotated with causal relations that are represented by verb pairs, the frequency of verb pairs in a causal relation is compared to the frequency of the verb pairs in a non-causal sentences, this generates a causal probability for the verb pair, also called bigram causal potential. Results indicate that only highly probable causal

verb pairs perform well for causal relation extraction, this supports the use of explicit causal relations in *Forest*.

Hybrids

Consider in this case an approach that uses the causal keyword and pattern approach to extract cause-effect pairs, then uses the cause-effect pairs, called causal actors, to extract new causal keywords and patterns. This approach is cyclical and can be bootstrapped with cause-effect pairs or the causal keywords and patterns approach. The main challenge of these approaches is that errors are propagated within the bootstrap cycle, therefore the performance varies dramatically. Another challenge is defining the limits to the bootstrapping cycle, this is a “multi-armed bandit” problem. There are also systems that extract multiple types of semantic relations[Butnariu et al., 2010, Chan, 2011] including causal relations, the performance of the system is comparable to the approaches presented above, yet the processing complexity is considerably larger.

In Learning Textual Graph Patterns to Detect Causal Event Relations[Rink et al., 2010] the authors aims to extract implicit causal relations using a parse tree with part of speech tags. Aggregated Parse trees are generated for causal sentences, these trees give a generic syntactic version of the sentence. Then pattern frequency is used to select causal patterns. To classify sentences as causal, the test sentence is matched to one or more of the extracted patterns and represented as a binary vector where each element of the vector is a pattern match. The vector is used in a Support vector machine, that has been trained on the causal sentence. The approach did not perform well, yet an error analysis indicate that the use of cue phrases would improve the approach. In *Forest* we will show a machine learning approach that uses cue phrase to classify causal sentences.

Another work that approaches causal relation extraction based on a machine learning is Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision[Riaz and Girju, 2010], this work deals with news stories and highlights the subjectiveness of the concept of causality. There are three stages to the approach. First, topics are modeled from a document collection, specifically Pachinko Allocation is used. Pachinko Allocation is similar to LDA, yet it includes not only relations between topics and words, it also includes relations between topics. Second, high relevance topic words are used to select candidate sentences from the corpus, these sentences are evaluated for the pattern: subject-verb-object, using a semantic role labeler. Third, the sentences matching the pattern are then clustered based on the verb in the pattern, the cluster of sentences is used to represent an event. Thereafter, event pairs are extracted from the corpus to be evaluated by association rules, the event pairs must occur together, this is in the same document, at least five times to be considered related. Causality is based on the assumption

that causal events have a temporal order and that temporal order can be obtained from the documents. Another assumption is that a causal event will appear independently of effect event in the news stories, while the effect is highly likely to appear with the cause. One more assumption is that the closer the events are in proximity, the more likely they are causal. The best accuracy achieved by the approach is 72.5%, yet the approach seems to highlight correlated events rather than causal, the author also notes the challenges in inter-annotator agreement of causal relations, and the usefulness of world knowledge for the task of identifying causal relations at a discourse level.

In following chapters we present our approach to causal relation extraction. Explicitly stated the originality of the proposed approach in contrast to similar works is the use of a machine learning to disambiguate explicitly causal sentences. Where other approaches focus on disambiguating implicitly causal sentences or classifying causal sentences using causal markers and part of speech patterns. We disambiguate sentences with causal markers, this is because explicit causal sentences are easier for users to understand, which is a goal of *Forest*. We select sentences based on the explicit causal markers then classify these sentences, that all contain a causal marker, as causal or non-causal with a machine learning approach.

Performance Factors

The performance of these systems is highly dependent on the field in which they are applied [Ittoo and Bouma, 2011]. There is the challenge of cross-domain over fitting due mainly to the differences in language models of the corpus. Another challenge is ambiguity in validating implicit causation. Systems to extract causal relations in a specialized domain can achieve a precision of 75% and a recall of 90%. Implementations that are used with language models that vary have an accuracy of 25% to 75% and the recall is often undefined due to variations in the definition of causation. More details about the work [Ittoo and Bouma, 2011] is given below.

In Extracting explicit and implicit causal relations from sparse, domain-specific texts [Ittoo and Bouma, 2011], the authors develop a minimally supervised bootstrap approach, where causal verbs are used to find causally related noun in Wikipedia, these noun pairs are then used to discover additional causal phrases, results show this straightforward approach performs well for causal relation extraction. The approach uses pointwise mutual information between causal nouns and causal phrases to asses the ambiguity of the extracted causal information, or causal purity as they call it. Using the least ambiguous extracted causal information they iterate using causal phrases to extract causal nouns and vice versa until an annotator decides that the results are no longer valid. The causal phrases extracted from Wikipedia are similar to those extracted by other authors such as Girju [Girju, R., Moldovan, 2002].

Another bootstrap approach is text mining for causal relations[Girju, R., Moldovan, 2002], where they use explicit causal phrases to find causal actors, the actors are then used to discover novel causal phrases from a document corpus, the novel phrases are manually classified as causal. With this approach they were able to find a set of causal phrases and evaluate the frequency and ambiguity of the phrases, this is to say that they classified causal phrases based on how often they are used and how often they refer to a causal relation when they are used. In comparison to human annotators this approach achieved an accuracy of 65.6%. In *Forest* we develop an approach based on this by using a subset of the discovered causal phrases.

2.4 News Text Processing

There are several works that should be mentioned that are in the area of study, as a reference and guide to the state-of-the-art. In the following section we review some of these works from the fields of document summarization, keyword extraction, named entity recognition and event extraction.

2.4.1 Document Summarization

There are several approaches to document summarization, many relevant approaches are listed in [Das and Martins, 2007]. Most of these works can be categorized into Multi- document summarization, such as MEAD[Radev et al., 2004] and Lex Rank [Radev et al., 2004]; and single document summarization, such as the well known The Automatic Creation of Literature Abstracts[Luhn, 1958] and more recently A financial news summarization system based on lexical cohesion[Oliveira et al., 2002]. Many of these approaches have been implemented by their originators and are available online. Below we mention some works that deal with summarizing news particularly at a sentence level.

In Storyline-based summarization for news topic retrospection[Lin and Liang, 2008] they use neural networks to cluster documents, the neural network uses the word vector model of the documents. To cluster the events they use growing hierarchical self organizing maps (HSOM), this type of neural network uses word vectors to represent documents and averages the vectors in a cluster to represent an event. The system architecture indicates that the document corpus is generated based on a single topic. Using the event vectors, paragraphs are selected to represent an event, similar but different events are selected for following paragraphs to generate a summary of the topic. The end result is sequence of paragraphs chosen based on word vectors, to provide a summary of a topic.

A work with sentence level temporal relations is temporal summaries of new topics[Allan et al., 2001], where a collection of documents is partitioned into sentences, and the sentence are grouped and ordered by time and placement in the document. Each time/place group is evaluated to find keywords and those keywords are used to select novel words in following time/place groups. Keywords and novel words are used to select and cluster sentences to generate a summary. Another work that uses keywords to generate summaries is [Wan, 2007] where sentences and words are selected based on three graphs the sentence to sentence graph, the word to word graph and the word to sentence graph. The links in each graph are computed differently, between sentences they are generated based on content similarity, for words is based on semantic information or knowledge bases. To link words to sentences a relevance score is given to the words, similar to Term Frequency Inverse Document Frequency (TF-IDF, explained below). An iterative process follows to select the keyphrases and generate the summary, in essence it selects salient sentences based on the sentence to sentence graph and the salient sentences are used to select salient words. Thereafter, salient words are used to select salient sentences. This repeats until there is a near-convergence on the results.

TF-IDF is a method to assign a numerical value to the importance or a word in a document collection, it is generally used in various information retrieval and text mining tasks. To calculate the number of times a term appears in a document is multiplied by the fraction of documents that contain that term, in this way a term that is often used in a single document increases in value and if the term is used in many documents it decreases in value.

2.4.2 Keyword Extraction

As we have seen in the works of document summarization, keyword extraction is a key components in several applications.

Previously we mentioned Maui[Medelyan et al., 2009] a work that that does controlled keyphrase indexing, this means that the tags or keyphrases for a document are from a limited vocabulary. On the other hand works such as Domain-specific keyphrase extraction[Frank et al., 1999] perform free indexing, such that the keyphrase can be freely chosen. These and other approaches are freely available online².

An approach that performed comparable to [Medelyan et al., 2009] is available in the Palladian toolkit³, called ControlledTagger[Katz, 2010]. This approach, as many others, are based on TF-IDF to select relevant terms. The approach follows the bag-of-words assumption that the order of the words does not affect the results. The words in the TF-IDF vector are

²<http://www.nzdl.org/Kea/index.html> accessed 1.1.2013

³<http://palladian.ws/> accessed 1.1.2013

process by stemming and removing stop words. To improve the results the terms are re-ranked based on the prior probabilities of the word occurring in the document, this way more popular terms are prioritized. They also correlate terms, so that terms that often appear together are considered in conjunction for the results.

Another relevant approach is called RAKE, Rapid automatic keyword extraction. The approach is described in [Berry, Michael W and Kogan, 2010], this approach uses stopwords and punctuations to partition the document into phrases, they include stop words if they are repeated within a phrase at least twice. Then a term co-occurrence matrix is developed for the words in the phrases. This matrix is used to calculate words ranking using the frequency of the word and the length of the phrase where the word is found. Keywords are selected based on the ranking such that longer phrases with words that are more frequent are prioritized.

2.4.3 Named Entity Recognition

There are several approaches[Nadeau and Sekine, 2007] and utilities for the task of named entity recognition^{4 5 6}. Though there are several organizations dealing with this task including TREC 2010[Balog et al., 2010], Message Understanding Conference (MUC 7)[Chinchor and Robinson, 1997] and CoNLL 2003 [Sang and Meulder, 2003], there is no consensus on the definition of an entity[Balog et al., 2010, Rössler, 2007]. Despite these challenges state-of-the-art named entity recognition approaches for English text perform at a level comparable to human annotators, for example the best performing system in MUC 7 achieved a 93.39% F_1 measure while human annotators achieved a 97.67% F_1 score. The F_1 measure along with precision and recall are metrics often used in information retrieval systems[Manning and Raghavan, 2009].

2.4.4 Event Extraction

This is the task of obtaining the key facts of an event, these typically include answering the 5W1H questions[Carmagnola, 2008]: Who, What, When, Where, Why and How; for a specific event. Events extraction can also be more specific, for example for disease outbreak extraction[Grishman et al., 2002] and conflict events[King and Lowe, 2003, Atkinson et al., 2008, Tanev, Hristo and Piskorski, Jakub and Atkinson, 2008] the number of victims and cost of the event may also be answered.

⁴<http://alias-i.com/lingpipe> accessed 1.1.2013

⁵<http://www.alchemyapi.com/api/entity-extraction> accessed 1.1.2013

⁶<http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions> accessed 1.1.2013

The Event Detection and Recognition (VDR) task in the ACE 2007 conference provided a set of templates to represent events, and the task was to fill these templates from natural language text found in news stories. The only participant (BBN technologies) obtained score of 13.4% in this task, this is an indicator of the difficulty of the task.

There are three main approaches to extract event information: the pattern based approach where events are obtained using predefined patterns such as “in an attack, *NUMBER* people were killed in *LOCATION* on *DATE*” to retrieve some of the values to fill an event template(*NUMBER*, *LOCATION* and *DATE*). This approach presents several challenges including manually developing the patterns and handling the trade-off between the precision and recall of the pattern, the more specific a pattern is the higher the precision but lower recall.

Another approach is based on event summarization, where a collection of topically related documents is statistically analyzed to retrieve keyphrases[Ji and Grishman, 2008, Filatova and Hatzivassiloglou, 2004, Li et al., 2006, Liu et al., 2007] and use these phrases to generate the event[Naughton et al., 2008]. Other works that generate event summaries are News in Essence which uses MEAD[Radev et al., 2004], the Columbia Newsblaster[Evans et al., 2004] and GISTexter[Harabagiu et al., 2002]. One of the challenges of this approach is that it fails at a finer granularity, the summary may contain relevant information but may lack cohesion and semantics.

A More recent approach is based on semantic role labeling[Yaman et al., 2009, Surdeanu and Harabagiu, 2003], where question answering systems are used to extract the 5W1H information. Semantic role labelers assign tags to words such as actor, patient, source and sender. This additional information allows clustering of event mentions on semantic representations of a phrase[McCracken et al., 2006] instead of syntactic representation.

Other systems focused on extracting the 5W1H information include News Cluster Event Extraction Using Language Structures (NEXUS)[Piskorski et al., 2007], where the focus is extraction of events that are either violent incidents or security-related events.

Another work that uses semantic role labelers is The Chinese News Fact Extractor (CNFE)[Wang et al., 2010]. They argue that semantic role labeling is computationally expensive and it does not scale to large news corpora, but still use it to semantically tag the text such as the actor *WHO* and the patient *WHOM* of a sentence. Consider the following to provide an intuition of the approach to extract the 5W1H information: *WHO/WHOM* is answered using patterns and named entity recognition, *WHAT* is answered using the highest ranking verb keyword, *WHERE* and *WHEN* are answered using named entity recognition and *HOW* is answered by combining the *WHO/WHOM* and *WHAT* results into a sentence “*WHO did WHAT to WHOM*”.

NewsX[Wunderwald, 2011], uses a step-by-step extraction approach based

on the concept of deferred commitment[Yangarber, 2006] to answer the 5W1H questions. In this approach candidate answers to each questions are considered closely coupled to each other. Candidates are extracted using patterns and named entity recognition, ranking is based on a combination of candidates, for example candidates for *WHO* are used to rank candidates for *What*. Their results indicate that there is low inter-annotator agreement for *WHY* and *HOW*, with Fleiss Kappa[Fleiss, 2003] scores of 0.2 and 0.25 respectively, these questions also had the highest proportions of partially correct or incorrect extractions.

2.5 News Understanding

In this section we review works that have goals similar to those of *Forest*, namely works that relate news items to facilitate understanding. We will show how news topics are defined at different levels, this is to say that the links between news topics are established at a document collections level, a document level, a paragraph level or at a sentence level. And we will show that different types of relations are established, including temporal relations and causal relations. All with the underlined goal of reducing information overload or improving understanding.

We begin with Incident Threading for News Passages[Feng and Allan, 2009] by Ao Feng and James Allen, the latter of which wrote the book on Topic Detection and Tracking [Allan, 2002]. In this work the authors address the challenge of multiple topics within a single story, by analyzing the story at a phrase, or passage ,level. Passages are clustered into events, called incidents, and then linked to establish a thread of events. Different threads are intertwined at common incidents to generate a network of events. The goal of this network is to provide a solution for “an increasing need for automatic techniques to analyze and present news to a general reader in a meaningful and efficient manner”, in other words to reduce news information overload for the user.

The method used to generate the network can be summarized as follows:

1. **Selecting the Stories:** In a preprocessing step, a collection of stories about a single topic is selected. The stories are manually selected from online sources. To simulate real world circumstances, namely imperfect topic clustering, a few stories from a different topic are added.
2. **Partitioning and Selecting:** Stories are parsed into paragraphs and classified, this reduces the amount of text to be processed, because only the passage is processed instead of the complete document. In this case classification was used to select only violent passages, called incidents. Several methods are used for the classification including Boostexter[Schapire, 2000], Support Vector Machines and Maximum

Clustering precision	0.14
Clustering recall	0.14
Link precision	0.03
Link recall	0.15

Table 2.1: Performance results of Incident Threading for News Passages

Entropy, using features such as violent action words and uncertainty terms.

3. **Clustering Similar Incidents:** The clustering is based on term overlap, entity coreference, and temporal metadata, the latter is also used to give a direction to the link.
4. **Linking Clusters:** The methods used for clustering are also used for linking, although the thresholds are set at different levels.

In these four steps we can see the main components of other related works, sometimes in different order and with variations in the methods, but ultimately the same steps. In the evaluation we can also find one of the main challenges of systems that aid understanding, namely how the evaluation improved understanding.

In Incident Threading for News Passages the evaluation is focused on clustering and linking performance, there is also a small evaluation of the usefulness of the system. A baseline system is provided as a comparison as there are no similar systems to compare to.

A summary result table is provided in table 2.1. This summary shows an F-Measure of 14% for clustering and 5% for linking. A small user study was performed to evaluate the usability. In the study users are given a set of questions regarding a topic, a fraction of users are given the incident network and the rest are given the original stories. Results show an improvement of approximately 25% in user understanding.

The link type in [Feng and Allan, 2009] is generic, this is to say that the link is based on common terms from different passages. In contrast *Forest* generates links that are explicitly causal. The node type in [Feng and Allan, 2009] is an incident, or passage, this is similar to *Forest* although, in *Forest* nodes are generated at a topic level and not at an event level.

A work that is close to *Forest* and is based on the work of [Feng and Allan, 2009], is causal network construction to support understanding of news [Ishii et al., 2010], as the name implies causal relations are established between news items, in this case the news item is a weighted word vector to represent a news topic event. This work is very similar in scope and goals as *Forest*, yet the methodology is considerably different and the results are only partially comparable. One interesting aspect of this system is that it not only clusters similar incidents, it also eliminates unimportant causal relations, this in order to reduce the complexity of the final network.

Causal relation extraction is based on “clue phrases” in Japanese, such as “Tame” (translated: for the sake of). The topic event vector is generated by extracting the keywords from the causal phrases or the title of the story, the keywords are obtained using a Japanese dependency structure analyzer and a case-frame dictionary. The weights in the word vector are based on the frequency of the words in the story, explicitly the more frequent the word the higher the weight. By clustering the vectors, based on cosine distance, weights are combined. Then the weights of linked clusters are used to determine the importance of the link.

It is evident that the system is heavily dependent on finding keywords to represent the topic well, as we have seen in the topic representation section, a vector representation is not optimal for clarity.

The evaluation was focused on extracting the topic from the story, specifically evaluating the keywords as being representative of the topic. Using only keywords extracted from the causal phrase an accuracy⁷ of 27% was reached, with the addition of keywords extracted from the title of the story the accuracy rose to 60%. The evaluation of topic clustering was done at different thresholds of cosine similarity. The results show that the accuracy is inversely correlated to the recall of the system and that the order of clustering topics affects the results. The F-measure for clustering ranged from 21 to 37%. The accuracy of causal relation extraction was not evaluated.

The main differences between [Ishii et al., 2010] and *Forest* is the language and the representation of the topic, in [Ishii et al., 2010] the topic is represented by a weighted word vector, in *Forest* a topic is represented by a semantically coherent phrase. As we will see in further sections the vector representation is prominent and very useful in many cases.

It is evident from the above works that the representations of topic or event can be very different. These different node types have advantages and drawbacks, in this section we will review some of these advantages and drawbacks.

In [Feng and Allan, 2009] a node in the resulting network is an event, generated based on one or more passages. A passage is a phrase or sentence that has been previously classified, in this case phrases must be classified as violent to be passages. One of the advantages of classification is the reduction of processing complexity, thus the performance of the system can improve. Passages generally offer a cohesive semantic unit, this is to say a passage is in itself complete and clear. This is an advantage for systems that intend to have coherence and comprehension in their results. One of the main challenges of using passages is that analysis is done at word level and it is possible to lose the semantics of the passage at this level. Another challenge is the processing complexity, while a document collection

⁷Defined by the author as number of correct topics divided by number of all extracted topics

may have a word dictionary length of 3000, the potential combinations of these words to generate a phrase have an exponential growth, and thus a phrase dictionary may be highly complex to process. The potential for high complexity is one of the reasons phrase level processing is often done on few selected phrases from the full corpus.

When the node type is a full story, as in Connecting the dots between news articles[Shahaf and Guestrin, 2010], the selection and processing is similar to a phrase level, and stories offer a cohesive semantic unit, but because several stories are brought together the complexity increased for the user. This is not a desired effect for a system to facilitate understanding.

Conclusions

In this chapter we have review several types of works closely related to *Forest*, we obtained a theoretical background as well as several definitions from Topic Detection and Tracking (TDT), this gave us a frame of reference to address topic modeling, we say how the vector space model evolved to probabilistic models and how how these models are not fine grained, for a sentence level topic modeling we saw works from the multi document summarization field and how these are still not well suited for the requirements of *Forest*, this motivated the development of news topic references, an approach developed for *Forest*. We also reviewed several works that link topics in different forms, including similarity and temporal linking, yet we focused on causal relations and presented two main types of approaches one that use causal phrases to find the causal participants, and one that uses the causal participants to find the causal phrases. Although we also mentioned other types of approaches, for the task defined in *Forest* we choose a causal phrase approach that deals with sentences that contain a causal phrase yet are not causal. We also presented several works in the field of text processing, including named entity recognition, keyphrase extraction, document summarization and event extraction to give an overview of the current state-of-the-art. Finally we reviewed works that have similar goal to *Forest*, namely helping users understand the news.

Chapter 3

Conceptual Architecture

In this chapter we present the conceptual architecture of *Forest*. First we present a general overview of the modules that make *Forest*. This overview provides an intuition of how some non-functional requirements are addressed and how the different modules are interconnected. Then, we present a process flow to illustrate the components used for extracting causal relations between news topics. Finally, we review each of the modules individually. This chapter provides a basis for further chapters where we give an in-depth review of the central components of *Forest*.

3.1 General Overview

In Figure 3.1, we present the high level architecture of *Forest* as an FMC¹ diagram. In further sections we will discuss the functionality of each module and how the modules will be modified for different tasks. The diagram depicts active components with sharp corners, active components are those that may alter the information. Passive components, such as storage, are depicted with rounded edges. Bidirectional communication is shown as a line with a circle, and curved arrows indicate read-write access. The Figure 3.1 shows the seven main modules of *Forest*, below we give a brief explanation for each module:

Presentation this module is responsible for presenting the information in a format that is suitable for the user.

Processing Components is the module that contains the processing steps to extract the information, for example the causal-relation-extraction component and the news-topic-reference-detection component. This component is tightly linked to the task manager.

¹<http://www.fmc-modeling.org/> accessed 6.6.2012

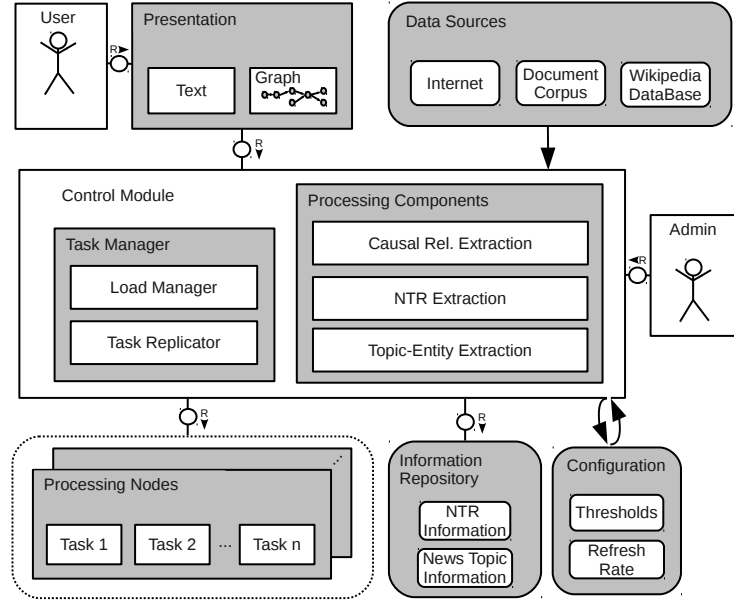


Figure 3.1: Conceptual Architecture

Task Manager uses the modularity of the processing components and features in the extracted information to improve performance, for example by executing process components in parallel. This module is responsible for defining the processing load distribution in order to maintain system performance. Together with the processing components, it makes the control module.

Processing Nodes represent one or more computers where the processing components are executed.

Data Sources is the access point to the distributed information sources, namely web sources. These sources may also be a fixed collection, fixed sources are useful for repeating tests and comparing the performance of different approaches on the same data.

Information Repository is where the extracted information is stored. Some of the stored information can be reused, therefore alleviating processing requirements.

Configuration is for the system to function optimally in different settings, for example to evaluate components or to set load capacities. User configuration may also be stored in this module.

3.2 Processing Components

While there are several components within the processing components module, only three are presented as examples. The components presented in this section are the functional components of the system. We present the different components, in Figure 3.2, as a process flow to provide an intuition of the functionality and to show how the components interact. In Figure 3.4 we show the process flow with intermediate results, in order to provide a more complete overview of each step in the process flow. In further chapters we will review the performance of two main components, namely causal relation extraction and also topic modeling with news topic references(NTRs). Several aspects are considered in the creation of the system architecture: in addition to the main focus of generating a network of causally related news topics, we also consider the systems modularity as a non-functional requirement, component modularity allows tasks to distributed to different processing nodes, thus minimizing processing latency. The process flow in Figure 3.2 does not reflect this modularity, instead it provides an intuition of how the system extracts information.

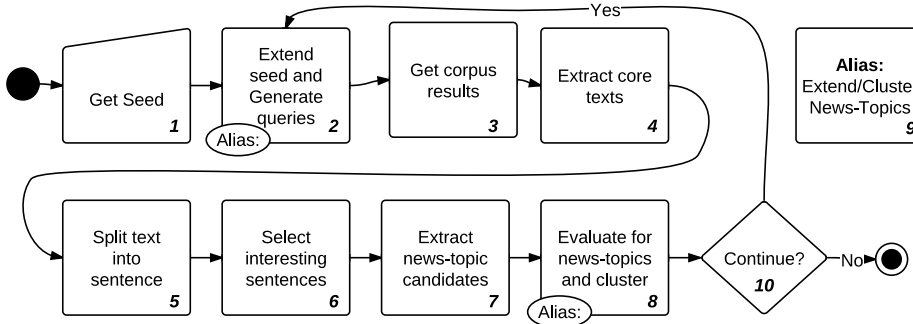


Figure 3.2: Forest process flow

3.2.1 Processing Steps and Intermediate Results

In this section we describe each component including the intermediate output that is produced. Below we present each step of the process flow in Figure 3.2. It is evident that these steps are part of the components mentioned in Figure 3.1. Before we present each component in the process flow, we note the distinction between modules and components:

Components are the low level processes that can be distributed and replicated.

Modules are the high level divisions of the system. They administrate the resources used by the components.

Step 1 The process begins by getting a seed news topic reference(NTR) from the user, this will be the first node of the causal network. There is a validation process in this step because the seed phrase given must be a NTR for the system to function. The validation is not only to verify that it is a NTR, some spelling or grammar corrections may also be done. The output of this step is a phrase that has been validated as a NTR. An in-depth analysis of NTR validation is given in the following chapters.

Step 2 The initial seed is extended with causal markers such as “caused” and “led to” in a method similar to Hearst’s in [Girju, R., Moldovan, 2002] to generate a query for retrieving relevant documents. The query is also extended with NTR aliases when they are available. An in-depth analysis of NTR aliases is given in the following chapters. The output for this step is a query that is extended with causal markers and expanded with aliases. The formulation of the query is intended to maximize the amount of relevant documents returned from the search engine.

Step 3 The extended query is used to retrieve relevant documents through the search engine. The top ranked results are selected for further processing. The main issue is to have up-to-date information available for processing, and to have the capability of distributing the processing load. An online search service, such as Google search², is preferred because these services are specialized in maintaining an up-to-date index, and the local processing requirements are lowered. For evaluation purposes, it is important to have a local index to be able to replicate experiments and compare different processing strategies. The output of this step is a collection of documents that may contain a causal relation about the initial seed NTR.

Step 4 The next step is to obtain relevant information from the top ranked results, this includes the human readable text and metadata including creation time and source information. The metadata information can later be used to refine the results in the causal network, for example by providing causal relations that are within a certain time period, or by providing information from selected sources. The output of this step is a short text linked to additional information, including the natural language text extracted from news articles, source information, author information, generation time-stamp and extraction time-stamp.

²<http://www.google.com/> accessed 6.6.2012

Step 5 The natural language text is split into sentences, this facilitates processing the results and allows summary results to be presented. This step reflects the assumptions that a single sentence contains causal relations between news topics. The output of this step is a collection of sentences linked to their respective metadata.

Step 6 This step is a preliminary selection process, to reduce processing requirements. The selected sentences must contain causal markers and they should also contain the known NTR or an alias of that NTR. The output of this step is a reduced collection of sentences linked to their respective metadata.

Step 7 In this step causal relations are extracted from natural language text. An in-depth analysis of causal relation extraction is given in the following chapters. In this step first we confirm there is a causal relation stated in the sentence, then if there is a causal relation, we extract the causal actors, namely the cause and the effect stated in the sentence. Afterwards we verify that one of the causal actors is the known NTR or an alias of that NTR. If one causal actor is the seed NTR or an alias then the counterpart is considered a NTR candidate. In this step we also register the direction of the relation, this is to say we register it as the cause or the effect of the seed NTR. If the sentence does not have a causal relation it is discarded. The sentence is also discarded if the seed NTR, or an alias of that NTR, is not a causal actor. The output of this step is a candidate NTR and the information extracted from the sentence, namely the causal relation with direction and relevant metadata.

The performance of causal relation extraction will be enhanced by analyzing the NTR causal actors. NTR features such as temporal order and part-whole relations can be used to improve causal relation extraction.

Step 8 The NTR candidates are evaluated using a validation similar to the one done in Step 1. Initially the sentence segment that is a candidate NTR is validated, if it does not pass the validation then the phrase is evaluated to find mentions of NTRs within the text, for example in the phrase “the recent conflict in the middle east known as *the Arab spring*”, we find the NTR “*The Arab Spring*”. The NTRs included in additional text should be extracted from the additional text and the complete information should be presented to the user for verification upon request. If a NTR is extracted it is compared to all other extracted NTRs to find aliases, if two NTRs are aliases they are clustered and only the more frequent NTR is given in the final presentation of the results. This is to simplify the results and make them more understandable for the user. The frequency information will also be used to rank different causes for the seed NTR. An in-depth analysis of NTR validation and NTR aliases is given in the

following chapters. The output of this step is one or more NTRs linked to their corresponding metadata including source sentence, source document and causal relation.

In the following chapter we will review how causal relations can be used to establish NTRs for very recent news topics. Using causal relations to establish new NTR is based on several assumptions; one of the assumptions is that the cause for a NTR has already passed therefore it may already be in a knowledge base, such as Wikipedia, while the effect may be a new NTR, not yet defined in the knowledge base but still referenced in online news sources.

Step 9 This step is complimentary to step 2 and step 9. In this step the NTR phrases are mapped to conceptual news topics. There may be several phrases that refer to a single news topics. This mapping is used when the user provides the initial seed NTR, this seed NTR is mapped to a news topic, then the query is expanded using all the phrases referring to that news topic. Each different phrase that refers to the same news topic is called an alias. Aliases are also used to reduce the result set, as stated in step 8. The output of this step is a mapping from NTR phrases to conceptual news topics. In some cases several news topics may be part of one another, this information is registered in the mapping when it is available.

Step 10 At the end of the extraction process we have a seed NTR and causally related NTRs, and the associated metadata. The extracted NTRs can be used to extend the causal network, this is to say the extracted NTRs can be used as seed NTRs for a new level in the causal network. At this step we can output the current level of the causal network, specifically the seed NTR and its causes and/or effects. This information can be presented as a very short text with the NTRs, or more coherent information as a sentence, or more complete information as a link to the information source. Additional information such as ranking of causal relations and aliases for NTRs should also be presented to make the information more complete.

The visualization in Figure 3.3 shows how an extracted NTR can be used as a news seed to extend the causal network. In the visualization we can see an initial seed for which a cause is extracted. The extracted cause is then used as a seed for another iteration of the system, thus extending the causal network.

Intermediate Results In Figure 3.4 we can see the process flow with intermediate results for each step. Because the components are modular and mostly independent it is possible to replicate and distribute the components as needed, we will show how this is done with the process control component. Ultimately the input to the system is a phrase that

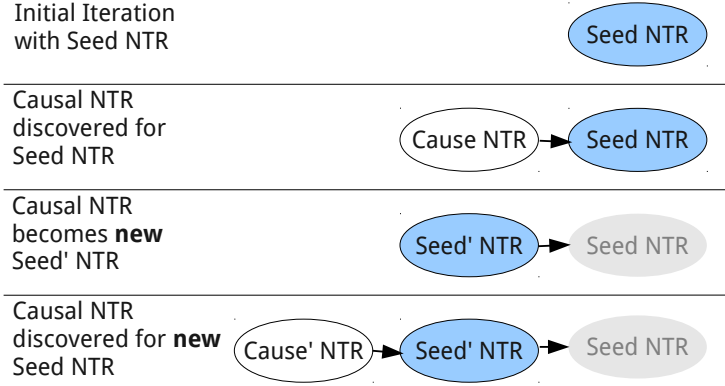


Figure 3.3: Example of reusing extracted information

is a seed NTR and the final result is a collection of phrases, along with the corresponding meta-data, that represent the causes or effects of the initial seed NTR. The resulting phrases are NTRs along with meta-data that includes source sentence, source story, story creation date, extraction creation date, and other information.

3.3 Additional Modules

Now that we have shown the main functionality of the system, we present the supporting components that help provide timely and user-friendly results. The eight support components are: Presentation, Control Module, Task Manager, Processing Components, Processing Nodes, Data Source, Information Repository, and Configuration.

3.3.1 Presentation

This component is the interface between the system and the user. The user can provide the initial seed NTR via this component. The system may also provide suggestions for the seed NTR based on prior extractions or current news. This component is also responsible for the presentation of the causal network. There may be several presentation formats for example a streaming JSON file or a graph visualization. Presenting the information in a visual format is useful to make the information easier to understand. Facilitating understanding is one of the main goals of this system.

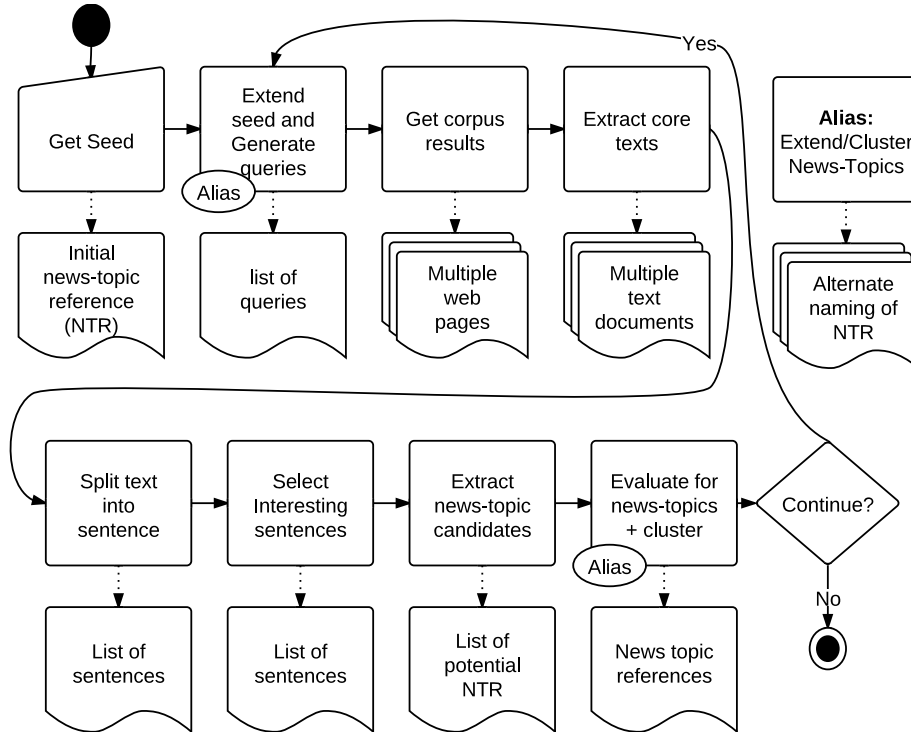


Figure 3.4: Process flow with result types

3.3.2 Control Module

This module is the main interface to the administrator of the system. The module is composed of the task manager and the processing components, this is because the functionality of these submodules is highly interdependent although considerably different.

3.3.3 Task Manager

Because of the modular design of multiple processing components it is possible to distribute these tasks to different processing nodes. The intention of distributing the processing load is to maintain a timely performance of the system. Depending on the processing load this component may choose to update recently extracted information or it may chose to present previously extracted information, therefore this component is responsible for the information freshness, this is to say that the information is up-to-date. Because many of the processing components enable parallelization, this module may assign the same task to multiple processing nodes, and once

the results are generated they can be consolidated. To illustrate consider the process of extracting causal relations from a collection of texts, the collection can be partitioned to be processed by several nodes. The results of the extraction can then be combined.

3.3.4 Processing Components

In the previous section we presented all the processing components in a flow format. The main components of this flow are the causal relation extraction and the news topic reference validation and extraction. There are additional components that are necessary to provide complete results and also to evaluate the hypotheses. This module is the repository of the methods and logic of each component. Each component can be deployed to different processing nodes to complete a task.

3.3.5 Processing Nodes

This module represents the available hardware resources to process the information. These resources are registered with the control module. This module addresses the need for resources to maintain scalability and near real-time performance of the system.

3.3.6 Data Sources

The processing components of *Forest* are used to extract information, therefore it is critical to have a clearly defined source of information from which the information is extracted, this module provides access to the sources of information. To evaluate the system it is important to replicate inputs to compare results of different processing strategies, in this case a static document corpus is appropriate as a source of information. To obtain the most up-to-date information an Internet source is well suited. This component is also responsible for filtering out sources that are out of scope, for example video sources or sources that are in a language other than English. This module also provides access to supporting data sources such as Wikipedia, these additional data sources are used to aid in different processes, such as NTR validation.

3.3.7 Information Repository

This component is for storing all the extracted information and its associated metadata. When the information is extracted from multiple nodes, it is the task of the control module to compile the information and store that information in this module. The information may be updated at

different intervals depending on configured requirements. This component will choose the NTRs that are presented to the user.

3.3.8 Configuration

The performance of the system may be tuned with different parameters for different configurations. For example, the configuration for testing causal relation strategies is different to the configuration for evaluating the scalability of the system. This module is intended to keep these configurations and the configurations of the user. An example of user configuration can be the information granularity such as maximum number of results, or setting an information freshness threshold, this is to say how often the information should be updated.

3.4 Conclusion

In this chapter we have presented the conceptual architecture of *Forest*. We have presented the processing steps to extract information and we have shown the supporting components that help the system to generate results. The architecture also presents components that are reviewed in detail in the following chapters.

Chapter 4

News Topic References

To model topics, specifically news topics, we require a topic model that can be found in a causal relation. In this chapter we present a topic model that addresses these requirements, we call this topic model a news topic reference(NTR). To better understand news topic references we provide a definition and some examples. We also present several approaches to extract and generate news topic references, along with their corresponding evaluations. We continue by demonstrating how these approaches can be composed and present the best approaches for different use cases. In conclusion we will show how the performance of these approaches is suitable for evaluating **Hypothesis 3** *By analyzing news articles at a sentence level is it possible to extract explicit causal relations between news topic references with a 95% accuracy.*

4.1 Introduction

The abundance of online information makes it difficult or impossible for a person to process this large amount of information, this problem is known as *information overload*. It is readily evident in online news, there are so many sources of information generating content that tools are needed to process all this information. These tools can include clustering similar news articles and organizing the articles into predefined categories, such as *sports* or *finance*. In this work we propose to organize the news into a network of causally related news topics. The system we developed to automatically extract this causal relations between news topics is called *Forest*. The goal of *Forest* is to facilitate understanding and navigation of the news. We facilitate understanding by giving a context to a news topic in the form of a causal network for a news topic. We believe that causally related news topics allow the user to navigate from one topic to the next while maintaining a coherent relation between the topics.

To extract a causal network of news topics we require two main components: a component to extract causal relations from natural language text, and a component to extract news topics from causal relations. In this chapter we review the methods to extract news topics references found in natural language text.

4.1.1 News Topic Reference Definition

A news topic reference can be defined as the phrase used to name a news topic, where a news topic is defined as a seminal event and the activities related to this event. To illustrate consider the phrase “*the mortgage meltdown*”, this phrase is used to name a set of financial problems worldwide in 2008. This phrase is a news topic reference, because it is used to name a seminal event and the activities related to this event. There may be phrases that present polysemy, one such phrase is “*hurricane Katrina*”, this phrase may refer to the hurricane itself or the activities and events surrounding the hurricane event. In this case the phrase is still considered a news topic reference. There is also a case where a single news topic reference may refer to multiple news topics, for example the phrase “*the gulf war*”, this phrase may refer to the Persian gulf war of 1990, or it may refer to the Iraq war of 2003. Temporal features are used to distinguish between different news topics that use the same phrase as a news topic reference. There may also be phrases that are present in several news sources that are not news topic references, for example “*man bites dog*”. This is not a news topic reference because it does not refer to a seminal event, in other words an event that was a precursor for other events.

4.1.2 News Topic Reference Types

For this work we will define two types of news topic references, namely current and former news topic references.

Former News Topic References are the news topic references that show little or no change. These types of news topic references are used for well defined or concluded news topics. In many cases these news topics have already passed and are not current news anymore.

Current News Topic References are the news topic references for ongoing events, or news topics that are not well defined therefore susceptible to changes in nomenclature.

There are two main reasons for defining these types of news topic references: First, in a causal relation the cause must precede the effect, therefore it is likely that the cause will be a former news topic reference. Secondly, the approaches to extract the different types of news topic references are considerably different. We will show how some approaches are better suited for certain types of news topic references.

4.2 Architecture Interface

Based on our definition of a news topic reference and the prior architecture chapter, we present the basic input and output requirement for this component.

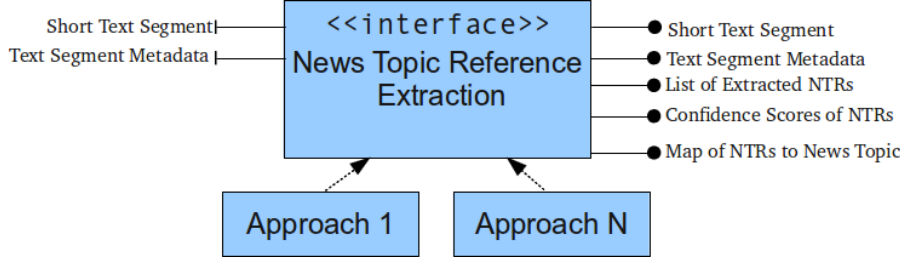


Figure 4.1: Interface for Extraction of News Topic References(NTRs)

In Figure 4.1 we can see that the main input is a text. This text is a segment of a sentence therefore it is often less than 20 words long. The output is the same sentence with additional information indicating the detected news topic references if there are any, a confidence score to indicate how probable it is for the extraction to be correct, and a map that links the extracted news topic references to a news topic identifier. The last output is used to map multiple phrases that refer to the same news topic, in other words news topic reference aliases.

The extracted news topic references can be stored in a format similar to the one shown in Figure 4.2.

The stored information is used to minimize processing costs, in particular for former news topics, as stated in the architecture chapter.

4.2.1 Semantic Relations

Several types of semantic relation may appear between news topic references, here we illustrate some of these relations and indicate how they are used in the overall system.

The foremost semantic relation in this system is a causal relation, this is the cause and effect relation between two news topics. Below is a list of additional semantic relations that may be found between news topics.

Subsumes Relation also known as *part-whole* relation. This semantic relation occurs when one news topic is included or absorbed in another news topic.

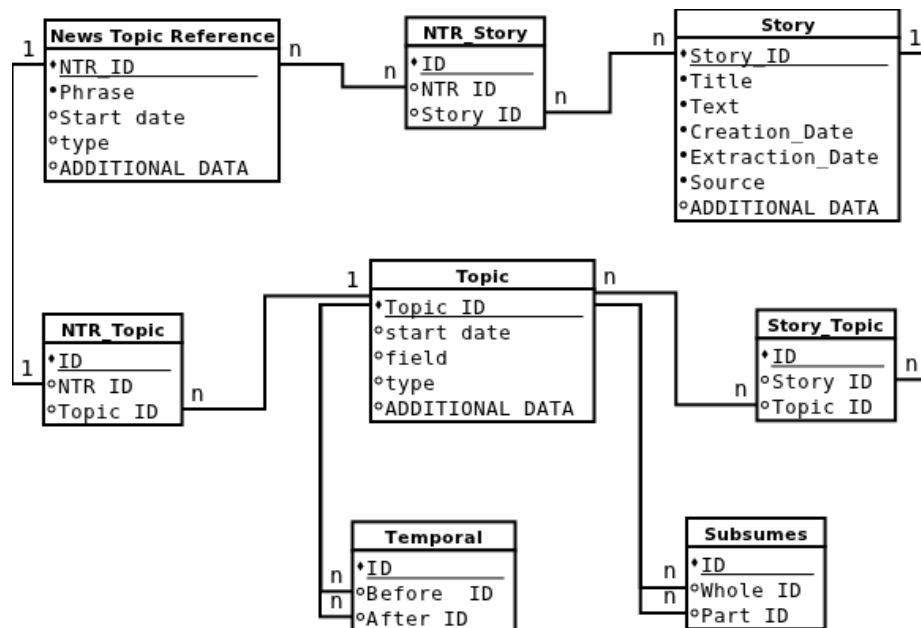


Figure 4.2: Conceptual database Scheme for relations between news topic references, news article stories and news topics

Temporal Relation provide the news topics with an order based on when the event occurred. For example, between a cause and effect there must be a temporal order of the cause happening before the effect. This semantic relation can help define and distinguish news topic references.

These semantic relations can be used to refine the results of a causal network of news topics. In the causal relation extraction chapter we will show how these two additional semantic relations are used to help detect causal relations.

4.3 News Topic Reference Validation

In this section we will present our contribution to the field of message understanding. We will present several approaches used to validate a phrase as a news topic reference. We will provide details about each approach and indicate its strengths and weaknesses.

The first approach uses the frequency and location of a phrase within a story to classify if the phrase is a news topic reference. The two following approaches use Wikipedia as a knowledge base to classify a text as a news topic reference. The final approach uses machine learning to classify the text.

4.3.1 Input Text

In order to validate if a phrase is a news topic reference or not, we assume that the input phrase has been preprocessed. This preprocessing is done in order to reduce the processing load on the system, it also facilitates the validation of the phrase.

Basic validation is done in previous steps, this includes text encoding filtering, language filtering and minimum-content filtering. These filters are important because the information is extracted from multiple sources and the data may not be consistent among these sources. The text input should not be more than 20 words long, because the input text is only part of a sentence and a sentence is often less than 20 words long.

After the initial filtering we should have a well formed text snippet that is part of a longer sentence.

4.3.2 Phrase Location and Frequency

Our first approach is based on the way users processed text snippets to find news topic references. In practice the user would use the NTR input text

as a query for an online search engine, such as Google¹. Some of the words from the input text would be highlighted in the results, these were often functional words and synonyms to these words, this step provided the users with semantically similar results, for example the query “*the economic meltdown*” would return highlighted text such as “*financial meltdown*”; in this result we can see that the stop word *the* is removed and a synonym for *economic* is used. The next step would be to use this highlighted text as a new query to find news articles about the initial text snippet, if an article was titled with the query phrase or the query phrase was in the summary text of an article, then the user considered the query phrase as a news topic reference.

The systematic approach to this user practice is stated below. In this approach the Google search engine was programmatically accessed, this facilitated processing that provided keyword tagging as well as synonyms and results ranking. The text snippet used as an input is a segment of a causal sentence. It is only a segment of the sentence because the sentence is split into the cause and effect fragments, and these are processed separately.

First Phase Keyword Selection

```
Input: preprocessed text snippet
1 use input as query for search engine
  result: ranked snippets with keyword tags
2 from the top results select the tagged keywords
3 exclude results that are exact matches to input
4 score keyphrases by frequency of appearance and word count
  score = frequency x word count
output: ranked list of candidate news topics references
```

Second Phase Keyphrase Validation

```
Input: ranked list of keyphrases
1 for the top N keyphrases in list
  use keyphrase as query for search engine
2   if the keyphrase is the title of at least one news article
    then it is a news topic reference
3   if the keyphrase is in the first paragraph of at least two news articles
    then it is a news topic reference
4   if the keyphrase is in the body of at least three news articles
    then it is a news topic reference
  else
5 it is not a news topic reference
Output: list of validated news topic references
```

The algorithm given above is dependent on the performance of the search engine for highlighting keyphrases and retrieving the correct articles. It is

¹www.google.com accessed 1.9.2012

possible to implement these features in a local search engine, for example with Lucene² and Solr³.

Approach Evaluation

A prototype implementation of *Forest* was developed and used to generate the input to this approach, namely the text snippets. The process of evaluating these text snippets with this approach is described below.

Initially a list of 20 seed news topic references were empirically selected. Using the process stated in the architecture chapter, a collection of 214 causal phrases were extracted from online news sources. One news topic reference was extracted from each of the 214 causal phrases, excluding the 20 initial news topic references.

The extracted news topic references and source sentence were provided to users for evaluation using an online survey system⁴. This online service called Crowdfunder uses a crowdsourcing approach to resolve repetitive tasks. Using the service, users are given a small set of tasks, see Figure 4.3. In this case each task given as a question, an example is given below "Is *Extracted_NTR* a reference to an event or activities? For example in the sentence: *source_Sentence*"

In this example, the extracted news topic reference is indicated by *Extracted_NTR* and *source_Sentence* indicated the sentence that was used to extract the news topic reference. Users could answer yes, no or unknown, the latter option is to allow for phrases that are unclear or incorrectly parsed from the original source.

Users were also given gold standard tasks, these are tasks where the results are known, to ensure that the results from the annotator are correct.

Results show that 54% of the extracted NTR were correctly classified as a reference to an event or activity, 32% were incorrectly classified and 14% were unknown.

Conclusion

There is room for improvement for this approach, yet it may be suitable for very novel current news topics, where information and terminology is actively changing. Further development is required to achieve a useful performance.

²<http://lucene.apache.org/core/> accessed 1.9.2012

³<http://lucene.apache.org/solr/> accessed 1.9.2012

⁴<https://crowdfunder.com/jobs/65082> accessed 1.9.2012

Flag references to event or activities

Instructions Hide

Given a phrase flag if it is a reference to an event or activities.

Is **backed securities** a reference to an event or activities?

for example in the sentence:

To understand how the subprime mortgage crisis led to the worst U.S. recession since the Great Depression, read How can mortgage-backed securities bring down the U.S. economy? For more information on mortgages and the financial system, please explore the links on the following page.

Choose one (required)

☐ Yes

☐ No

☐ Unknown

Figure 4.3: A screen capture of an task from Crowdfunder

4.3.3 Knowledge Based

In this section we present two approaches that use Wikipedia as a knowledge base. The first approach uses structured-information features to classify a Wikipedia article as an event. The second approach uses a special purpose page, called the *Current Events Portal*⁵, to extract a list of news topic references. Both approaches are based on the fact that the titles to select Wikipedia articles, namely the articles about seminal events, are news topic references.

Wikipedia as a Knowledge Base

Multiple academic works have used Wikipedia as a knowledge base, because of many reasons including its accessibility and structured data. The information found in Wikipedia is not guaranteed to be correct, although it has been shown that the accuracy of the information is comparable to other commercial encyclopedias⁶. By using Wikipedia as a knowledge source we intend to demonstrate the viability of the approach and not the correctness of the information found in Wikipedia.

⁵http://en.wikipedia.org/wiki/Portal:Current_events accessed 1.9.2012

⁶http://en.wikipedia.org/wiki/Reliability_of_Wikipedia accessed 7.20.2012

Based on our definition of a news topic reference as the phrase used to name a seminal event, we can safely assume that the titles to articles about seminal events can be considered news topic references. For the task at hand, the challenge is distinguishing between the Wikipedia articles that are about seminal events and those that are not, in order to use the titles to these articles as news topic references.

Wikipedia Event Navigation

In this approach we used markers, found in Wikipedia articles, for binary classification of an article as regarding an event or not. Markers are given a positive and negative weight, if the sum of the weights is above a predefined threshold then the article is classified as an event article. The values of the weights and the thresholds were empirically selected. There are two main types of markers: structural information markers, for example the presence of an infobox with an event field, and keyword markers, such as article sections containing the word “aftermath” or “timeline”.

The completed system presents the results in a dynamic visualization of Figure 4.4. The visualization is used to compare the performance of a graphical representations versus a textual representations of the information found in Wikipedia.

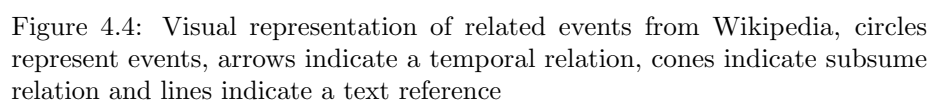
To classify an article certain markers were given a weight, then if the article was above a predefined threshold weight the article was classified as an event article. The weight values were empirically defined as well as the threshold value. Not all markers are present in the articles, some positive and negative markers are given below as examples.

Positive markers are indicators that the article is about an event, negative markers indicate the article is not about an event. A combination of certain marker would give the article a weight above the threshold level, for example if the article has time, location, action and participants markers then the article is classified as an event.

Many of the markers were taken from the infobox section of the Wikipedia articles, see Figure 4.5. Infoboxes are form type structures that contain summary information about an article, and they are often found in event type articles.

Positive markers included the presence of location, time, participants and action information in the infobox. Specific fields used as markers include: *location*, *date*, *time*, *timezone*, *perps*, *target* and *type*. The value of the field is not a marker, only the presence of the field itself. To make some of the fields more clear consider the following: *perps* indicates the people or entities that generated the event and *type* indicates the activity in the event, for example “natural disaster” or “conflict”.

Other positive markers include keywords present in the article title including “civil war”, “uprising” and “scandal”. A strong positive marker



Munich massacre




Image of terrorist looking over the balcony of the Israeli team quarters at Building 31 of the Munich Olympic village. This is the most widely recognizable and iconic photo of the event.^{[1][2]}

Location	Munich, West Germany
Date	5–6 September 1972 4:30 AM – 12:04 AM (UTC+1)
Target	Israeli Olympic team
Attack type	mass murder , massacre , hostage-taking
Death(s)	17 total <ul style="list-style-type: none">• 11 Israeli athletes• 5 members of Black September• 1 West German police officer
Perpetrator(s)	Black September

Figure 4.5: An infobox that contains a *location*, *date* and *perp*(Perpetrator) fields, among others.

	Positive (predicted)	Negative (predicted)	Recall
Positive (actual)	81	138	37%
Negative (actual)	19	3686	
Precision	81%		Accuracy 82%

Table 4.1: Wikipedia event article classification

is a time reference in the title, for example the “*April 9 tragedy*”, or “*Mississippi flood of 1927*”. More positive markers include article sections with keywords including: *timeline*, *chronology*, *background*, *prelude*, *causes*, *impact*, *aftermath*, *reaction* and *response*.

Negative markers are also used, these types of markers reduce the possibility of an article to be classified as an event type article. Negative markers included titles that start with the following keywords: *list of*, *timeline* and *history of*. Additional negative markers are titles that contain the keywords: *disambiguation*, *song*, *album*, *film*, *river* and *place*.

The sum of the positive and negative markers is then compared to the threshold value, if the sum is above the threshold then the article is classified as an event type article, if the article is about an event then we consider the title to this article a news topic reference.

Evaluation An empirical test was conducted to evaluate the event classification algorithm. For this, We classified random Wikipedia articles as events, until we completed a dataset of 100 articles about events. The dataset was then manually annotated to identify false positives. We also manually checked the Wikipedia articles that where not classified as events, to find false negatives. The results are presented in Table 4.1.

The performance metrics for the selected articles is precision 81%, recall and an overall accuracy of 82%. An analysis of the results shows that the performance is highly affected by the quantity of information in articles, for example articles that contain many sections or articles that do not contain an infobox were in many cases misclassified. In future work we consider information quantity as a feature for classification.

Conclusions This approach is geared toward former news topic references. It provided valuable insight into the surface form of news topic references, namely the types of words that are used to name news topics. In the following approach we will see how this insight was used, and how it takes us toward a machine learning approach.

Curated News Story Collection

Here we present an approach that uses hand crafted news article collections to extract news topic references. These online collections are often curated by expert editors, such as *BBC news special reports*⁷. The online collections distinguish themselves from news aggregators in focus, news aggregators focus on novel content, while news collection focus on seminal events in specific fields.

While several news collection sites are available online, we selected Wikipedia's current events portal⁸, because of the insight we discovered in the event detection approach. In figure 4.6 we can see four main sections to the Wikipedias portal:

Topics in the news (above) provides a short summary of the most current news topics, many terms given in this summary are linked to Wikipedia articles, for example locations and people.

Ongoing events (right) presents an outline of seminal events in the news, the titles of these events link to Wikipedia articles. The titles to these events can be considered news topic references.

Wikinews articles (below) are news articles that are produced in a similar way to Wikipedia articles, this is to say they are generated by a diverse community of volunteers and vetted by online community participation.

Non-Wikipedia articles (center) is a section that presents short text summaries of a news topic, similar to the topics in the news section, except that it also includes a link to a news article about the topic. These news articles are generated by a source that is not Wikipedia. Often the articles are from well known online news portals, like BBC.com or CNN.com.

From these four sections we selected the non-Wikipedia articles, because the summary contains links to Wikipedia articles and it also provides a link to a relevant news article. An overview of the process to extract news topic references from the selected section is presented below.

1. The content of this web page is scraped, from the current month to the month it began.
2. The relevant sections are processed into a table format that contains: the text, the links to Wikipedia, and the links to other sources.
3. Internal links are further processed, these are links to Wikipedia articles. Many of these links are references to the entities mentioned in a news topic, we wish to exclude these references from the results.

⁷<http://www.bbc.co.uk/news/special-reports/> accessed 1.9.1012

⁸http://en.wikipedia.org/wiki/Portal:Current_events accessed 1.9.1012


Portal:Current events

From Wikipedia, the free encyclopedia

Worldwide current events · Sports events

Topics in the news

- In sumo, Mongolian wrestler **Harumafuji Kōhei** (*pictured*) is formally promoted to become the 70th yokozuna.
- At the **64th Primetime Emmy Awards**, *Homeland* and *Modern Family* win the awards for Outstanding Drama and Comedy Series, respectively.
- In Gaelic football, Donegal defeat Mayo in the **2012 All-Ireland Senior Football Championship Final**.
- After 40 suicides of victims trigger a parliamentary inquiry, the Roman Catholic Archdiocese of Melbourne confirms the sexual abuse of 618 children over 80 years.
- The extent of Arctic sea ice falls to a record minimum.
- Anticipating violent reactions to a series of cartoons depicting Muhammad in the magazine *Charlie Hebdo*, France announces plans to temporarily close its embassies in 20 Muslim countries.

 Syrian civil war – Wikinews – Recent deaths – **More current events...**

September 28, 2012 (Friday) edit history watch

Armed conflicts and attacks

- The Kenyan Air Force bombs the Somali city of Kismayo with claims that an armoury belonging to the al-Shabab movement was destroyed. *(Mmegi)* [ⓘ]

Disasters

- A Sita Air plane with 16 passengers and three crew members on board crashes on the outskirts of the Nepalese capital of Kathmandu, killing everyone on board. (BBC) [ⓘ] (AFP via Google News) [ⓘ]

September 26, 2012 Wikinews articles

- Rockets, mortars fired from Syria land in Israel
- Maharashtra, India state transport bus falls in river; at least seventeen dead

About this page · Suggest a headline
News about Wikipedia

Ongoing events

Political

- Arab Spring
 - Syrian civil war
 - Bahraini uprising
- Mexican Anti-Imposition Protests
- Senkaku Islands dispute

Economic

- Bardays LIBOR manipulation
- Global financial crisis

Figure 4.6: An edited screen capture of the Current events portal from Wikipedia

4. As we have learned from the previous approach, a date reference in the title is a strong indicator of an event title. Based on this information we select internal links that begin with a date in the title.
5. Upon review of the list at this stage, several dated entities are detected, for example “2012 US Olympic team”, these are manually filtered out.
6. Date references are removed from the titles to make them more similar to the references used in news articles.
7. Filtering is done to obtain minimal content per news topic reference, for example “September 11 attack” was reduced to “attack” by removing the date, then filtered out because it does not provide enough information.

The result of this process is a list of over 1000 news topic references with related Wikipedia article and the corresponding links to online news articles.

News Topic Reference Analytics Using part of speech tags and other tools, we analyzed the resulting set of 1235 news topic reference and discovered the following information:

- All news topic references contain a noun.
- Close to one in ten news topic references contain a number.
- Close to one in ten news topic references contain a verb.
- Close to four in ten news topic references contain adverbs.
- On average there are 3.25 words per news topic reference.
- The standard deviation is 1.26 words per phrase.
- The minimum word count is 2 and the maximum is 13.
- Approximately 2/3 of the words in a news topic reference are nouns.
- Approximately eight in every ten nouns are proper nouns.

Conclusions This approach makes it possible to obtain a small set of current news topic references and a larger set of former news topic references. The approach, with some variation, can be applied to other curated news collection sites, for example BBC special reports. The results of this approach aids in the development of a more refined machine learning approach.

4.3.4 Machine Learning

In this final approach we use a text classifier suitable for short text classification. The classification strategy that was selected is called a dictionary classifier, implemented in the Palladian toolkit⁹. Essentially this approach assigns probabilities for each word, or n-gram, to be a news topic reference, these probabilities are then used to classify a phrase as a news topic reference.

To implement this approach, a collection of positive and negative examples are required to train the system. The positive examples were taken from the results of knowledge based approach. To generate the negative examples, specifically phrases that are not news topic references, a small collection of news articles were manually reviewed and all news topic references were removed, then the articles were partitioned with a similar distribution to that of the extracted news topic references, namely an average of 3.25 words per phrase and a standard deviation of 1.25 words.

The positive and negative examples were combined to generate the training and testing data. The performance of the dictionary based classifier are promising, based on a conservative training and testing split, a precision and recall of approximately 92% is achieved, see figure 4.8.

Implementation

To implement this approach an agile prototyping tool was used. This tool, called Knime¹⁰, provides a user friendly development environment for information mining. It provides access to several implementations of information extraction and processing techniques including the text classification implementations found in Palladian.

A process flow was generated in Knime¹¹ to evaluate several classification methods, see Figure 4.7.

Results show that the best performing approach is the dictionary classifier. We assume that this is due to several factors including text length and type of language used in news topic references. Many text classification methods rely on extensive text segments, long text documents are not available for news topic references, therefore the performance of these classification approaches is reduced. In contrast a dictionary classifier is word based and does not require long text documents, such as a complete news article.

⁹<http://palladian.ws> accessed 1.9.2012

¹⁰<http://www.knime.org/> accessed 1.9.2012

¹¹<http://www.knime.org/> accessed 1.9.2012

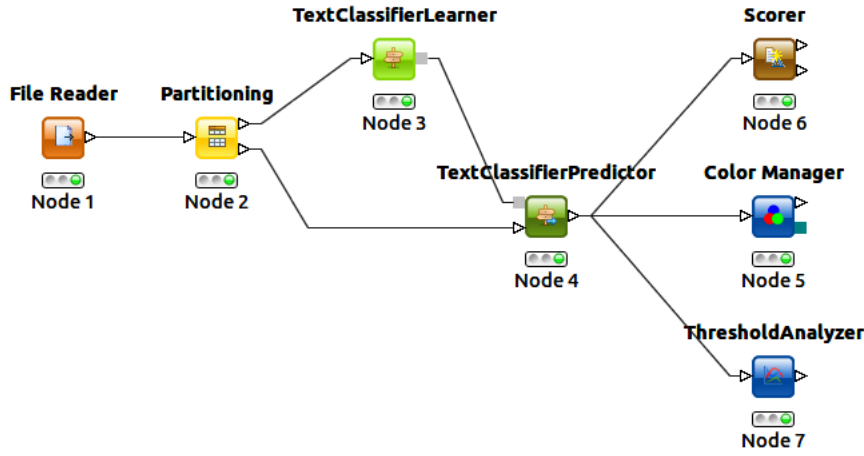


Figure 4.7: Screenshot of the Knime classification process flow

Evaluation

The results of this approach are presented in Figure 4.8. The positive and negative examples were equally split into training and testing sets, this is a conservative split, in many cases the training set is substantially larger than the testing set. The complete gamma of thresholds were analyzed and the best results were found at 61%, the F_1 score¹² at this threshold is 92.5%. We believe that this performance level is suitable to obtain useful results from *Forest*.

Conclusion

The performance of this approach is very useful in many aspects. It can be used for current news topic references, the results can be ranked and provide a probability of being correct. It is scalable and can be continually updated. Additional features can be implemented and evaluated in future work. These features could include part-of-speech tags, that have been shown to be relevant in previous approaches.

4.4 Approach Combination

We have shown how the approaches presented in this chapter can complement each other, and how this composition of approaches generates a

¹²http://en.wikipedia.org/wiki/F1_score

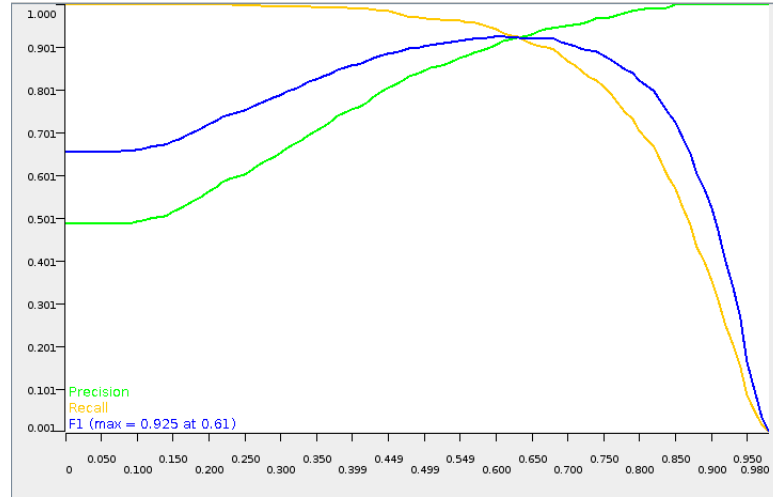


Figure 4.8: Screenshot of classification results: Dark line indicates the F_1 measure, Y axis is percentage, X axis is threshold value

suitable system for extracting news topic references from natural language text. In this section we consider the aspects that surround the approaches.

4.4.1 Additional Approaches

There are several approaches that have not been fully developed due to time and resource constraints. One of these approaches is based on textual and temporal similarity of search engine results. In this approach a candidate phrase is given as a query to a search engine, the top ranked results are then evaluated for temporal and keyphrase similarity, this is to say the generation dates for the news articles are close to one another and that certain keyphrases such as locations, activities and entities are found in most of the top ranked results. While the performance of such an approach may be limited, it may still be useful for current news topics. Another unimplemented approach is based on generating a summary for a news article. The summary consisting of automatically answering the questions who, what, when, where and how regarding the content of a news article. The summary content could be used to generate news topic references. This system to automatically answer the questions has already been partially developed[Wunderwald, 2011]. In the following section we will see how news topic reference aliases can be generated. We assume that if a news topic reference alias can be generated for a phrase, then that phrase is a news topic reference itself.

4.4.2 Complimentary Approaches

In this chapter we have seen that approaches for current news topics differ to approaches for former news topics. We have seen that by using the curated collections approach we can train the machine learning approach. The curated collection approach retrieves former news topics, and the machine learning approach can be used for current news topics. The current news topics will, over time, be converted into former news topics and this information can be used to evaluate and improve the results of the machine learning approach.

In this way we can develop a continuously updated repository of news topics and news topic references, as mentioned in the architecture chapter.

In considering how these approaches fit together, we should also consider the source of the phrases, namely the causal relation extraction component. This component will be reviewed in the following chapter, here we only consider that it provides a phrase that is part of sentence, to evaluate if this phrase contains news topic references.

4.5 News Topic Reference Alias

In this section we review the aliases of news topic reference. We briefly provide a definition of aliases and an approach to generate aliases.

Aliases are different news topic references that refer to the same news topic events, for example the news topic reference “*Mortgage meltdown*” is an alias of “*Subprime crisis*” because both refer to the same events in late 2007. Aliases can be defined as alternate names or synonyms to a news topic.

Aliases can be used to generate multiple queries for a news topic and also to reduce a result set by consolidating multiple aliases to one resulting news topic reference, aliases may also be used to classify a phrase as a news topic reference, if an alias can be generated for a phrase and that alias is a news topic reference then the original phrase can be considered a news topic reference.

4.5.1 Wikipedia Redirect

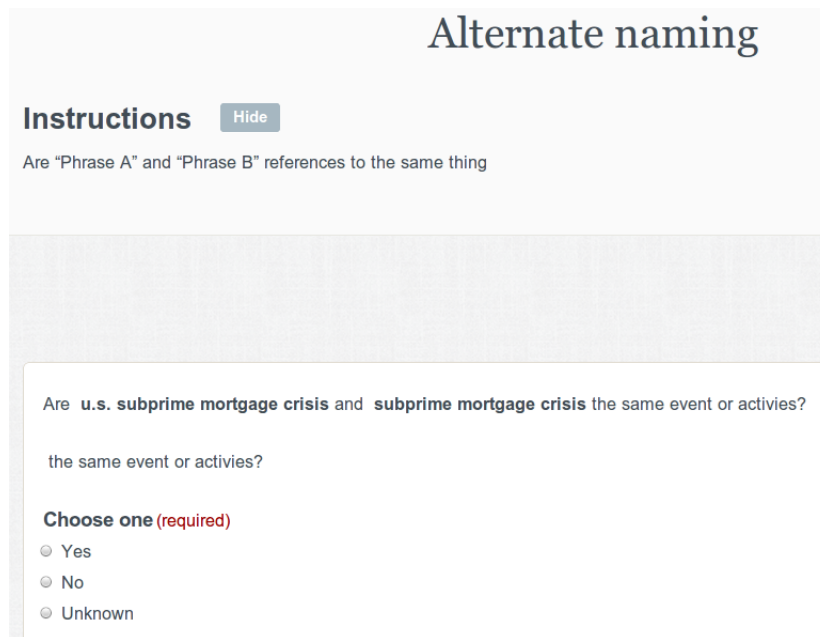
One approach to generate aliases is based on the structured information found in Wikipedia, namely redirect pages. Redirect pages are defined in Wikipedia as “a page which has no content itself, but sends the reader to another article, section of an article, or page, usually from an alternative title”. For example if you search for “*Subprime crisis*” or if you select an internal Wikipedia link with the same term, then the result will be the article titled “*Subprime mortgage crisis*”, therefore both links reference the

same page, this indicates that they are aliases of one another. In essence multiple phrases will be directed to the same article, because these phrases refer to the same concept.

The titles of the redirect pages are considered aliases if they link to the same article, and that article is titled with a news topic reference.

Using the redirects information from Wikipedia it is possible to obtain aliases for many former news topic references, though it may not be possible to obtain aliases for all news topic references.

An implementation of this approach was evaluated and results showed an overall accuracy of 65%. The evaluation was done using an online survey system¹³. In this survey users were given a set of phrase pairs that are news topic reference aliases, see Figure 4.9. Users were then asked to confirm if both phrases refer to the same event or activities. In 65% of the cases the user answered *yes*, in 27% of the cases the users answered *no* and in 8% of the cases the user answered *unknown*. After reviewing the results we found that because of obscure terminology or terms given in a language other than English, many of the answers were considered incorrect or unknown.



Alternate naming

Instructions Hide

Are "Phrase A" and "Phrase B" references to the same thing

Are **u.s. subprime mortgage crisis** and **subprime mortgage crisis** the same event or activities?

the same event or activities?

Choose one (required)

- ☐ Yes
- ☐ No
- ☐ Unknown

Figure 4.9: Screen capture of online survey for alias evaluation

¹³<https://crowdfunder.com/jobs/65082> accessed 1.9.2012

4.5.2 Additional Approaches

The task of finding news topic reference aliases is a subtask of finding semantically similar phrases. It follows that additional approaches to find aliases are based on approaches to find semantically similar phrases.

One of these intended approaches uses the correlation between query patterns to find similar phrases. An example of query patterns is given in Figure 4.10. In the figure we can see that the X axis is time, and the Y axis is a normalized value. The value is based on the number of times the term is used as a query in a search engine, the value is normalized so that patterns with different scales can be compared. This approach is based on the assumption that query patterns of news topic reference aliases are similar, or closely correlated to one another.

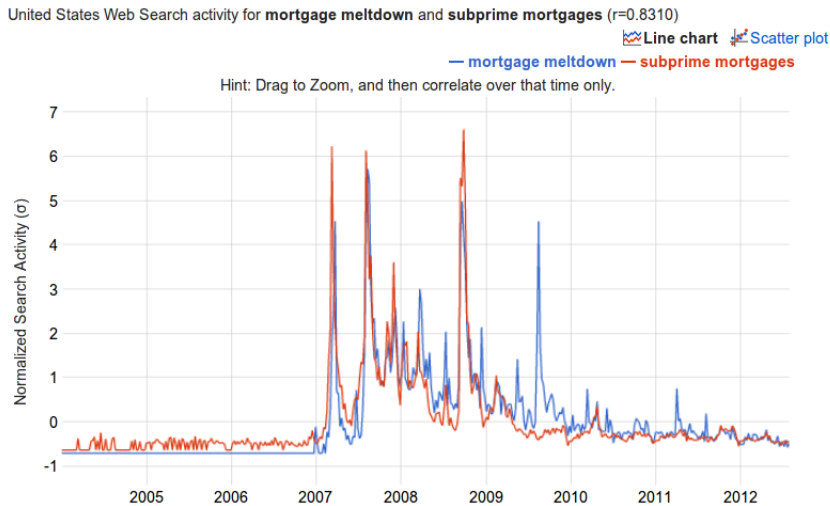


Figure 4.10: Example query pattern of “*Mortgage meltdown*” and “*Subprime mortgages*”

If we assume that the queries used to search for a news topic are news topic reference aliases, then different queries are aliases of one another. We assume that query patterns for aliases follow a similar distribution over time, then we can use the correlation between query distributions to find news topic reference aliases. For example in Figure 4.10 the news topic reference aliases “*Mortgage meltdown*” and “*Subprime mortgages*” are correlated queries, because the query patterns have similar spikes and valleys. Therefore if we use the news topic reference “*Mortgage meltdown*” to find all the correlated queries, one of the resulting queries will be “*Subprime mortgages*”. Other results may not be aliases, these results

must be filtered out. Google correlate¹⁴ provides access to the correlated information of search queries.

We assume that the popularity of a news topic will change over time, and that this change will give a news topic a characteristic distribution, and that news topic reference aliases will follow this characteristic distribution. Therefore if we plot the distribution of a news topic reference and find all the distributions that are similar, the queries used for these similar distributions could be news topic reference aliases.

A preliminary evaluation of this approach has shown that some relevant results are generated, though there is some filtering required to remove results that are not aliases. A possible way to filter out these results is by comparing the similarity measures for the query results, this is to say that different queries to a search engine would return similar results.

4.5.3 Outlook

In this chapter we have given a definition of news topic reference aliases and potential approaches that have in part been implemented, but not fully evaluated.

4.6 Conclusion

In this section we have presented our technical contributions to the field of message understanding, we have defined news topic references and given several approaches to extract them from natural language text. We have provided an evaluation for some of these approaches and shown that the performance is suitable for the proposed system to function. The completed system will then be used to evaluate the hypotheses given in the introduction chapter.

In this chapter we answered the supporting thesis: How can coherent news topic references be extracted from short text segments, this supports **Hypothesis 3** *By analyzing news articles at a sentence level is it possible to extract explicit causal relations between news topic references with a 95% accuracy.*

¹⁴<http://www.google.com/trends/correlate/> accessed 1.9.2012

Chapter 5

Causal Relation Extraction

In this chapter we review causal relations. We provide several definitions of causal relations, we also present a naïve approach to extract causal relations from natural language text, then we show the development of an evaluation dataset and we also show a machine learning approach for causal sentence classification. before we conclude, we provide insight into semantic role labelers.

5.1 Introduction

In this work we intend to extract causal relations between news topics, linking many topics together to make a network. To generate this network a system called *Forest* is developed. The network is intended to help the user navigate and understand the news. One of the central components to *Forest* is the causal relation extraction component. This component is used to establish a connection between news topics.

The goal of this component is the following: Given a sentence, classify the sentence as causal or non-causal, in other words to indicate if the sentence encodes a cause and the corresponding effect. If the sentence is causal, then the component should parse the sentence into cause and effect.

The input to this component is a well formed sentence. By well formed we mean: a sentence that has been preprocessed to meet certain requirements, such as English language and consistent text encoding. The output is the sentence classified as causal or non-causal, if the sentence is causal, then the cause and effect should be tagged.

5.1.1 General Definition

Although many people have an intuition of what a causal relation is, there are often confusions regarding this type of relation, for example confusing correlation with causation. We have provided a short definition of causal relations in the related works chapter. In this chapter we provide a more functional definition.

5.1.2 Definition in literature

There are multiple formal definitions of causation in literature, some definitions are used for natural language understanding, other definitions may be used in different fields such as logic, mathematics and philosophy. We focus on the definition used for natural language understanding, and we include the distinction between implicitly stated and explicitly stated causation. Explicitly stated causation is present when a sentence contains a causal marker, also known as a causal keyword, such as “because” or “consequently”. Implicitly stated causation is present when a sentence is causal though there is no causal marker, for example in the sentence “he was very tired and went to sleep”, the connective “and” does not encode causation, but the sentence still encodes causation.

5.1.3 Interpretation

For this work we focus on explicit causal relations. Explicit causal relations are selected because they are often easier to understand, compared to implicit causal relations. Facilitating understanding is one of the goals of *Forest*.

An additional advantage to using explicit causal relations is the following: Causal sentences contain causal markers, there is a low processing cost to locating causal markers in text, it follows that sentences can be pre-filtered, to reduce the processing load of causal relation classification.

We should note that many causal markers are ambiguous, and do not always encode causality, for example the marker “since”, can encode causality, but it can also encode a temporal anchor. To illustrate consider the following examples: In the sentence “he bought an umbrella since it was raining”, the marker “since” encodes causality; in the sentence “it has not rained since he bought an umbrella”, the marker “since” encodes a temporal anchor. This example illustrates one of the main challenges in explicit causal relation classification.

5.2 Additional Semantic relations

While *Forest* focuses on causal relations, there are other semantic relations that may be useful to detect a causal relation. Two semantic relations that support discovery of causal relations are:

- Temporal order, also known as timeline or temporal relations.
- Subsumes relation, also known as part-of or part-whole relations.

Temporal order aids in discovering causal relations because the cause often precedes the effect. An exception to this order is when an activity is preemptive. Personal experience shows that most causal relations are reactionary and not preemptive, therefore we will consider this case negligible. The set of possible causes to an effect is reduced to those that occur before the effect. Because news topics often have temporal information, it is possible to use this information to establish a causal relation.

News topics may consist of multiple events or activities. We define a news topic as coarse grained if it includes many events, as compared to other news topics. A news topic is fine grained if it consists of few events, as compared to other news topics. It is also possible for one news topic to be subsumed by another, for example the “*Lehman brothers collapse*” news topic is part of the “*credit crunch*” news topic. This granularity information can be used to establish a causal relation from natural language text. In work done by [Mulkar-mehta2011a] information granularity is used to detect causal relations. News topic granularity can also be used to define the scope of the results presented in *Forest*. When many results are available, the granularity level can be used to define what results to initially present. Our main source of supporting-semantic-relation information is Wikipedia, where many events contain structured information regarding temporal and subsumes relations. The supporting information can be used to validate the extracted information. For example, temporal and subsumes information about news topics, can be used to validate extracted causal relations between the news topic references.

5.3 Causal Markers Approach

In this section we present a basic approach to detect sentences that encode causal relations. The causal marker approach relies on the presence of phrases that indicate causality, to classify a sentence as causal or non-causal. There are many names to this approach in literature, one name is *lexico-syntactic pattern matching* another is *cue phrase approach*. To provide an intuition we present the process below:

1. Generate a list of causal markers, for example “caused” and “because”.

2. Generate a pattern where the causal markers should appear, for example the pattern *[causal_phrase] caused [effect_phrase]*
3. Classify the sentences that contain the causal marker in a corresponding pattern as a causal sentence.

Parse trees are often used to define the pattern, for example the pattern *NP1 CV NP2* is used to detect causation between nominals, in this case *NP1* is defined as the first noun phrase, *NP2* as the second noun phrase and *CV* is the causal marker. In this case the classification is dependent on the accuracy of the parse tree. The pattern could also be a minimal text length before and after the causal marker. There is a trade-off between precision and recall based on the amount of causal markers. A small set of low ambiguity causal-markers will have a high precision and low recall, A large set of causal markers will include ambiguous causal markers, this will provide a higher recall but lower precision. Therefore the selection of causal markers is key in the performance of this approach. In *Forest* we use this approach with a large set of causal markers as a pre-process step. We use this approach to reduce the amount of sentences evaluated with a more resource-intensive approach.

There are several sources for the set of causal markers [Girju2003, Girju2002a]. In the related works section we noted some of these sources. In *Forest* we focus three causal marker sets:

- A *small set* of four markers, generated by Eduardo Blanco[Blanco et al., 1991] used in a machine learning approach.
- A *larger set* of 22 causal markers, generated by Roxana Girju[Girju, R., Moldovan, 2002] considered low ambiguity and high frequency.
- An *extended set* of 88 markers, based on Girju's set, that includes all the word forms of the initial 22 markers.

One of the largest marker sets is generated based on WordNet[Butnariu and Veale, 2008], where causal markers are words that have a gloss that contain the term “cause”. A gloss gives a definition and/or an example sentence of the word. For example one of the glosses for the word “rainbow” is “*rainbows are caused by rain*”, this gloss contains the term “cause”, therefore the word is considered causal. This example also illustrates a highly ambiguous causal term.

Forest also uses causal markers for the query extension component, see figure 3.2, in the architecture chapter. To illustrate consider an initial seed news topic reference such as “credit crunch”, a query is generated to retrieve the relevant documents, the query includes the news topic reference and the causal markers, with the intention of retrieving documents that contain causal relations, in particular a causal relation between the initial seed and a different news topic reference.

The prototype implementation was evaluated based on a small user study. To evaluate causal relation extraction between NTRs, human annotators were provided information that was extracted by the prototype. The information consisted of a seed NTR, a phrase that the system classified as causally relating the seed NTR to another NTR, and the additional NTR extracted from the causal phrase. Because the system is intended to causally relate NTRs it follows that the performance of the causal relation extraction is dependent on the performance of NTR validation. The triplets of seed NTR, causal phrase and additional NTR were generated based on 20 seed NTRs. For evaluation, 40 generated triplets were randomly selected. Annotators were asked to confirm if the seed NTR and the additional NTR were causally related in the given phrase. The system showed an accuracy of around 24%, an analysis of the results show that many of the extracted NTR were incorrect, and that question phrases were miss-classified as causal.

5.4 Causal Dataset

The dataset is used to validate the component, and it is used to train a machine learning approach.

In this first stage we classify a sentence as causal or non-causal, in later stages we will extract the causal actors, this is to say the cause and the effect. In *Forest* only sentences considered interesting would require parsing, for a sentence to be interesting it should be causal and should also contain a news topic reference. To obtain this dataset we first searched for freely available collections, the only suitable collection found was from Kira Radinsky. This collection contains a set of more than 6000 sentences, the sentences are taken from the titles of New York Times articles, additionally the sentences are annotated with parse tree information, for example the sentence “songwriter’s death lead to suits against preacher”:

```

songwriter's death
COE:subjattr:songwriter,
COE:subj:death,
COE:subjattr_uri:dbpedia:Songwriter,
COE:subj_uri:dbpedia:Death,
COE:date:01/01/2007,
CO:source:nytimes.com,
COE:originaltext:songwriter's death
CO:leadto
suits Against preacher
COE:subj:suit,
COE:prepobj_Against:preacher,
COE:subj_uri:dbpedia:Suit,
```

```
COE:prepobj_Against_uri:dbpedia:Preacher,
COE:date:01/01/2007,
CO:source:nytimes.com,
COE:originaltext:suits Against preacher
```

The definitions for the encoding is not given, though we can infer that *CO:* indicates the causal marker, in this collection we find seven causal markers:

- CO:after
- CO:as
- CO:becauseof
- CO:caused
- CO:causedby
- CO:despite
- CO:leadto

The reason this dataset is not used is because many of the sentences are ambiguous, and there are many non-causal sentences in the list. for example “Nazi U-boat imperils norwegians after the war”, in this case “after” is a temporal anchor and not a causal marker.

Other datasets were requested, for example from Eduardo Blanco and Roxana Girju both replied to the request, yet neither provided a dataset.

5.4.1 Generated Dataset

In order to develop this component a dataset was generated. This dataset addresses the challenges of the reduced set of causal markers and the ambiguous causal markers found in the Radinsky dataset. In this dataset all the sentences contain causal markers, close to 16% of the sentences were manually classified as non-causal, the rest are causal.

The process to generate the datasets is as follows:

1. The data source is a small subset of news articles from the Reuters corpus[Lewis et al., 2004]. The Reuters Corpus is a collection of Reuters news stories. The subset is a collection of 2755 news articles with metadata. The metadata includes when and where the article was published.
2. From the news articles a collection of more than 32000 sentences were extracted, using the *Palladian* toolkit.
3. The *extended set* of 88 causal markers was used to pre-filter the sentences.
4. More than 900 unique sentences contained the causal markers.

5. A total of 629 sentences were annotated as causal or non-causal, 341 sentences as causal and 287 sentences as non-causal.

This dataset mimics the process flow in *Forest*, because it uses causal markers as a pre-filter. These pre-filtered sentences are then classified as causal or non-causal. The pre-filtering reduces the processing load, it may also affect the recall. For this reason we used the *extended set* of causal markers, and rely on the machine learning approach to maintain precision.

5.5 Machine Learning Approach

In this section we present an approach for explicit causal relation extraction from news-domain natural language text. the approach uses the dataset to train and validate several machine learning classifiers. Results show a performance with an F_1 measure of approximately 74%. We propose several features that may be used to classify the sentences, and present the features that are extracted. We evaluate the features and compare the performance of several classifiers. Before we conclude, we present a cross validation of the best performing classifiers.

5.5.1 Proposed Features

Here we explain some of the features used in the machine learning approach. Many of the features are selected because they follow simple writing rules, for example a sentence should not be too long or overly complex, a sentence should have a clear idea in and of itself, a sentence should be well written and have a correct grammatical structure.

Features are also selected based on characteristics of causal sentence, for example a causal sentence will link at least two actors, the cause and the effect. For the sentence to be clear, the causal actors must be clearly defined. Causation that is encoded throughout a document instead of a single sentence, is not within the scope of this work. We focus on causation that is encoded in a single sentence,

Below we present several features that are explored for the approach, and we provide a short description of why they are presented:

Causal Marker Used This is the basic feature of causal sentences, all explicit causal sentences contain causal markers. A caveat is that some markers are ambiguous, therefore not all sentences with causal markers are causal sentences.

Contains Quotes Many sentences contain a quoted section, the quotes are often important declarations, causal relations may be correlated to important declarations.

Word Count Since causation relates two actors in a sentence and the actors should be clearly defined, it follows that causal sentences may be longer than average.

Begins with Causal Marker If a sentence begins with a causal marker then the cause and effect should be separated within the sentence, for example with a comma “because of all the trouble, there has been a crisis”.

Major Sentence Before and After The Causal Marker If the causal actors are clearly defined then the text segments before and after the causal marker should be major sentences, major sentences contain a subject, verb and predicate.

Contains Numbers Encoding causation in a sentence implies a level of complexity in the sentence. Numbers in a sentence imply a level of detail that also implies complexity. It follows that to maintain the complexity of a sentence low, a causal sentence will not provide details such as numbers.

Stop Word Count Because of its complexity a causal sentence may contain less stop words than a non-causal sentence.

Contains Proper Nouns A causal sentence should contain at least two nouns, one for the cause and one for the effect.

Verb Count A causal sentence should contain at least two verbs, one for the cause and one for the effect.

Noun Location Relative to Causal Marker A noun should be present before and after the causal marker, or if sentence starts with a causal marker, then the nouns should be separated by a punctuation.

Verb Location Relative to Causal Marker A marker verb should be present before and after the causal marker, or the causal marker should start the sentence and the verbs should be separated by punctuation.

Verb Tense Mis-match Because of the temporal order characteristic of causal relations, we may find a temporal mismatch between the verbs in a sentence, for example “the broken economy leads to a rising of entrepreneurship”, where “broken” is past and “rising” is present.

Contains Temporal Phrase Many causal markers can also be used to indicate temporal order, a sentence encoding temporal order may include a temporal reference, for example “the greatest financial crisis since November has been the CITIGroup crash”, in this example we can see the temporal phrase “November”, other phrases may include the year or day of the week.

Count of Entities More than two entities referenced in a sentence may indicate the sentence is not causal.

Contains News topic references If the sentence contains more than two news topic references then it may be causal.

Temporal Order Between News Topic References If the sentence contains two news topic references that are temporally ordered, then it may be a strong indicator of causation.

Subsumed Noun Pair The causal actors are often of the same type of event, but of a different granularity level, for example a war may cause a battle or vice versa, this may be an indicator causation.

Synonyms or near-synonyms Using synonyms or near synonyms in a sentence increases the complexity of the sentence, it follows that this would be avoided in a causal sentence.

Contains Question Term There are often inquiries about the causes of activities. Sentences that contain a question word or a question mark are often questions. These sentences are non-causal.

Contains Negation The negation of the cause or the effect of an activity makes the sentence non-causal.

Position of Causal Marker If a sentence ends with a causal marker it may not be causal, for example “there has been no one since”, the causal marker “since” is not used to encode causality.

Causal Marker Ambiguity and Frequency Work has been done to quantify the frequency and ambiguity of many causal markers [Girju2003], this information may be useful to define causal sentences.

Count Of Definite Article “The” The marker “the” indicates a unique activity, having more than one unique activity may be an indicator of causality.

In this list we also include discovery features, these are features that have no clear correlation to causation, but may still provide insight into how causal sentences are formed.

Character Count Simple words often have a small character count, this may be an indicator of complexity.

Ratio of Upper to Lower Case Letters How many upper or lower case letters are found. or the ration of upper case to lower case.

Number of Punctuation Marks How many punctuation marks are found.

Number of Adjectives How many adjectives are found.

Parse-tree Pattern An analysis of the parse tree information may provide insight into causal sentences. In particular, what patterns appear before and after each specific causal marker. Because each marker may have its own pattern. By pattern we mean the order of the part-of-speech tags in the sentence.

Part-of-speech Tag Count and Pattern An analysis of the part-of-speech information may provide insight into causal sentences. In particular, what patterns appear before and after a causal marker in a causal sentence.

Repeated words The intuition is that to maintain a distinction between the cause and the effect, there will be no repeated terms.

Words Before and After Causal Marker are Related and Causal

In WordNet there are several definitions or examples of each word, each example is called a gloss, a gloss may contain a causal marker and also another word from the sentence, this may be an indicator of causality.

The features proposed here are intended to aid in discovering what features are valid and useful, not all the features may seem relevant, yet they may present latent information, therefore we consider and extract those that are available.

5.5.2 Feature Extraction

Here we present an overview of the selected features that are extracted. Many of the features can be extracted in a straight forward way. Features that require more processing to be extracted include the following:

Contains Proper Nouns

Contains Question Term for example in the sentence “the location where that occurred”, this sentence contains a wh-word “where”, but it is not used as a question word.

Verb Tense Mis-match To extract the verb tenses a part of speech tagger is required.

Major Sentence Before and After The Causal Marker This is to say that there are at least two major sentences linked by a causal marker. To extract the sentence structure a parse tree generator is required. The parse tree generator annotates the grammatical structure of a sentence, for example it annotates the subject or object of a verb within a sentence. If the sentence is too long the probabilistic parser can not perform correctly. Because the parse tree could not be generated for all the sentences, features based on the parse tree information were not further developed.

A complete list of the extracted features is given in the following section.

5.5.3 Feature Evaluation

We use two methods to evaluate the features, backward feature elimination and Chi square feature evaluation, We present the results of these evaluations here. To conclude this section we will show what features are useful for the classification task.

Backward Feature Elimination

We use backward feature elimination to rank the features. A framework was developed for this evaluation, The framework was used to evaluate the dataset on several iterations. Results of several iterations were not consistent. This may be due to the size of the dataset. To analyze the discrepancies in the results, the backward feature elimination was run on the same dataset in ten iterations. The results were averaged out for analysis.

The backward feature elimination framework used a Naïve Bayes classifier, each feature elimination iteration used 310 sentences for classifier training and validation; and 310 sentences for testing the features. Each iteration of a classification with one feature eliminated used 310 sentences, the classifier was trained on 210 sentences and it was validated on 100 sentences. A cross validation of the feature eliminations was not possible due to the small size of the dataset. The sentences were selected based on a stratified random sample of the dataset.

On average of the feature ranking is as follows:

1. Contains Quotes.
2. Stop Word Count.
3. Character Count.
4. Verb Tense Mis-match.
5. Contains Negation.
6. Verb Count.
7. Contains Numbers.
8. Count of Definite Article “The”.
9. Contains Question Term.
10. Position of Causal Marker.
11. Begins with Causal Marker.
12. Word Count.
13. Causal Marker Used.
14. Contains Proper Nouns.

In backward feature elimination the lowest ranking features are the first features to be eliminated, based on the average of the ten iterations of the feature elimination. The first three features to be eliminated reduced the error rate, The following seven features do little or no affect to the error rate. The last four features to be eliminated increased the error rate.

Chi Square Feature Evaluation

In this evaluation we compare the value of the Chi square statistic of each feature, to the classification of the sentence, namely causal or non-causal. In this evaluation we use five fold cross validation to confirm the feature ranking. The ranking is as follows:

1. Causal Marker Used.
2. Stop Word Count.
3. Character Count.
4. Verb Count.
5. Word Count.
6. Contains Numbers.
7. Contains Quotes.
8. Count Of Definite Article "The".
9. Position of Causal Marker.
10. Verb Tense Mis-match.
11. Begins with Causal Marker.
12. Contains Negation.
13. Contains Question Term.
14. Contains Proper Nouns.

The rank average based on the cross validation is given in the Table 5.1

Rank	Distribution	Feature Description
1	1 +- 0	Causal Marker Used.
2	2.2 +- 0.4	Stop Word Count.
3	3.2 +- 0.75	Character Count.
4	4 +- 0.63	Verb Count.
5	4.6 +- 0.8	Word Count.
6	6.8 +- 0.75	Contains Numbers.
7	7.6 +- 0.8	Contains Quotes.
8	8.8 +- 2.32	Count Of Definite Article "The".
9	9.4 +- 2.8	Position of Causal Marker.
10	10 +- 1.1	Verb Tense Mis-match.
11	10.2 +- 0.98	Begins with Causal Marker.
12	10.4 +- 1.2	Contains Negation.
13	13 +- 0.63	Contains Question Term.
14	13.8 +- 0.4	Contains Proper Nouns.

Table 5.1: Cross validated feature ranking

The ranking is based on a correlation value, the top ranking feature (Rank 1) has a value higher than the following four features (Rank 2,3,4,5) combined, though these four features show a correlation to the classification. The following four features (Rank 6,7,8,9) have a low correlation to the classification and there is a high variance in the cross validation. It follows that these features have a low impact on the classification. The following three features (Rank 10,11,12) have a very low correlation value. The last two features (Rank 13,14) show no correlation to the classification.

These results indicate that the top eight features are correlated to the classification. Therefore, they are useful for classification.

Based on the backward feature elimination and the Chi square correlation, we have selected the following features:

1. Contains Quotes.
2. Stop Word Count.
3. Character Count.
4. Verb Count.
5. Contains Numbers.
6. Count of Definite Article “The”.
7. Position of Causal Marker.
8. Word Count.
9. Causal Marker Used.

These markers were selected based on the correlation value to the classification and the average error rate increase produced in the feature elimination.

5.5.4 Classification Technique Evaluation

Based on the selected features we evaluated several classification techniques and present their performance.

- Dictionary Classifier F_1 0.78
- Naïve Bayes F_1 0.74
- K Nearest Neighbor F_1 0.58
- Support Vector Machine F_1 0.67
- Bagging F_1 0.71
- Decision Tree F_1 0.70
- Probabilistic Neural Network F_1 0.71

As we can see the best performing classifiers are the Dictionary Classifier and the Naïve Bayes classifier. The dictionary classifier does not use the extracted features, instead it used the text in each sentence of the dataset.

5.5.5 Cross Validation

Once we have selected a classification technique, we validate the classification approaches using five fold cross validation.

Dictionary Classifier F_1 0.74

Five folds cross validation in Table 5.2.

Percent	Line_count	Error_count
36.50%	126	46
35.71%	126	45
38.09%	126	48
40.80%	125	51
34.40%	125	43

Table 5.2: Five folds cross validation for the Dictionary classifier

Naïve Bayes F_1 0.72

Five folds cross validation in Table 5.3

Percent	Line_count	Error_count
45.23%	126	57
39.68%	126	50
40.47%	126	51
34.4%	125	43
36.0%	125	45

Table 5.3: Five folds cross validation for the Naïve Bayes classifier

The results show that a stable performance of the Dictionary classifier is slightly better than that of the Naïve Bayes classifier. Therefore we chose the Dictionary Classifier for the task of causal relation extraction and for the task of news topic reference extraction.

5.6 Conclusions

In this chapter we presented the development of the causal-relation-extraction component. We provided a definition of causal relations, and presented the development of a causal sentence dataset. We showed what features were extracted from the dataset, and how those features were used in several machine learning classifiers. The best performing classifiers were validated and selected for use in *Forest*.

Chapter 6

Evaluation

In this chapter we present the evaluation of the hypotheses given in the introduction chapter. In essence these hypotheses aim to answer three basic questions about causally related news topics: first, is this type of information useful for understanding news topics; second, is this type of information available for most if not all news topics; and finally, can the system, *Forest*, extract this information accurately.

For the evaluation of the first hypothesis we employ a user survey. This survey indicates that annotators find this type of information useful to understand news topics, we call this a *usefulness rating*. Users rated the usefulness on average 3.69 on a scale of 1 to 5, where 1 is not useful and 5 is very useful.

To evaluate the second hypothesis we analyze a large collection of news articles, to find out the availability of causal information for news topic references. We found that approximately 25% of news topics are causally related to another news topic. We also found that the likeliness of finding a causal relation for a news topic is correlated to the number of occurrences of that news topic in news articles. For example, the top 10% most frequently occurring news topics have a likelihood of 85.71% of being found in a causal relation to another news topic, and the bottom 50% least frequently occurring news topics have a likelihood of 3.22% of being found in a causal relation to another news topic.

We analyzed the distribution of selected news topics in a sample of news articles. We found that there is approximately a 25% probability that a news topic is causally related to another news topic. And that there is a correlation between the number of articles for a news topic and the likelihood of finding that news topic in a causal relation with another news topic. We also found that the distribution of number of articles (frequency) of the news topics is similar to a long tail distribution, where the occurrences of the top 5% of news topics outnumber the following 95% in frequency. It

follows that there is high likelihood of finding a causal relation for the top 5% of news topics.

To evaluate if we are able to extract the information in a way that is useful to users, we completed another survey. In this survey we, once again, ask users to rate the usefulness of the information, we also ask users to verify that the extraction is accurate, namely, if the sentence contains a causal relation between news topics. Given certain constraints we are able to extract the information accurately in 94.44% of the cases. Users rated the usefulness on average 6.34 on a scale of 1 to 10, where 1 is not useful and 10 is very useful.

In the rest of the chapter we present the evaluation of each of the three hypotheses, these evaluations contain a brief summary of the findings, then a detailed description of how those findings were obtained, and a short conclusion to each hypothesis. Before we end the chapter we give our general conclusions where we state that the system performs well, under given constraints, for the task of aiding users understanding news topics.

6.1 Hypothesis I Evaluation

In this section we present the evaluation of the first hypothesis, *Causal relations between news topics provide a context that aids the user to understand the news*. The results will show that users find causally-related news topics helpful for understanding the news. Results were generated with a user survey where 40 causally-related news topic sentences were manually selected. These manually selected sentences along with related information are called evaluation units, these units were judged a total of 137 times, each judgement consists of 12 survey questions. In following section we present the details of how we obtained the evaluation units and the results of the survey. The questions in the survey focus on validating the units and obtaining user opinion, there is also a question to obtain user comments. The central survey question is “*How helpful is Text X to understand the subject?*”, where *Text X* is the causally related sentence and *the subject* is a news topic. The average response for this rating question is 3.69 on a scale of 1 to 5, where 1 is *not at all*, and 5 is *very much*, Indicating that these results are useful for understanding the news.

In this section we will present how the survey evaluation units were generated. Then we present the details of the survey, including questions and participation, finally we present the results from the survey, before giving our conclusions.

6.1.1 Survey Evaluation Units

In order to ask users if *causal relations between news topics provide a context that aids the user to understand the news*, we require sentences that contain causal relations between news topics. We obtained a small collection of 40 such sentences, by searching through a news article collection with *Forest* and manually selecting sentences that match our description. The manual selection is to ensure that we obtain *causal relations between news topics* sentences, in the evaluation of the third hypothesis we discover how well we are able to automatically extract these sentences.

Sample Result

The following is an example of a sentence that encodes a causal relation between news topics: *Gold in European trade soared to historic highs as speculators and investors snapped the metal up on the back of strong oil prices.* The sentence contains two news topics: *Gold in European trade soared to historic highs* and *strong oil prices*. The sentence also contains a causal marker: *as*. There is additional information assigned to this sentence including: an identifier for the extraction, the source document or news article, the uniform resource identifier (URI) associated with the document. For the evaluation we only require the sentence, the news topics and the causal marker. The sentence is for the users to evaluate the hypothesis, the news topics and causal markers are for users to validate the correctness of the extractions.

Below is an itemized description of the elements of a result set.

ID An identifier for the extracted relation.

Document The source document, this may be text extracted from a web page.

Sentence The segment of text that contains a causal relation and at least two news topic references.

Causal Marker The text segment that triggered the preliminary selection.

First News Topic Reference The news topic reference extracted from the first section of the sentence, the sentence is sectioned based on the causal marker, the first section often encodes the cause, and the second section often encodes the effect.

Second News Topic Reference The news topic reference extracted from the second section of the sentence, the sections are intended to correspond to the cause and effect in a causal relation, due to inflections in the sentence this may not always be the case.

A result set is the raw information extracted by *Forest*, all the evaluations are based on these result sets. In this evaluation multiple result sets are generated by *Forest*, then several of these sets are manually selected. Additional information is obtained for the manually selected sentences, for users to compare results generated from *Forest* to results generated by an online news portal, namely Google News. In the Evaluation Unit section below we present how the additional information is obtained and we provide an example.

Data Sources

Result sets can be extracted from different types of sources of information: a static corpus of news articles, or a dynamic corpus of articles downloaded from the Internet. The static corpus allows us to duplicate and analyze results, and the dynamic corpus allows us to obtain updated and near-real-time results.

In this evaluation we use the static corpus, though we present both types of sources for future reference.

Static Corpus To be able to repeat results, we selected the Thomson Reuters Text Research Collection (TRC2). This corpus contains 1,800,370 news stories dating from January 1, 2008 to February 28, 2009. All of the stories originated from the Reuters news agency. Stories span multiple domains, though the domains are not annotated in the corpus.

Each story is composed of 3 elements: the creation date, the story title and the main text. Empirical evidence shows that the corpus contains near duplicate stories.

Dynamic Corpus We obtained a list of relevant documents from online search engines and retrieve the documents to obtain up-to-date information. To obtain the documents, one of several online search engines is used. These include Google News¹, Bing², and DuckDuckGo³. For each query, up to 100 documents are retrieved if enough relevant documents are available. The queries consist of causal markers, a news topic and temporal features to limit the results.

Data Selection

In this section we present the process to obtain the result sets. In the following section we present the process to obtain additional information to generate the evaluation unit used in the survey.

¹news.google.com accessed 1.1.2013

²www.bing.com accessed 1.1.2013

³duckduckgo.com accessed 1.1.2013

Although there are 1.8 million articles that can be processed, it is not necessary to process all the articles to obtain the 40 manually selected sentences. Initially we processed large batches of 100,000 articles, there were some memory limitations to processing batches of this size. after some processing optimization we were able to process batches of 10,000 articles. From one batch of 10,000 articles, the system extracts approximately 2,500 result sets. Many of the articles in the corpus are near duplicate, often the difference in near duplicates was the use of abbreviation, or the use of a synonym within the article, we filtered duplicates at a sentence level to reduce redundancy, and manually selected results based on the following criteria:

- A clear and short sentence, the sentence should not be more than 50 words long
- A causal relation between news topics defined by a causal marker, the causal relation established by the causal marker should be clearly evident.
- Both news topics that are causally related should be clearly defined and comprehensible.

The manual selection was to exclude miss-classified news topic references and to exclude wrongly parsed sentences. Many of the bad parses were due to non-readable characters included in the sentence. Some of the miss-classifications were due to bad sentence parsing, where the sentence would include the introduction text such as author and publisher. We also excluded near duplicates that the system did not detect at a sentence level. We excluded the results because they are not valid for the evaluation of the thesis.

We selected sentences that met the criterion, from a random sample of the results. An example sentence including the extracted news topic references, is given below.

Causal Sentence Gold in European trade soared to historic highs as speculators and investors snapped the metal up on the back of strong oil prices.

First News Topic Reference Gold in European trade soared to historic highs

Second News Topic Reference of strong oil prices

A total of one hundred and one sentences, including this one, were taken from a batch of 1,000 articles from the static corpus. These results were further processed to generate the collection of data used in the user survey. In the following section we explain how the 40 evaluation units are generated from these sentences.

Evaluation Unit

For the user survey we provide the users with sentences that encode causally related news topics, we also outline the news topic references. Additionally we provide a short text that is not causal, taken from an online news portal. The short text is provided to obtain user preference between causal and non-causal information, user preference is not within the scope of the hypothesis evaluation, yet it does provide an insight as to the usability of this type of information.

In this section we explain the generation of the 40 evaluation units including some of the challenges in obtaining valid material.

Evaluation units are based on results from *Forest* as previously described, this includes a causal sentence, and the first and second news topics that are causally related. In addition to this information we provide a short text segment, also known as snippet, that is non-causal. These snippets are often given along with the title of the article by online news portals. They are the short text summary about the content of the article, often they are taken from the first paragraph of an article where the main points, about the story, are given. For each evaluation unit we systematically obtain a short text summary, or snippet, about at least one of the two news topics, the news topic that is common to the causal sentence and the snippet is called the commonality.

An example evaluation unit is given below:

Causal Sentence Gold in European trade soared to historic highs as speculators and investors snapped the metal up on the back of strong oil prices.

First News Topic Reference Gold in European trade soared to historic highs

Second News Topic Reference of strong oil prices

Snippet Euro, Gold, Oil Surge to Record Highs - Naharnet Newsdesk
Naharnet - Mar 4, 2008 In late European trading, the dollar recovered slightly from its record low to trade ... alongside jitters over record *high oil prices* and rising inflation, ...

Commonality of strong oil prices

In this example we find the snippet taken from Google News, the process to obtain these snippets is described below. We also find the news topic reference *of strong oil prices* in the example, this news topic reference is partially defined by temporal features. While strong oil prices may occur in several points in time, this occurrence is limited by a time frame. Time frames are also used in the process to obtain the snippets.

To find non-causal comparison data, or snippets, we used a web-news search engine, namely Google News. The search query used for each result

set is the concatenation of the first and second news topic references. The queries were customized by limiting the dates for the results, so the results correspond to the time frame of the static corpus.

The criterion to select the snippet are the following:

- The snippet should not be causal, this allows us to compare causal to non-causal text.
- The snippet is the top ranked result from a news source.
- The snippet should have a common text to one of the news topics in the result set.

In the process of obtaining the snippets we found the following information. The query did not provide any results in 11% of the cases. The first result from Google News was not online news in 93% of the cases, instead the result is a relevant web page that is not from a news source.

In 36% of the cases the top result was a duplicate or near duplicate of the causal sentence, The comparison data is intended to be similar to the original, with the main difference that comparison data should not encode causal relations, therefore these results were not included. We consider that finding duplicates in the comparison data is a consequence of searching for similar data.

The duplicate or near duplicate results were excluded because they encoded causal relations. The commonality was manually assigned for each of the 40 evaluation units.

After removing the unwanted results we obtained 40 evaluation units, with the following structure:

Causal Sentence The sentence from the result set that was classified as causal and contains two news topic references in a causal relation.

First News Topic Reference The sentence segment that was classified as a news topic reference.

Second News Topic Reference The sentence segment that was classified as a news topic reference, found in a causal relation to the first news topic reference.

Snippet The text snippet provided from a news search engine⁴, to describe a news article.

Commonality The news topic reference that can be found in both the causal sentence and the text snippet.

In the following section we discuss how these evaluation units are used in the survey.

⁴news.google.com accessed 12.1.2013

6.1.2 User Survey

The survey was completed using an online survey system⁵, where contributors generated a total of 137 judgements. A judgement is when all the survey questions have been answered for an evaluation unit. We obtained at least three judgements for each evaluation unit. Each judgement is composed of 12 questions plus a comments section. There are two types of questions: validation questions and opinion questions. The validation questions are to confirm that the evaluation units are formulated correctly, the opinion questions help us prove or disprove the hypothesis. Opinion questions are also used to obtain user preference between causal and non-causal text.

Survey Questions

Here we present the validation and opinion question given for each judgement in the survey. There are four types of answers for the questions:

Trinary Where the possible answers are: *Yes*, *No* and *Unknown*.

Rating Where the answers are on a scale of 1 to 5, where 1 stands for: *No* or *Not at all*; and 5 stands for *Yes* or *Very much*.

Preference Where the possible answers are: *Text X*, *Text Y* or undecided; In this case *Text X* is a reference to the causal sentence and *Text Y* is a reference to the snippet.

Open Where the user can type any answer.

In the survey the terminology is simplified for the user. In the survey the news topic that is common to the causal sentence and the snippet is called *the subject*. The causally related sentence is called *Text X*, and the snippet obtained from an online news portal is called *Text Y*. This nomenclature is used in presenting the questions below.

There are five validation questions listed below, along with the type of answers that are possible for each question.

- Is the Subject in Text X? *Trinary*
- Is the Subject in Text Y? *Trinary*
- Is Text X a clear sentence? *Rating*
- Does Text X contain a causal relation? *Trinary*
- What type of relation does Text Y have, if any? *Open*

⁵<https://crowdfunder.com/jobs/159355>

We present these questions to confirm that the texts are comparable and suitable for the thesis evaluation.

There are seven opinion questions, three of the questions are regarding the hypothesis, the additional questions are to obtain user preference. the opinion questions are as follows:

- How helpful is Text Y to understand the subject? *Rating*
- Does Text Y give a context to the subject? *Rating*
- Which text about the subject do you like better? *Preference*
- Can you tell us why you like it better? *Open*

These questions are intended to give us a better understanding of user preference. The order of the questions presented here is not the order provided to the users, in the survey we first ask about the usefulness of the causal sentence, then the usefulness of the short summary, then we ask for the user preference.

The three thesis evaluation questions are as follows:

- How helpful is Text X to understand the subject? *Rating*
- Does Text X give a context to the subject? *Rating*
- Does the context help to understand the subject? *Rating*

Based on the results of these questions, we prove or disprove the thesis.

The last question in the survey is for comments, so the user can provide additional information. Users did not provide any relevant feedback in the comments question.

Survey Participants

Contributors were validated by answering gold standard questions, these are questions where the answer is known and the contributor must match the correct answer. Using the gold standard, users are given a trust value, and only users with a value above a predefined threshold contributed judgments to the evaluation. In Figure 6.1 we can see that 19 Users submitted 304 judgments. The evaluation is based on the judgements from three users, these three users are above the trust threshold and completed judgements for all evaluation units. Four additional users were trusted, yet they did not complete judgements for all evaluation units. The rest of the users were not trusted. The trusted users are those with at least 80% gold standard questions answered correctly.

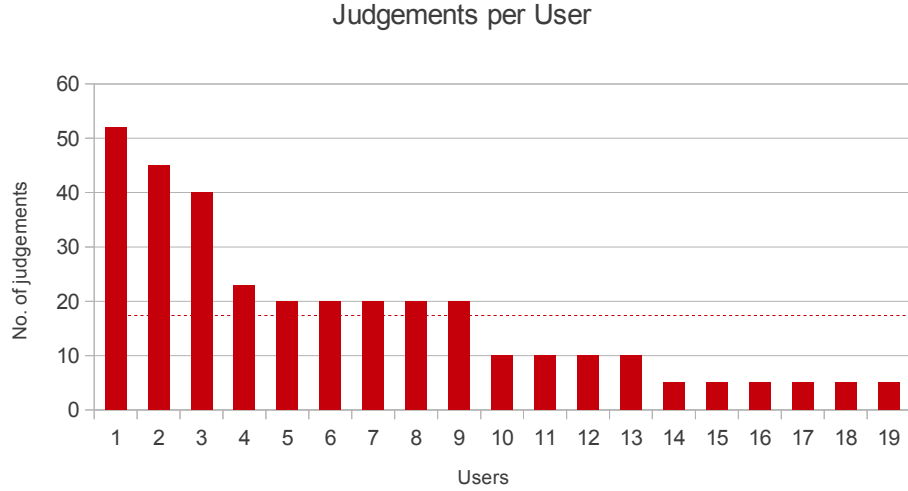


Figure 6.1: Numbers of judgments per User

6.1.3 Survey Results

The survey indicates that the first hypothesis *Causal relations between news topics provide a context that aids the user to understand the news* is valid. This is based on users giving a positive rating for the question “How helpful is Text X to understand the subject?” in 100% of the cases. In the question *Text X* refers to the causal sentence and *the subject* refers to a news topic. On average the rating for this question is 3.69 on a scale of 1 to 5, where 1 stands for *not at all* and 5 for *very much*. In contrast users gave a positive rating for the questions “How helpful is Text Y to understand the subject?” in 99% of the cases, where *Text Y* refers to the text summary and *the subject* refers to a news topic. The average rating for this question is 3.73 on a scale of 1 to 5, where 1 stands for *not at all* and 5 for *very much*. Therefore, on average the non-causal text is better for understanding, yet as we will see, the users prefer the causal sentence.

Users also indicate that causal relations between news topics do provide a context for the news. This is based on users giving a positive rating for the question “Does Text X give a context to the subject?” in 98% of the cases, where *Text X* refers to the causal sentence and *the subject* refers to a news topic. On average the rating for this question is 3.68 on a scale of 1 to 5, where 1 stands for *not at all* and 5 for *very much*. They also indicate that the context helps users understand the news. Based on users giving a positive rating for the question “Does the context help to understand the subject?” in 100% of the cases, where *the subject* refers to a news topic. The average rating for this question is 3.72 on a scale of 1 to 5, where 1 stands for *not at all* and 5 for *very much*.

Aiding in Understanding the News

The following graphs address the central point of the hypothesis, restated here *Causal relations between news topics provide a context that aids the user to understand the news*. We have broken down the hypothesis into basic components for evaluation, these components are:

- Causal Relations.
- News Topics.
- Information context.
- Aid in understanding.

The survey shows that 98% if the causal sentences were judged as causal by users.

In Figure 6.2 we can see that both causal sentences and text snippets are highly rated as providing a context to the news topic, with an average rating of 3.68 on a scale of 1 to 5, we can also see that the rating for context of the text snippet is lower than the rating for the causal sentence in the highest rank (5 *Best*), but in the rest of the ranks the text snippet is rated higher.

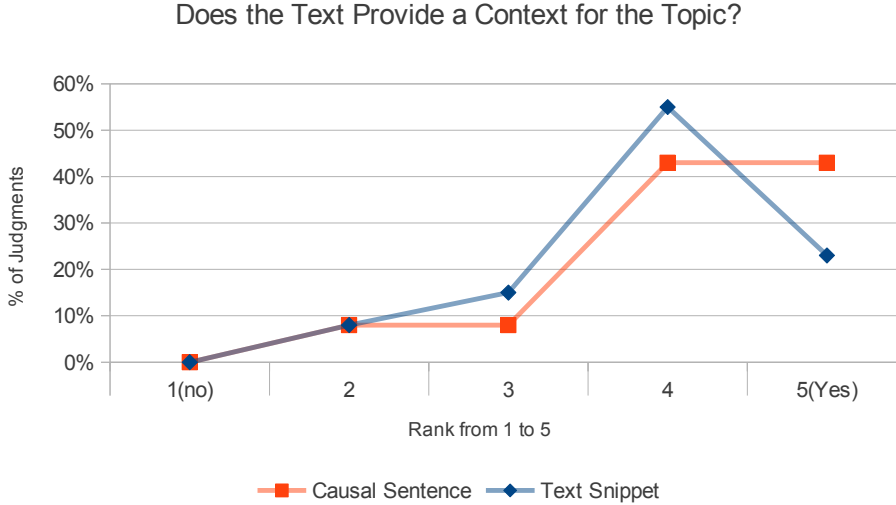


Figure 6.2: Rating distribution of the texts for providing a context.

Users have rated the text snippet as marginally better in aiding understanding, with an average rating of 3.73 versus 3.69 on a scale of 1 to 5. In Figure 6.3 we show user judgment distributions,

These ratings indicate that both results are useful for the users to understand a news topic and that both results provide a context for the news topics.

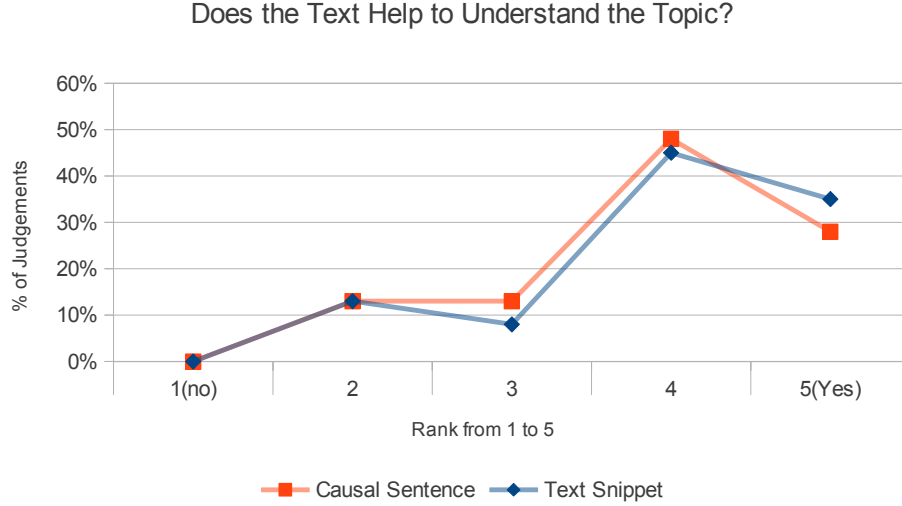


Figure 6.3: Rating distribution of text for aiding the understanding of news topics

The users positively confirmed that context does aid in understanding. In Figure 6.4 we show that most users view context as an aid in understanding. Based on these results we can state that Hypothesis I is true, and that the type of information provided by *Forest* is useful for users to understand the news.

Additional Findings

The evaluation also shows that users preferred the causal sentence over the text summary, and that users did not evaluate all the validation questions correctly.

In Figure 6.5 we see that the fraction of users that prefer the causal sentence over the text snippet, almost 33% of users preferred the causal sentence, compared to the 8% that selected the text snippet, though 59% of the users were undecided.

For the open question “Can you tell us why you like it better?” the main reason for the preference is the clarity of the text, the top two reasons given by users are “*it’s clear*” and “*clearer answers*”. Though many users provided clear responses, some provided non-legible or no answer to this question.

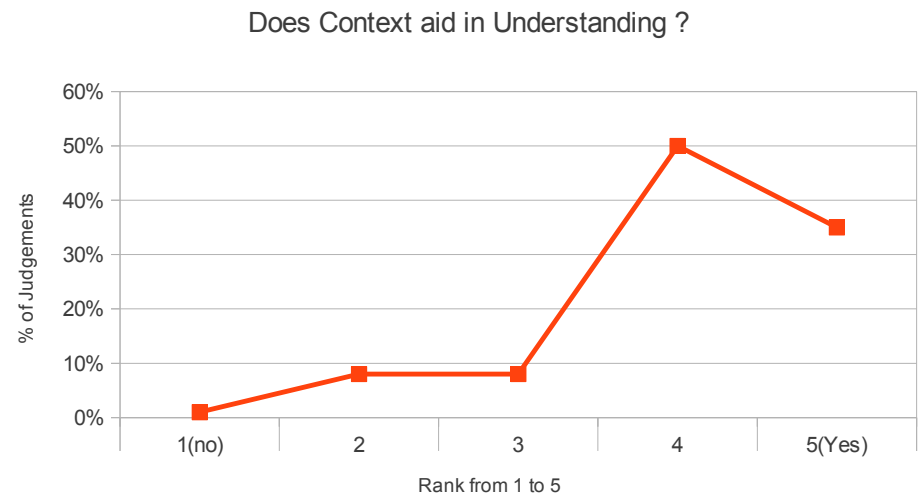


Figure 6.4: Rating distribution of user consideration of context in aiding understanding of news topics

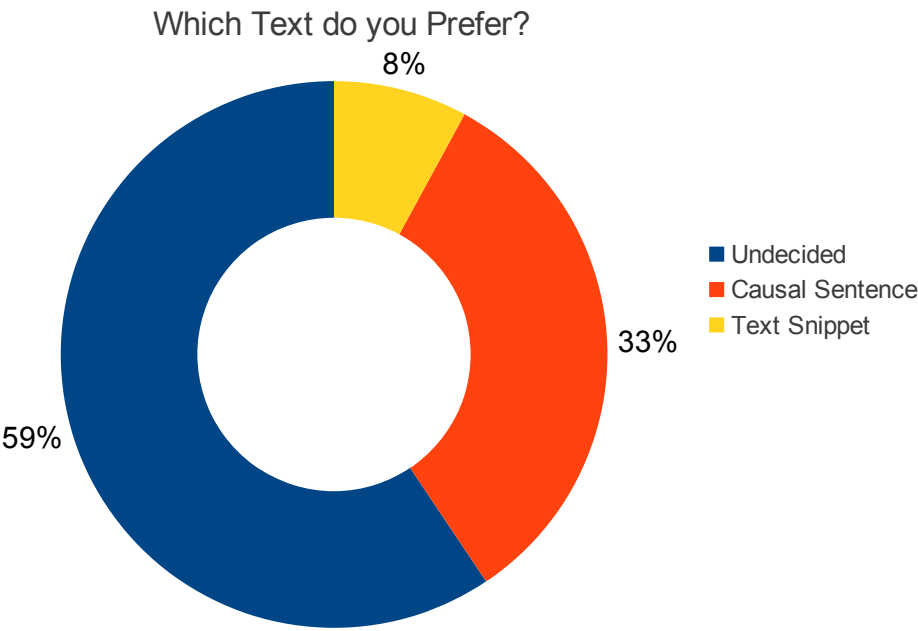


Figure 6.5: Distribution of judgments showing user preference

Question	Agreement (%)
Is the Subject in Text X?	78.34
Is the Subject in Text Y?	66.25
Is Text X a clear sentence?	41.86
Does Text X contain a causal relation?	74.04
What type of relation does Text Y have, if any?	18.28
How helpful is Text X to understand the Subject?	39.88
How helpful is Text Y to understand the Subject?	38.75
Does Text X give a context to the Subject?	39.43
Does Text Y give a context to the Subject?	39.51
Does context help understand the Subject?	42.3
Which text about the Subject do you like or prefer?	54.05
Can you tell us why you like it better?	22.91
What type of relation does Text Y have, if any?	18.28

Table 6.1: Table with user agreement per question.

Accuracy and Agreement

In this section we review the agreement between users and user accuracy evaluated by the gold standard. Excluding open questions, user agreement was approximately 50%, this includes ranking and multiple choice questions. User accuracy, measured using the gold standard, was on average 95%. Users with low accuracy received low trust scores, therefore contributors' judgments were not included in the results. Untrusted contributors accuracy is on average 8%.

In Table 6.1 we present the agreement values for each question. A high agreement value indicates that users frequently gave the same answer to a question, for example if all the users answered "yes" to a question the agreement is 100%. This agreement is calculated by averaging the percentage of users that gave the same answer to a question. Low agreement values may indicate that the question is difficult, ambiguous or subjective. Additionally, ranking (from 1 to 5) or open questions may present a low agreement due to minor differences.

In Table 6.1 we can see that open questions have an agreement of approximately 20%, this can be considered low. We can also see that basic quality questions have a high agreement, approximately 75%. The thesis questions are answered in the form of a ranking from 1 to 5, for these questions the agreement is 40% this can be considered medium to high agreement, based on the multiple possible answers.

6.1.4 Conclusion

The survey results show that the thesis is valid and that the results from the system are useful. the survey was completed by multiple users on an online survey system, where users and user responses were evaluated.

6.2 Hypothesis II Evaluation

The second thesis, *At least 95% of news topics can be found in a causal relation to another news topic*, is intended to discover if the desired information is available for extraction. In the first thesis we found that causal relation information is useful for users to understand the news, for this thesis we want to find out if the information is available, in the next thesis we will show how we can extract this information.

Our analysis shows that approximately 25% of news topics can be found in a causal relation to another news topic. We also found that the 10% most frequent news topics have a 94% probability of being found in a causal relation to another news topic. These results are based on a sample of news articles and a sample of news topics. For news articles that are available online, we took our static corpus as a sample. For news topics we took a collection of validated news topic references as a sample. We then extracted results from these samples to obtain an insight about the news topics, particularly the characteristics of a news topic that is in a causal relation to another news topic.

The evaluation will show that the amount of articles per news topic follows a long tail distribution, where the top 10% news topics account for 89.2% of all the news articles. This is relevant because there is a strong correlation between the amount of articles for a news topic and the amount of causal sentences for that news topic. For instance if we take half of the news topics, those that occur less frequently, for those news topics there is a probability of 3.44% of being found in a causal relation to another news topic.

We conclude that causal sentences are not generated for more than 95% of news topics and that there is a correlation between how frequent a news topic appears in news articles and the likelihood of that news topic to be found in a causal relation to another news topic. These conclusions disprove the hypothesis, yet they allow us to estimate the news topic that may be found in a causal relation.

6.2.1 News Topic References in Corpus

During our analysis we found 224 news topic references in the corpus, the total number of validated news topic references is 1,224, these are phrases

we have verified as news topics. The reason for finding only a fraction of the validated news topics in the corpus is a temporal mismatch between the validated news topic references and the corpus. The validated news topics range from January 1, 2000 to August 8, 2012; and the corpus contains news articles from January 1, 2008 to February 28, 2009.

In the corpus we found 932,187 sentences that included at least one of the 224 news topics. these sentences were found in a total of 64,769 articles. The distribution of occurrences for these topics can be seen in Figure 6.6.

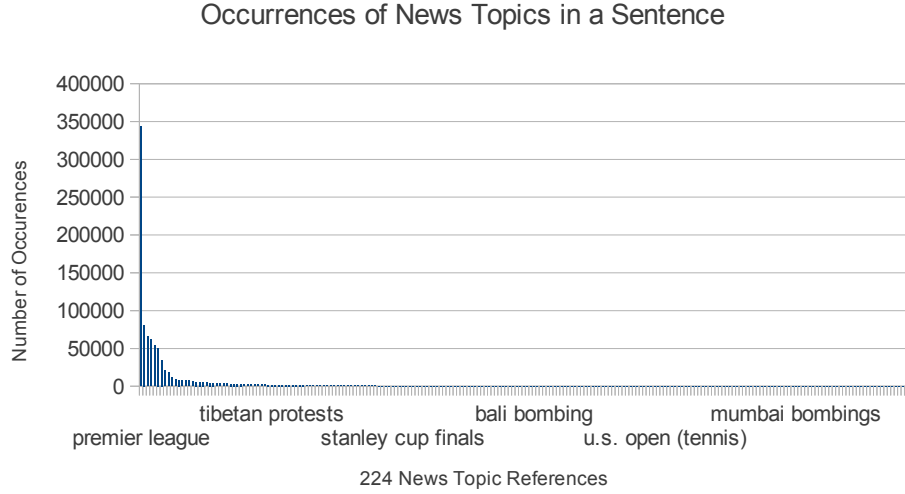


Figure 6.6: Number of occurrences of news topics in a sentence from the static corpus.

The figure shows that the maximum number of occurrences of a news topic is 343,766, while the median number of occurrences is 265.5. The distribution shows that the top 2% news topics have more occurrences than the following 98%. This may be explained by the way news articles are generated, where articles about novel news topics are generated constantly, this makes the long tail; and articles about popular news topics are generated more abundantly, this makes the top 2% of news articles.

Current News Topics in Dynamic Corpus

As a reference we also present the frequency distribution for a small number of current news topics. In this case the documents were obtained from an online news portal. We call this the dynamic corpus, for news topics that users considered to be popular in the news at the date of extraction⁶.

⁶February 20, 2013

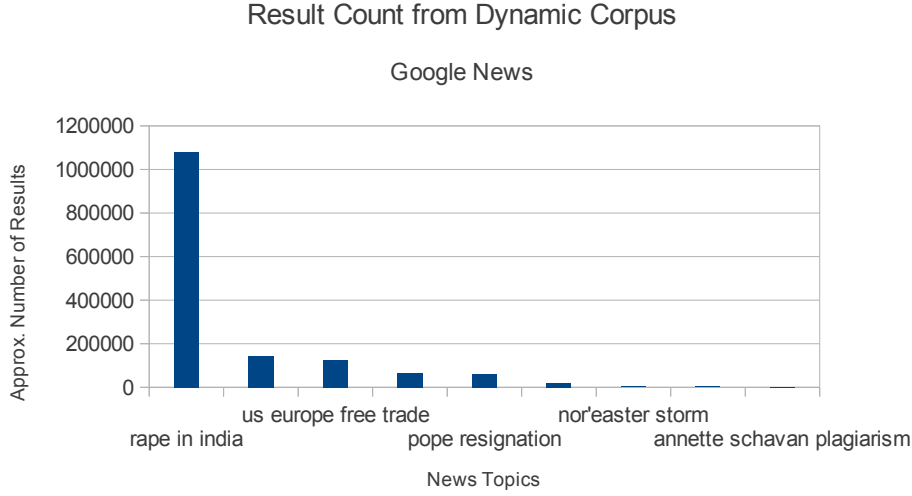


Figure 6.7: Approximate number of news articles found for current news topics.

In Figure 6.7 we can see that the distribution is similar to that of the static corpus. Where the top 1% news topics have more occurrences than the following 99% news topics. The approximate maximum number of results for a single news topic is 1,080,000, the median is approximately 62,100.

In the following sections we will show how the number of occurrences of a news topic is correlated to finding causal information about that news topic.

6.2.2 Causally Related News Topic References in Corpus

Based on the 932,187 sentences that included at least one of the 224 news topics, we extracted 12,040 sentences that also contained a causal marker. We then extracted the news topics from these causal sentences and found 159 news topics. It follows that out of the 224 news topic references found in the corpus, only 159 were also found in a causal sentence.

We processed the 12,040 causal sentences to find causal relations between news topics, we found that 56 news topics were causally related to another news topic. It follows that out of the 224 news topics found in the corpus, only 25% of them were found to be causally related to another news topic.

This indicates that while approximately 71% of news topics have causal information, only 25% of the news topics have causal relations that can be used by *Forest* because the scope of this system is focused on causally correlating news topics, not causal information in general.

Correlation between Frequency of Occurrence and Causal Information

In our analysis we found that there is a correlation between the frequency with which a news topic appears in a news article and the frequency of that news topic being found in a causal sentence. The correlation is calculated using the Pearson index⁷. The correlation index is 0.65, with a sample size of 57 units. A Pearson index value of 0 indicates no correlation and a value of 1 indicates a complete linear correlation, values above 0.50 are considered strongly correlated.

In Figure 6.8 we present the distribution of occurrences of news topics that are found causally related to another news topic.

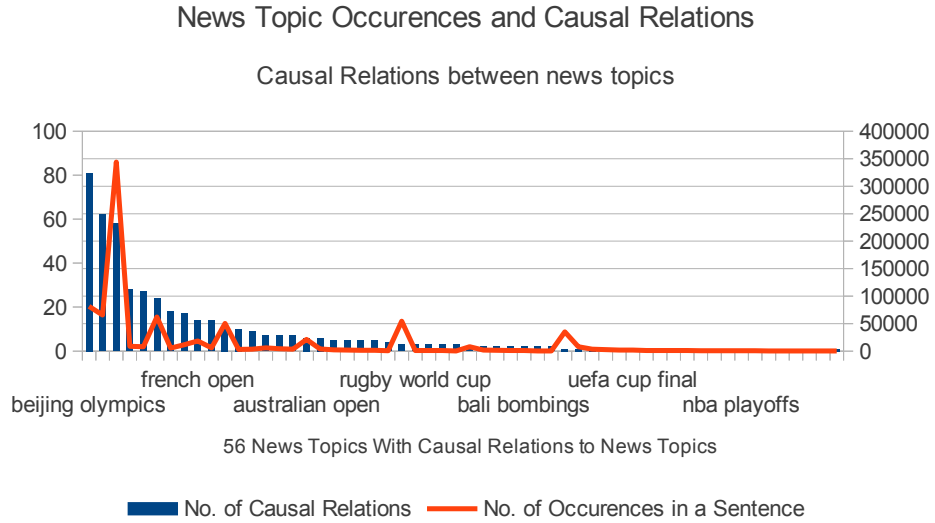


Figure 6.8: Frequency distribution of news topics in causal sentences.

In the figure we see that the maximum number of causal relations to another news topic is 81, and that the maximum number of occurrences of a news topic in a sentence is 343,766. The figure shows that news topics that occur more frequently also have a higher number of causal relations.

In Figure 6.9 we can see all the news topics found in the corpus, the news topics are ordered by frequency of appearance, this is to say the most frequent are on the left and the least frequent are on the right. Following this order we calculated the news topics that are in a causal relation in the following way: We divided the number of causally related news topics that have appeared so far by the total number of news topics that have appeared so far. To illustrate, consider the 20 most frequent news topics,

⁷http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

these are the 20 left most news topics. From these 20 news topics only 17 have been found causally related to another news topic, therefore we calculate the percentage as 17 divided by 20 or 85%.

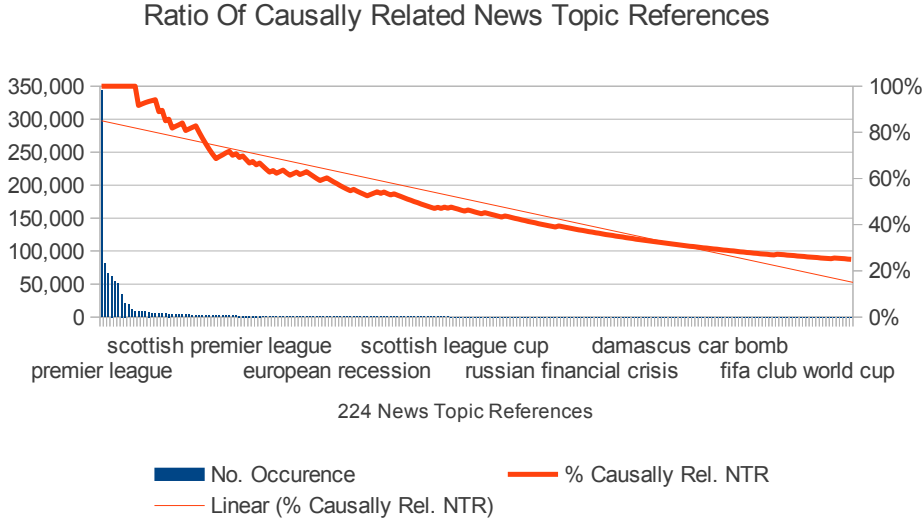


Figure 6.9: News topics ordered by frequency with percentage of causally related news topics and trend line

The graph shows that it is very likely to find a causal relation for the most frequent news topics, and that causal relations are found for 25% of the news topics in general. In the graph we can see that among the first 5% most frequent news topics, 100% of them are causally related. We can also see that among the first 10% most frequent news topics, 85.71% of them are causally related. We can see the trend line that shows the trend of finding causal relations decreasing as the news topics are less frequent.

6.2.3 Conclusion

This evaluation, based on sample data, showed that we can not find causal relation information for 95% of the news topics, it shows that there is causal information for approximately 70% of the news topics, and that the causal information is within the scope of this work in 25% of the news topics. The scope of this work is to find causal information that relates news topics. We also found a correlation between the frequency with which a news topic is mentioned in the news and the number of causal relations found for that news topic. The correlation is measured with a Pearson index of 0.65, this is highly correlated. Therefore we can state that it is highly likely (95% probability) to find a causal relation for the 8% most

frequent news topics.

6.3 Hypothesis III Evaluation

For the evaluation of the hypothesis *By analyzing news articles at a sentence level it is possible to extract explicit causal relations between news topic references with a 95% accuracy*. We employed a user survey⁸, where we asked users to validate the correctness of the extractions. We obtained a 94% accuracy in the extractions. To obtain this level of accuracy we limited the amount of results provided to the user, by ranking the results and providing only the top 5 results for each news topic. The results were generated using nine current news topics. These are news topics that have occurred recently. The news topics were used in the dynamic corpus, this is to say we used an online news portal to obtain 576 of the relevant news articles. We used current news topics and the dynamic corpus to simulate a real-world use case, the main use case of the system. In this use case a user queries the online system to obtain the relevant information.

Results also show that users find the automatically extracted information useful, rating it on average 6.34 on a scale of 1 to 10, where 1 is not useful and 10 is very useful. Though we also found that the usefulness rating is not correlated to the ranking provided by the system, presenting a Pearson index of less than 0.05 which is not statistically significant based on the small sample size. It follows that we can not estimate what results are useful for the user.

Before we began the survey, we addressed some of the challenges found during the previous evaluation, namely selecting the most relevant results automatically and selecting news topics where causal information may be available.

To select the most relevant results automatically, we developed the rating system, described further in this chapter. To select news topics where causal information may be available, we requested six participants to provide a popular news topic. The news topics mentioned are the nine news topics used to develop the causal information. To provide the reader an intuition about the performance of the system regarding all the results, not only the top ranked results, we present a theoretical evaluation where we focus on the precision of the extraction. We focus on the precision of the extractions because we assume that users would prefer fewer correct results than many potentially erroneous results.

⁸<https://crowdfunder.com/jobs/172977>

Component	Precision	Recall	$F_{0.5}$
Sentence extraction (SE)	99%	99%	99%
Causal relation Extraction (CE)	88%	98%	90%
News topic reference extraction (NTRE)	97%	94%	96%

Table 6.2: Performance metrics for the main components of *Forest*

6.3.1 Theoretical Evaluation

To generate the theoretical results we used the performance metric of each individual component of *Forest*. The main components are: sentence extraction from the news article, causal relation extraction and news topic reference detection. In the practical evaluation we also consider the source of the news articles, in this case we do not consider the source because the metrics were generated with the static corpus, that does not change. In Table 6.2 we present the precision with the highest $F_{0.5}$ score, for causal relation extraction and for news topic reference extraction. For sentence extraction we provide the available F_1 measure.⁹ The $F_{0.5}$ measure puts more emphasis on precision than on recall.

Based on precision values that correspond to the highest $F_{0.5}$ score, we calculate the theoretical precision of the system. The calculation uses Equation 6.1, given below.

$$p(Forest) = p(SE) \cdot p(CE) \cdot 2p(NTRE) \quad (6.1)$$

In this equation the overall system precision $p(Forest)$, is given by the product of of each component precision: The sentence extraction precision $p(SE)$, The causal relation extraction precision $p(CE)$, Twice the news topic reference extraction precision $p(NTRE)$, because news topics are extracted once for the cause and once for the effect in a causal relation.

It follows that the precision of *Forest*, when there is an emphasis on the precision, is approximately 82%.

These results indicate that obtaining a high level of precision for all results is challenging. In the *Forest* use case, only the top results are provided to the user. The selection of the top results is to minimize complexity for the users, thus minimizing the cognitive load.

In the following section we evaluate the performance of the system using top ranked results. This allows us to obtain an intuition of the performance of the system in a real-world use case.

⁹The F_1 measure provided instead of the $F_{0.5}$ because of unavailability

6.3.2 Practical Evaluation

In this evaluation we present a configuration based on the main use case, where users submit a news topic and the system provides the top results for that news topic. The evaluation shows that *Forest* is able to achieve an accuracy of 94% for a defined scope of news topics, namely topics that are prominent in the news, where the results are limited to the top N results, in this case the top 5 results for each news topic.

6.3.3 Dataset Generation

To generate the dataset the system is configured to generate the results automatically. First, a list of news topics is provided to the system, these are the nine current news topic references. Then, the system retrieves 576 relevant news articles from the dynamic corpus. Finally, the system processes the articles and ranks the results. These 45 results are used in the survey.

Current News Topic Selection

To obtain the initial news topic references, several individuals were queried for a current news topic. We gave the query “tell me something that is popular in the news right now?...with few words”. The informal format was intended to illicit natural, straight forward responses from the individuals. The responses had minimal preprocessing to be used as news topic references, for example we removed the beginning “about the”, and replaced verbs with the simple form, for example “arresting the pope” was changed to “pope arrest”. the query was made on February 20, 2013. The responses provided are listed below:

- Annette Schavan plagiarism
- Horse meat contamination scandal
- Japan trade deficit
- Nor’easter storm
- Pope arrest
- Pope resignation
- Rape in India
- US Europe free trade
- US Germany agreement

These responses reflect current news topics at the time.

Result Generation

To generate the results the system was configured to use the dynamic corpus to retrieve news articles. To rank the relevant articles, an online news portal was used, namely Google News. 64 articles were retrieved for each news topic, the articles were the top ranked articles from the online news portal. The natural language content for each article was extracted and split at a sentence level. The sentences are then classified as causal and a confidence score is assigned to each sentence. In the following section we explain how the confidence score is used to rank the results. The sentences classified as causal are then processed to find the causally related text snippet, the text snippets are processed and classified as news topic references, this classification also issues a confidence score.

At the end of this step, we have a collection of approximately 900 sentences, for the nine initial news topic references. The sentences have been classified as causal and as relating two news topic references. Data linked to each sentence includes source information and confidence scores.

Top Ranked Results

To simulate a real world scenario, where only the top results are provided to the user initially, we rank the results and provide only the top five ranked results.

The ranking of each sentence is based on the following values:

- The probability index of the causal marker, this is the probability of the causal marker to be in a causal relation.
- The confidence score of the causal sentence classifier.
- The similarity score between the initial news topic reference and the extracted news topic reference.
- The confidence scores of the extracted news topic references.

Based on empirical evidence we weighted each of the values mentioned above to generate a rank value. This rank value is used to sort and select the results. The weights reflect the relevance of the information found in the sentence. To illustrate consider the following. When providing a results it is important to find an unambiguous causal relation in the sentence, yet it is more important to find the relevant news topic in the sentence. Therefore, the weight of the news topic similarity score is higher than the weight of the causal sentence confidence score.

Using the ranking values we selected the 5 top ranked results for each initial news topic reference. These top ranked results are then used in the survey for annotators to evaluate the system.

A sample result is presented below:

Initial News Topic Reference annette schavan plagiarism

Source <http://www.panarmenian.net/eng/news/145123/...>

Sentence Merkel said she had accepted the resignation of Annette Schavan “with a heavy heart” after her former university stripped her of her doctorate , saying she had plagiarized parts of her thesis ”Person and Conscience” 33 years ago.

Cause resignation of Annette Schavan” with

Effect her former university stripped her of

Rank value Rank:2 (value 0.56)

In the survey the terminology was changed for the user, the terms used in the survey are listed below:

Initial News Topic Reference The Query

Source The Source

Sentence The Sentence

Cause TextA

Effect TextB

The rank values were not provided to the user, therefore they are not listed. This nomenclature is used in the survey.

6.3.4 User Survey

For the user survey we used an online survey system¹⁰, where users provided their opinions about the system results. Due to the objective nature of the questions and the openness of the survey, many invalid results were detected and discarded, namely evaluations that showed no variation in the answers, this is to say evaluation sessions where all the rankings were 1 and all the answers were yes. We believe that the evaluation session with no variation do not reflect user opinion, but were systematically generated to complete the task.

There are several types of answers for the questions:

Trinary where the possible answers are: *yes*, *no* and *unknown*.

Selection where the possible answers are: *Yes*, *TextA caused TextB*; *Yes*, *TextB caused TextA*; *Yes*, *it is causal but not TextA and TextB*; *No*, *there is no causal relation*.

Rating where the answers are on a scale of 1 to 10, where 1 stands for: *not likely* and 10 stands for *very likely*.

Ranking where the answers are from *1 Best* to *5 worst*.

¹⁰<https://crowdfunder.com/jobs/172977>

Multiple where the possible answers are: *Text A*, *Text B* or *Undecided*;
Open where the user can give any answer.

Below is the list of questions provided to the users, with the type of answers that are possible:

- Is the sentence clear? *Trinary*
- Is the sentence about the query? *Trinary*
- Is the sentence causal? *Selection*
- Is textA an event or activities? *Trinary*
- Is textA something that might be mentioned in the news? *Rating*
- Is textB an event or activities? *Trinary*
- Is textB something that might be mentioned in the news? *Rating*
- Does the sentence help you understand the query? *Trinary*
- Does the sentence help you understand the query? *Rating*
- There are 6 sentences for this query, can you rank this one? *Ranking*
- Would more causes and effect of the query help you understand it? *Trinary*
- Would more causal sentences be useful to understand the query? *Rating*
- Does the sentence say what are some causes or effects of the query? *Trinary*
- Does the sentence state some of the causes or effects of the query? *Rating*
- Any comments? *Open*

There are three types of questions in the survey.

First, questions to insure that the result is well formulated. Specifically, that the result is a clear sentence, that it is about the query news topic reference, and that it is causal. We also confirm if the cause and effect texts can be defined as news topics, namely, if it is an event that might be mentioned in news. Because this last question is very subjective we repeat the question in a different format, first, the possible answers are *yes*, *no* and *unknown*; then the answer is a rating from 1 to 10 where 1 is *not likely* and 10 is *very likely*.

The second type of questions are to obtain the user opinions about the results. The opinions about the usefulness of the results for understanding the news topic, and about the completeness of the results, as in being a sentence that states causal relations for a news topic. These questions are given in two formats: as a close ended question, where the possible answers are *yes*, *no* or *unknown*; and as a rating scale, where the scale ranged from a negative answer of 1 to positive answer of 10.

The close ended questions were intended to facilitate the decision process for the user, then the scale response was intended to obtain a more fine grained value.

The third type of questions are supplementary question, to discover if the users have enough results for each news topic, and if the users see multiple results for a single news topic as a single collection. And finally any comments the user may have.

Survey Responses

We obtained a total of 598 completed surveys from 23 users, from these surveys 90 presented variations in the results, the rest were not included in the analysis and we assume they are invalid. These 90 valid results were generated by nine users. We assume that the rest of the users did not contribute valid evaluations because there was no consequence to invalid participation. Users were only instructed to *Evaluate the performance of a system to extract information from online news, by providing your opinion and ranking of the text snippets.*

6.3.5 Result Analysis

Results show that the system performs with a high accuracy, when providing the top ranked results. The results also indicate that users consider the extracted results useful. The system achieved a 94% accuracy for extracting sentences with a causally related news topics, given that only the top five results are provided. The usefulness rating of the automatically generated results present a similar distribution to the usefulness rating of the manually selected results presented in the evaluation of Hypothesis I, Figure 6.3.

System Accuracy

The system accuracy is evaluated with the answer distribution for the trinary question *does the sentence say what are some causes or effects of the query*

Users were also instructed to provide a rating value regarding how well the sentence presents a causal relation to the initial news topic. The specific question is *Does the Sentence state some of the causes or effects of the Query?*, with possible rating from 1 to 10, where 1 indicates *No, not at all* and 10 *Yes, a lot*.

In Figure 6.11, we can see the rating distribution from the users, This distribution of similar to the usefulness rating in Figure 6.13 given in the next section.

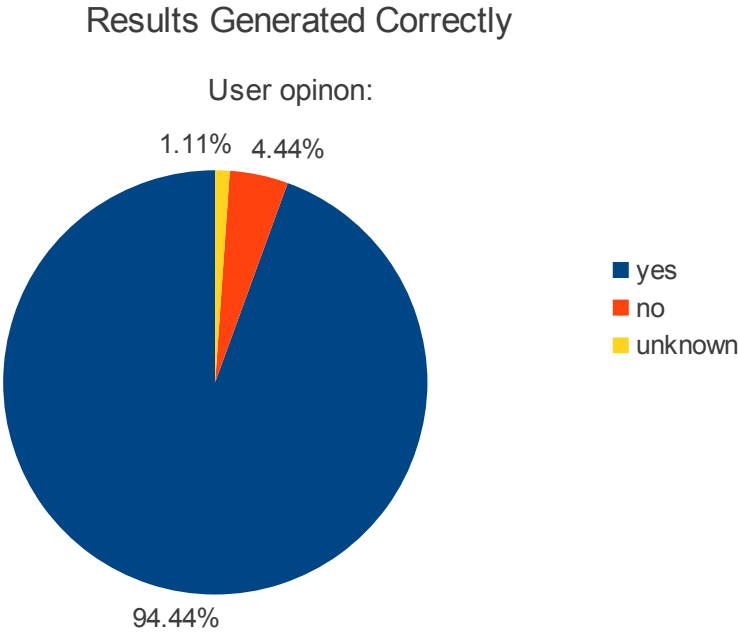


Figure 6.10: User opinion on accuracy of extractions

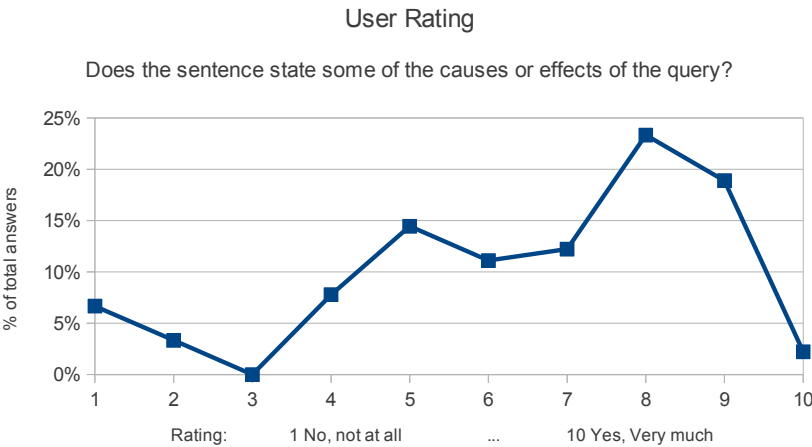


Figure 6.11: User opinion distribution about the accuracy of extractions

Usefulness Rating Distribution

We also asked the user for a rating to the question *does the sentence help you understand the query*. This question is a follow up on Hypothesis I, where we ask users if the the information in the sentence is useful for understanding.

First we present the results for the closed question, in Figure 6.12

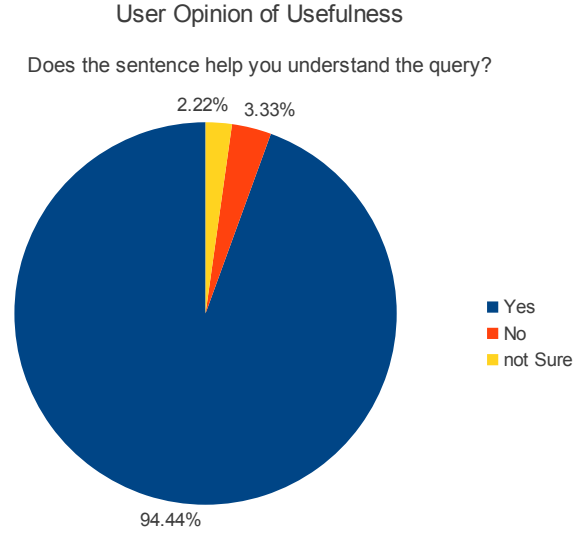


Figure 6.12: Trinary user classification of causal sentence usefulness

In Figure 6.13 we can see that the distribution is similar to the distribution found for Hypothesis I evaluation in Figure 6.3, except for some incorrectly generated results, the distribution spikes close to the three-quarter rating.

The usefulness rating distribution is also similar to the accuracy rating distribution, this indicates that results that the users consider correct are also useful.

6.3.6 Conclusion

Results show that the system is able to achieve an accuracy of 94% using the top ranked results, the results are useful for understanding, with an average rating of 6.34 on a scale of 1 to 10. Given these results imply that the hypothesis *By analyzing news articles at a sentence level it is possible to extract explicit causal relations between news topic references with a 95% accuracy* is true given certain conditions.

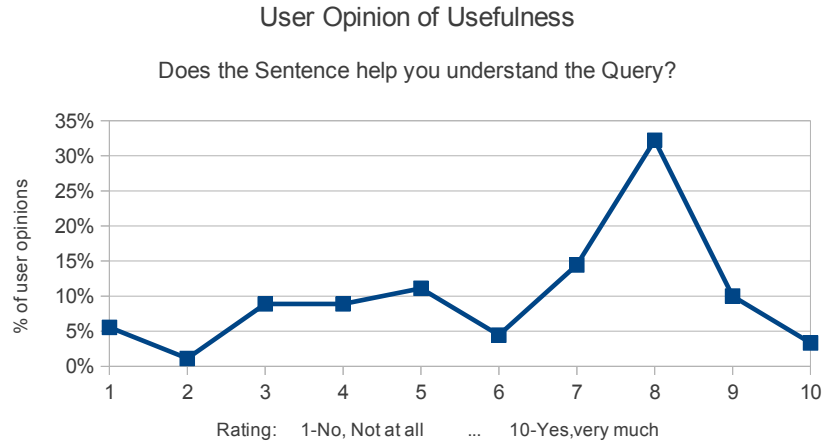


Figure 6.13: User opinion distribution about usefulness of causal sentence for understanding a news topic

6.4 Conclusions

In this chapter we have presented the evaluation of the three hypotheses presented in the introductory chapter, we show that causal information is useful for understanding the news, that we are not able to extract causal information for all news topics, and that by presenting only top ranked results we can achieve a high accuracy in the results.

Chapter 7

Conclusions

The overwhelming amount of online news presents a challenge called news information overload. To mitigate this challenge we propose a system to generate a causal network of news topics. The causal network has some advantages over currently used techniques, such as text-similarity clustering, for relating news articles. First, the news topics are put into a context the user may already know, thus facilitating understanding. If the causally related topics are unknown to the user, these novel topic may subsequently be of interest, therefore novel interesting material may be presented. Also, it may be easier for a user to navigate the news on a causally related network instead of a traditional list-column layout.

To generate the aforementioned network, several components are necessary. These include a component to extract causal relations from news article text and defining a news topic in a way that can be discovered in causally-related text. All this while maintaining journalistic features such as source reference and author attribution, that allows the user to verify the information. We have developed these components in a system we call *Forest*.

To illustrate *Forest*, consider a user interested in reading the news. The user finds a novel news topic, namely *The Dodd-Frank Act*. To get a quick intuition about the topic the user generates a causal network for *The Dodd-Frank Act*. The generated causal network shows that *The Dodd-Frank Act* was created in response to *The Late-2000s Financial Crisis*, that was in turn cause by *The Subprime Mortgage Crisis*. Because the user already knows about some of the causally related topics beforehand, the novel topic gets an interesting context. The causally related topics the user does not know about have already past, but now they become newly interesting because of the novel context. For the above illustration the causal chain of news topics is as follows: *The Subprime Mortgage Crisis* caused *The Late-2000s Financial Crisis* which in turn caused *The Dodd-Frank Act*.

This example presents some of the assumptions we have made; We assume that the causally related news topics help the user understand the news. Therefore, we have evaluated if users find this type of information useful. We also assume that because of the abundance of news we can extract this kind of information for most if not all news topics, so we verified if this type of information is available. Another assumption is that we are able to extract causal relations between news topics with an accuracy that is acceptable to users. Consequently, we tested the accuracy of our system to extract this type of information. These assumptions are the basis for the hypotheses evaluated in this work.

To evaluate if causal relation information is useful for users to understand the news, we used a survey where users gave 137 judgements about the usefulness of this type of information, users rated it positively, giving it a rating of 3.69 on a scale of 1 to 5, where 1 is not useful and 5 is very useful. To gain a perspective about these results we also asked users to rate non-causal information, in this case a short text summary obtained from a popular online news portal, namely Google News. Users rated this non-causal information at 3.73 on the aforementioned scale. We also asked users if they prefer the causal or non-causal information, only 41% of users provided a response, but from those 80% preferred the causal information. We concluded that this type of information is useful for users to understand the news, and that the results are as useful as the alternative.

To evaluate if causal relation information is available for a large majority of news topics, we analyzed a selection of 224 news topics found in a sample of 64,796 news articles. We found that approximately 25% of the news topics are causally related to another news topic. We also found that there is a positive correlation between the number of times a news topic appears in the news and the likelihood that the news topic is causally related to another news topic. For example, the 10% most frequent news topics have a 94% likelihood of being found in a causal relation, while the least frequently occurring news topics, 50% of all news topics, have a 3.44% likelihood of being found in a causal relation. This analysis indicates that causal relation information is not available for most news topics, and it also indicates that causal relation information is likely available the news topics with high occurrence frequencies.

To evaluate if we can accurately extract causal relation information, we used another survey. In this survey we asked users to validate if the results were generated correctly. By providing users only the top five results, based on the systems ranking, we were able to obtain an accuracy of 94.44%. The survey shows that users found the results useful, rating them at 6.34 on a scale of 1 to 10, where 1 is not useful and 10 is very useful. Based on 90 judgements given by nine users, we can conclude that the system performs at a level users find acceptable.

Given these three evaluations we can conclude that causal relation information between news topics is valuable for users, it helps users to understand

the news and it can correlate past news topics to current news topics, potentially placing the news topics in relation to a news topic that is known. Though we cannot extract this type of information for all news topics, it is likely that we can extract this information for the most popular news topics.

These evaluations and conclusions are based on results generated by *Forest*, the system we developed to extract sentences that encode causal relation between news topics. As the description indicates the main components of the system are causal relation extraction and news topic extraction. In our related work research we found multiple systems that performed these tasks. Yet, there were gaps that we needed to fill for our system to perform as intended.

Some of the works that dealt with linking news topics are: *Incident Threading for News Passages*[Feng and Allan, 2009] and *Causal Network Construction to Support Understanding of News*[Ishii and Ma, 2010]. Both of these works were based on research done for a seminal project called Topic Detection and Tracking[Allan, 2002]. *Incident Threading for News Passages* gave us an example of relating multiple news articles. In this case the news articles were about a single news topic and the “threading” was between events that happened in the news topic, we were searching for work that linked different news topics, where the link was specifically causal. In *Causal Network Construction to Support Understanding of News* we found causal relations, but they were in the Japanese language and the approach could not be adapted for use in English. Other drawbacks of this work include the loss of source information, This is to say that when topics are extracted the source article was no longer correlated to the extracted information, this was mainly due to the model used to represent news topics, namely a term vector.

Based on the drawbacks we found, we began researching causal relation extraction, we found that there are four main types of approaches, including one we developed.

The causal keyword approach, where a sentence was analyzed to contain a causal term, such as “because” or “due to” and if these terms were within a given pattern the sentence was classified as causal. An example of a pattern could be finding nouns before and after the keyword “because”. These nouns would be classified as causal actors. This approach presents the challenge of finding the causal keywords.

The causal actors approach, where known causally-related actors are used to find the patterns that encode a causal relation, for example the actors “HIV” and “AIDS” are causally related, therefore we use them to find causal patterns, for example in the sentence “*HIV* leads to *AIDS*” we can extract the causal pattern “*actor 1* leads to *actor 2*”. One of the main challenges of this approach is that it produces ambiguous results, this is to say that the causal relations may not be explicitly stated, they may be implied, this

may lead to miss-classification.

The third approach is a hybrid of the last two approaches, where keywords/patterns are used to find actors and then actors are used to find more keywords/patterns. The approach may also begin with the causal actors. These three approaches present the same challenge of over-fitting to a given domain. Though by specializing on a domain they are able to achieve an F_1 score of 0.81.

To overcome this challenge we developed a machine learning approach that uses causal keywords to preselect sentences, then we use N-Gram features to classify the sentences as causal. In the development of this approach we evaluated several types of features used in several classification techniques. The features including part-of-speech tags, verb tense and negation or question terms. We used Chi-square feature selection and backward feature elimination to select the most relevant features. We then implemented six well-known machine learning approaches using the selected features when applicable. We also implemented a text classification approach developed within our research group. We found that our text classification approach performed better than the other machine learning approaches. Using five fold cross validation we confirmed that our text classification approach obtained an F_1 score of 0.78. We also cross validated the best performing machine learning approach and we found that the Naïve Bayes classifier obtained an F_1 score of 0.72. One of the contributions of this work to the research community is this causal relation classification approach along with the dataset generated to develop the approach.

Another contribution to the research community is the development of news topic references as a topic model. In our related work research we found that the models for topics were not very straight forward to understand. For example, in Latent Semantic Indexing a topic is modeled as a term vector. And while lists of words with a scalar value are useful to cluster document into topics, they are not straight-forward for users to understand. This issue was addressed in the field of multi-document summarization, where an approach called Topic Themes and Signatures uses sentence fragments to define one or more phrases to represent a topic. Some of the issues of this approach is that it requires a document collection about a topic to extract the topic phrases. We use a machine learning approach to classify text as a news topic reference, which allows us to process text without previously requiring a document collection about the topic. During the development of the machine learning approach we defined news topic references and generated a dataset for evaluation. This dataset, definitions and the machine learning algorithm could be used by researchers to further develop news topic models.

7.1 Outlook

There are many directions that can be taken for the further development of *Forest*, for example making the results of the system openly accessible to users, the results from this system may help users take a different perspective of the events that surround them and users may also gain a perspective on the importance of events by understanding how influential these events become over time. These changes in the perceptions of news may lead to more critical thinking about past and present events, this may also affect the way news is written, making news an event in a continuum instead of an isolated happening.

Some of the open challenges of *Forest* include improving the classification of cause and effect in a causal relation. Our classification approach can classify a sentence as causal, but defining what text segment is the cause and what text segment is the effect is still a challenge. In future work we hope to extend the approach to include additional semantic relation extraction. For example temporal or coherence relations. The methodology that we followed to create the causal relation extraction could be adapted for these additional semantics. The additional semantic relations would allow us to generate results such as time-lines or summaries of events, which would enhance the semantics in a causal networks of news topics.

Ranking causal relations has shown us how we can quantify semantics, this may be applied to the quantification of intrinsic semantics of discourse. An example of intrinsic semantic of discourse is connotation, for example a product may be described as *new* or *experimental* giving it a different connotation where *new* is more reliable than *experimental*, we can be quantify the connotation to determine the objectivity of the text. The field of automatic discourse analysis would be enhanced by the proposed developments.

In our news topic extraction chapter we presented aliases for the news topics. We did not evaluate all the proposed approaches, one of these approaches, the query pattern correlation, may be used to discover latent semantics in text phrases. It would be interesting to use the extended semantics to generate large networks of interconnected news topics. During the development of the system we found potential approaches for the continuous validation and extraction of news topic references. More research is still required to perfect the extraction and validation of news topic references and their aliases.

Due to the limited sources of information we are not able to continuously update *Forest*. This is another open challenge, and also to develop a mechanism to extend the type of information *Forest* can process. During our development we found that there are valuable information sources that are not news articles. Yet, with some adaptation we can still process these information sources. The system may continue to be adapted to process

heterogeneous types of information, not only textual such as tweets or academic reports, but also audio and video formats that are available online. Providing automatically standardized, validated and linked information from heterogeneous sources, may aid in the development of the Semantic Web.

Though the type of information generated by *Forest* is linked text from different source, the presentation of this information lends itself to be graphical, we have not developed the graphical representation of the information *Forest* extracts, it follows that we do not know how users will interpret this representation. We believe that there is an open area of research in the field of information interpretation. Given that we can quantify semantic information, it is valuable to know how users can interpret this information intuitively via a graphical representation. Relating information with a novel technique may require a novel technique in presenting this information. Information presentation technique may also be used to reduce information overload. We hope that developments in this field will provide a new paradigm in navigating information.

We strongly believe that the results from *Forest* can be analyzed for journalistic content, for example to find what authors present unique information, or what publisher present the same information as most other publishers. Developments in this area could afford users an overview of not only the news but of the providers of the news.

Further development in *Forest* with regards to processing news topics may allow the system to process user submitted content. This would make the user a prosumers of the news. The usage pattern of a network of news could provide an insight into what users consider important, not just currently relevant. When the system processes information from heterogeneous sources, analyses its journalistic content from those sources, extracts the relevant information and presents it in a way that is intuitive and easy to understand; And thereafter, it analyses the user submitted information as well as the usage patterns to discover what users consider relevant; Then, we will have a different kind of news portal, one that is not focused on what is new but on what is important, one that would help users gain a perspective on what happens in their world, not the perspective of few central authorities. A news portal so that people can see the Forest for the trees.

Bibliography

- [Allan, 2002] Allan, J., editor (2002). *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA.
- [Allan et al., 1998] Allan, J., Carbonell, J., Doddington, G., and Yamron, J. (1998). Topic detection and tracking pilot study final report. *Evaluation*.
- [Allan et al., 2001] Allan, J., Gupta, R., and Khandelwal, V. (2001). Temporal Summaries of New Topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '01*, pages 10–18.
- [Atkinson et al., 2008] Atkinson, M., Piskorski, J., Pouliquen, B., Steinberger, R., Tanev, H., and Zavarella, V. (2008). Online-monitoring of security-related events. *22nd International Conference on Computational Linguistics: Demonstration Papers*.
- [Balog et al., 2010] Balog, K., Serdyukov, P., and Vries, A. (2010). Overview of the TREC 2010 entity track.
- [Beamer and Girju, 2009] Beamer, B. and Girju, R. (2009). Using a bigram event model to predict causal potential. *Computational Linguistics and Intelligent Text Processing*, pages 430–441.
- [Berry, Michael W and Kogan, 2010] Berry, Michael W and Kogan, J. (2010). *Text Mining: Applications and Theory*. Wiley.
- [Blanco et al., 1991] Blanco, E., Castell, N., and Moldovan, D. (1991). Causal Relation Extraction. *Machine Learning*, pages 310–313.
- [Blei and Ng, 2003] Blei, D. and Ng, A. (2003). Latent dirichlet allocation. *The Journal of Machine Learning*, 3:993–1022.
- [Blei, D.M. and Griffiths, T.L. and Jordan, M.I. and Tenenbaum, 2004] Blei, D.M. and Griffiths, T.L. and Jordan, M.I. and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems 16: proceedings of the 2003 conference*, 16:17.

- [Bollegala, 2007] Bollegala, D. (2007). Measuring Semantic Similarity between Words Using Web Search Engines. *Science And Technology*, pages 757–766.
- [Butnariu et al., 2010] Butnariu, C., Szpakowicz, S., and Veale, T. (2010). SemEval-2010 Task 9 : The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. *Computational Linguistics*, (July):39–44.
- [Butnariu and Veale, 2008] Butnariu, C. and Veale, T. (2008). On the categorization of Cause and Effect in WordNet. *afflatus.ucd.ie*.
- [Carmagnola, 2008] Carmagnola, F. (2008). The five ws in user model interoperability. *Proc. IUI 2008 Workshop on Ubiquitous User Modeling*.
- [Chan, 2011] Chan, Y. S. (2011). Minimally Supervised Event Causality Identification. *Computational Linguistics*, pages 294–303.
- [Chim et al., 2008] Chim, H., Deng, X., and Member, S. (2008). Efficient Phrase-Based Document Similarity for Clustering. *Knowledge Creation Diffusion Utilization*, 20(9):1217–1229.
- [Chinchor and Robinson, 1997] Chinchor, N. and Robinson, P. (1997). MUC-7 Named Entity Task Definition. *Proceedings of the Sixth Message Understanding Conference MUC6*, page 21.
- [Cieri et al., 2000] Cieri, C., Graff, D., and Liberman, M. (2000). Large multilingual broadcast news corpora for cooperative research in topic detection and tracking: The TDT2 and TDT3 corpus efforts. *Proceedings of*, (January 1998).
- [Das and Martins, 2007] Das, D. and Martins, A. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU 4*, pages 1–31.
- [Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- [Dempster and Laird, 1977] Dempster, A. and Laird, N. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*, 39(1):1–38.
- [Evans et al., 2004] Evans, D., Klavans, J., and McKeown, K. (2004). Columbia newsblaster: multilingual news summarization on the Web. *Demonstration Papers at HLT-NAACL 2004*.
- [Feng and Allan, 2009] Feng, A. and Allan, J. (2009). Incident threading for news passages. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, page 1307.
- [Filatova and Hatzivassiloglou, 2004] Filatova, E. and Hatzivassiloglou, V. (2004). Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization*, volume 111.

- [Fleiss, 2003] Fleiss, J. L. (2003). The Measurement of Interrater Agreement. pages 598–626.
- [Frank et al., 1999] Frank, E., Paynter, G., and Witten, I. (1999). Domain-specific keyphrase extraction.
- [Girju, R., Moldovan, 2002] Girju, R., Moldovan, D. (2002). Text mining for causal relations. *Proceedings of the FLAIRS Conference*, pages 360–364.
- [Grishman et al., 2002] Grishman, R., Huttunen, S., and Yangarber, R. (2002). Real-time event extraction for infectious disease outbreaks. *Proceedings of the second international conference on Human Language Technology Research -*, pages 366–369.
- [Harabagiu and Lacatusu, 2005] Harabagiu, S. and Lacatusu, F. (2005). Topic themes for multi-document summarization. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 202.
- [Harabagiu et al., 2002] Harabagiu, S., Lacatusu, V., and Morarescu, P. (2002). Multidocument Summarization with GISTexter. *LREC*, 1(c):1456–1463.
- [Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM*, pages 50–57.
- [Ishii and Ma, 2010] Ishii, H. and Ma, Q. (2010). Causal Network Construction to Support Understanding of News. *System Sciences (HICSS), 2010*, pages 1–10.
- [Ishii et al., 2010] Ishii, H., Ma, Q., and Yoshikawa, M. (2010). Causal Network Construction to Support Understanding of News. *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.
- [Ittoo and Bouma, 2011] Ittoo, A. and Bouma, G. (2011). Extracting explicit and implicit causal relations from sparse, domain-specific texts. *Natural Language Processing and Information Systems*, pages 52–63.
- [Järvelin and Kekäläinen, 2000] Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.*, pages 41–48.
- [Ji and Grishman, 2008] Ji, H. and Grishman, R. (2008). Refining Event Extraction through Cross-Document Inference. *ACL*, (June):254–262.
- [Joshi et al., 2010] Joshi, S., Pangaonkar, M., Seethakkagari, S., and Mazlack, L. (2010). Lexico-syntactic causal pattern text mining. In *Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference*, volume II, pages 446–451.
- [Katz, 2010] Katz, P. (2010). NewsSeecr Clustering und Ranking von Nachrichten zu Named Entities aus Newsfeeds. *Dresden University of Technology*, page 69.

- [Kim and Oh, 2011] Kim, D. and Oh, A. (2011). Topic chains for understanding a news corpus. *Computational Linguistics and Intelligent Text Processing*, pages 163–176.
- [King and Lowe, 2003] King, G. and Lowe, W. (2003). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization*, 57(03):617–642.
- [Knoth, 2011] Knoth, P. (2011). Mining Cross-document Relationships from Text. *Conference on Advances in Information Mining*.
- [Knoth et al., 2008] Knoth, P., Novotny, J., and Zdrahal, Z. (2008). Automatic generation of inter-passage links based on semantic similarity. *Contract*.
- [Lee, 1999] Lee, L. (1999). Measures of distributional similarity. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*.
- [Lewis et al., 2004] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.
- [Li et al., 2006] Li, W., Wu, M., Lu, Q., Xu, W., and Yuan, C. (2006). Extractive summarization using inter-and intra-event relevance. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, (July):369–376.
- [Lin and Liang, 2008] Lin, F.-r. and Liang, C.-H. (2008). Storyline-based summarization for news topic retrospection. *Decision Support Systems*, 45(3):473–490.
- [Liu et al., 2007] Liu, M., Li, W., Wu, M., and Lu, Q. (2007). Extractive summarization based on event term clustering. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, (June):185–188.
- [Liu et al., 2009] Liu, S., Merhav, Y., Yee, W. G., Goharian, N., and Frieder, O. (2009). A sentence level probabilistic model for evolutionary theme pattern mining from news corpora. *Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09*, page 1742.
- [Luhn, 1958] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts.
- [Manning and Raghavan, 2009] Manning, C. D. and Raghavan, P. (2009). An Introduction to Information Retrieval.
- [McCracken et al., 2006] McCracken, N., Ozgencil, N., and Symonenko, S. (2006). Combining techniques for event extraction in summary reports. *AAAI 2006 Workshop Event Extraction and Synthesis*.

- [Medelyan et al., 2009] Medelyan, O., Frank, E., and Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. (August):1318–1327.
- [Medelyan and Witten, 2008] Medelyan, O. and Witten, I. (2008). Topic indexing with Wikipedia. *Proceedings of the AAAI WikiAI workshop*, pages 19–24.
- [Mei and Zhai, 2005] Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207. ACM.
- [Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, (1991):1–20.
- [Nallapati et al., 2004] Nallapati, R., Feng, A., and Peng, F. (2004). Event threading within news topics. *Proceedings of the thirteenth*, pages 446–453.
- [Naughton et al., 2008] Naughton, M., Stokes, N., and Carthy, J. (2008). Investigating statistical techniques for sentence-level event classification. *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, (August):617–624.
- [Oliveira et al., 2002] Oliveira, P. D., Ahmad, K., and Gillam, L. (2002). A financial news summarization system based on lexical cohesion. *Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France*.
- [Perrin et al., 2008] Perrin, T., Kawai, H., Kunieda, K., and Yamada, K. (2008). Global Dynamics Network Construction from the Web. *2008 International Workshop on Information-Explosion and Next Generation Search*, pages 69–76.
- [Piskorski et al., 2007] Piskorski, J., Tanev, H., and Wennerberg, P. O. (2007). Extracting violent events from on-line news for ontology population. In *Business Information Systems*, pages 287–300. Springer.
- [Porteous et al., 2008] Porteous, I., Newman, D., Ihler, A., and Asuncion, A. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. *Proceeding of the 14th*, page 569.
- [Radev et al., 2004] Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.
- [Radinsky, 2011] Radinsky, K. (2011). Learning causality from textual data. *Proceedings of Learning by Reading for Intelligent Question Answering Conference*.
- [Riaz and Girju, 2010] Riaz, M. and Girju, R. (2010). Another Look at Causality: Discovering Scenario-Specific Contingency Relationships

- with No Supervision. *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368.
- [Rink et al., 2010] Rink, B., Bejan, C. A., and Harabagiu, S. (2010). Learning Textual Graph Patterns to Detect Causal Event Relations. *Artificial Intelligence, (Flairs)*:265–270.
- [Rössler, 2007] Rössler, M. (2007). *Korpus-adaptive Eigennamenerkennung*. doctoral, Universität Duisburg-Essen.
- [Saito et al., 2007] Saito, H., Kawai, H., and Tsuchida, M. (2007). Extraction of Statistical Terms and Co-occurrence Networks from Newspapers. *Proceedings of NTCIR-6 Workshop Meeting*.
- [Salton et al., 1975] Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11).
- [Sang and Meulder, 2003] Sang, E. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.
- [Schapire, 2000] Schapire, R. (2000). BoosTexter: A boosting-based system for text categorization. *Machine learning*, pages 135–168.
- [Shahaf and Guestrin, 2010] Shahaf, D. and Guestrin, C. (2010). Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM.
- [Surdeanu and Harabagiu, 2003] Surdeanu, M. and Harabagiu, S. (2003). Using predicate-argument structures for information extraction. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics*, pages 8–15.
- [Tanev, Hristo and Piskorski, Jakub and Atkinson, 2008] Tanev, Hristo and Piskorski, Jakub and Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems*, pages 207—218. Springer.
- [Wan, 2007] Wan, X. (2007). Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. *Computational Linguistics*, (June):552–559.
- [Wang and Li, 2011] Wang, L. and Li, F. (2011). Story Link Detection Based on Event Words. *Event (London)*, pages 202–211.
- [Wang et al., 2010] Wang, W., Zhao, D., Zou, L., Wang, D., and Zheng, W. (2010). Extracting 5W1H event semantic elements from Chinese online news. *Web-Age Information Management. Springer Berlin Heidelberg, 2010*, (20080440260):644–655.

- [Wolff et al., 2002] Wolff, P., Song, G., and Driscoll, D. (2002). Models of causation and causal verbs. In *the Meeting of the Chicago Linguistics Society, Main Session*, volume 1, pages 607–622. Citeseer.
- [Wunderwald, 2011] Wunderwald, M. (2011). *NewsX: Event Extraction from News Articles*. Diplom, Dresden University of Technology.
- [Yaman et al., 2009] Yaman, S., Hakkani-Tür, D., and Tür, G. (2009). Combining semantic and syntactic information sources for 5-w question answering. *INTERSPEECH*, pages 2707–2710.
- [Yang et al., 2009] Yang, C., Shi, X., and Wei, C. (2009). Discovering event evolution graphs from news corpora. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 39(4):850–863.
- [Yangarber, 2006] Yangarber, R. (2006). Verification of facts across document boundaries. *Proc. International Workshop on Intelligent Information*.
- [Zhang et al., 2008] Zhang, K., Li, J., Wu, G., and Wang, K. (2008). Term Committee Based Event Identification within News Topics. *Analysis*, (1):821–829.
- [Zhao and Wang, 2010] Zhao, H. and Wang, F. (2010). Research into the Topics Representation in Topic Tracking. *English*, (Fskd):2498–2502.