

FACULTY OF ARTS
DEPARTMENT OF LANGUAGES
UNIVERSITY OF HELSINKI

**Information extraction and
linguistic characteristics of texts:
exploring scenarios and text types**

Silja Huttunen

*Doctoral dissertation, to be presented for public discussion with
the permission of the Faculty of Arts of the University of Helsinki
at Porthania, room P673, on the 10th of October, 2020
at 12 o'clock*

HELSINKI 2020

Supervisor

Professor Lauri Carlson, University of Helsinki, Finland

Pre-examiners

Professor Robert Gaizauskas, University of Sheffield, UK

Dr Jussi Karlgren, KTH Royal Institute of Technology, Sweden

Opponent

Dr Jussi Karlgren, KTH Royal Institute of Technology, Sweden

Custos

Professor Lauri Carlson, University of Helsinki, Finland

Contact information

Department of Languages
Unioninkatu 40
FI-00014 University of Helsinki
Finland

Email address: info@helsinki.fi

URL: <http://www.helsinki.fi/>

Copyright © 2020 Silja Huttunen
ISBN 978-951-51-6640-1 (paperback)
ISBN 978-951-51-6641-8 (PDF)
Helsinki 2020
Unigrafia

Information extraction and linguistic characteristics of texts: exploring scenarios and text types

Silja Huttunen

Abstract

Information Extraction (IE) is the systematic harvesting of information from natural language text and speech into structured form, e.g., into a database, for further downstream use. The most typical use cases are related to media monitoring. Research in IE is driven by the need to find accurate information about a particular topic in massive collections or streams of text.

In addition to the traditional methods of evaluation in IE, we introduce a second measure of quality, which indicates the relevance, or usability, of the extracted facts for an end-user. An extracted fact may be correct, but irrelevant from the user's perspective.

This dissertation presents work on two problems: 1. porting an IE system from one topic to another, and 2. assessing the user-oriented relevance of results produced by an IE system.

All tasks are not equally responsive to IE, and performance on some tasks remains worse than on others, despite extensive customization. The first part of this study is motivated by the gap between performance obtained by IE systems for different topics. Our experience with customizing IE confirms the intuition that different domains exhibit different kinds of complexity, e.g., the business-related domain vs. the domain relating to natural events.

The underlying reason is the variation in the language that is used to report the topics. The aim of this thesis is to improve IE results by determining which linguistic and structural features should be taken into consideration when customizing an IE system to a new topic.

In the process of adapting the IE system to several domains and building their knowledge bases, we analysed the linguistic and structural characteristics of the

domains, and the style of reporting. Information extraction is used as a methodological tool for linguistic observation, as it enables us to expose and explore how linguistic variation affects the IE results.

The second part focusses on measuring relevance of the IE results, that is, how well the extracted information satisfies the user's interest. We identify which linguistic and structural features are useful for improving the performance on these scenarios.

It has been observed elsewhere in NLP settings, that taking the features into account can produce better results. Thus, the findings presented in this work can be beneficial for a variety of approaches to IE, including those based on machine learning techniques.

Acknowledgements

I am grateful for the help and support of numerous people. First, I want to thank my advisor, Professor Lauri Carlson at the University of Helsinki. I would especially like to thank my pre-examiners, Professor Robert Gaizauskas and Dr. Jussi Karlgren, for invaluable feedback and insightful comments and suggestions regarding my work. They helped to illuminate the structure and the main points of my work.

I am grateful to Fred Karlsson, who gave me Biber’s “Variation across Speech and Writing,” which sparked my interest in types of speech and writing. Further inspiration came while I was working on my Master’s thesis with Orvokki Heinämäki, to whom I am also very grateful. I thank my colleagues and friends at the Linguistics department: Juha Heikkilä, Tarja Heinonen, Timo Järvinen, Pasi Tapanainen, Jussi Piitulainen, Matti Miestamo, Antti Arppe, Kimmo Koskeniemi, Seppo Nyrkkö, and so many others.

As I got a life-changing opportunity from Ralph Grishman to pursue the interest in spoken and written “genre,” I did not hesitate to join the Proteus group at New York University to help with the customization of an IE system to news reports about natural disasters. My interest expanded into observing linguistic differences in news reporting in other news topics, and the results are now here in my thesis. I would like to thank my colleagues in the Proteus group at the Department of Computer Science: Adam Meyers, Catherine Macleod, Michael Gregory, Winston Lin, Kiyotaka Utchimoto and so many others for inspiration and company. I am extremely grateful to Professor Ralph Grishman at New York University for the opportunity to work with him, which was a memorable and pleasant experience, and a turning point in my career.

Back in Finland, I joined the PULS project to continue the work on my thesis. I am grateful to my colleagues Peter von Etter, Mian Du, Lidia Pivovarova, Arto Hellas, and many others, as well as to our project leader Roman Yangarber, for inspiring and fun company.

I thank my friends in Linguistics, the inspiring Aspekti crowd of the early days, among others Soili Turro, Marja Pälsi, Mickael Suominen, Jarno Raukko, Risto Widenius, and especially Mila Engelberg for her encouragement and support, and Johanna Ratia, the master of information services.

This work was supported in part by Tekes, the Finnish Funding Agency for Technology and Innovation, through Project: *ContentFactory*; the Stockholm Environment Institute (SEI), Sweden, through Project: *Automatic Data Gathering for the Arctic Resilience Report*; and Frontex: European Border and Coast Guard Agency, through Project: *Capturing Structured Information on Illegal Immigration Events, Cross-border Criminal Activities and Related Crisis Events from Online News*. I am grateful for their support.

My deepest gratitude goes to my family, my sisters Soili and Heli for their emotional support, and especially to my dear and lovely daughter Eli, whose humour and great ideas are at the same time both inspiring and grounding.

Contents

1	Introduction	1
1.1	Background and preliminaries	2
1.1.1	Information extraction	2
1.1.2	Evaluation	4
1.1.3	Judging the relevance of the IE results	6
1.2	Objectives and scope	7
1.3	Linguistic background	8
1.3.1	Genre and text types	8
1.3.2	Cohesion, coherence and discourse structure	12
1.4	Related work	16
1.5	List of original publications	19
1.5.1	Publication I	20
1.5.2	Publication II	20
1.5.3	Publication III	21
1.5.4	Publication IV	21
1.5.5	Publication V	21
1.5.6	Publication VI	22
1.5.7	Publication VII	22
1.6	Structure of the research	23
2	Analysis of cross-domain issues in IE	25
2.1	Scenario characteristics and customization	26
2.1.1	Nature domain	27
2.1.2	Business domain	29
2.1.3	Security and Cross-Border Crime	31
2.2	Data	34
2.3	Setup and tools	36

2.4	Results of the linguistic analysis	38
2.4.1	Lexical analysis	40
2.4.2	Structural analysis	47
2.4.3	General and scenario-specific pattern types	59
3	Assessing relevance of IE results for the user	69
3.1	User-centric relevance of IE results	69
3.2	Data	73
3.3	Setup and tools for relevance	73
3.4	Results: domain features	75
3.4.1	Discourse features	76
3.4.2	Combining lexical and discourse features	86
3.4.3	How scenarios reflect discourse features	87
4	Summary and conclusion	93
4.1	Summary of results and discussion	93
4.2	Conclusions and implications for future work	96
	References	99
	Appendices	
	Appendix A Examples of extraction	111
A.1	Intra-sentential	111
A.2	Intra-paragraph inter-sentential	111
A.3	Inter-paragraph	112
	Appendix B Examples of launch	113
	Appendix C Real-world application of IE	115

Chapter 1

Introduction

Automated information extraction (IE) arose from the need to quickly find accurate information about a particular topic in large masses of natural language texts, for example, text in news sites around the internet. Most information extraction systems are manually or semi-manually customized for each new topic. This customization work is slow and expensive. Further, all IE tasks are not equally responsive, and the performance remains poor despite extensive customization. This study is motivated by the gap between results obtained in evaluation of IE systems for different topics. The results of the customization projects described in this study confirm the intuition that different kinds of complexity emerge when IE is applied to different domains, for example, to the Business domain vs. domains relating to natural events, such as natural disasters. In this study, I apply linguistic analysis to determine what needs to be taken into consideration for successfully customizing an IE system to a new topic. As the differences in IE results made linguistic differences between domains more clear during the customization processes, on one hand, information extraction is used as a methodological tool for linguistic observation. On the other hand, identifying good linguistic features helps improve the performance of IE systems. The findings can be useful for a variety of IE approaches, including those based on machine learning techniques, such as neural networks.

The experimental and implementational work presented in this manuscript has been performed several years earlier, and current (2020) research on text and speech analysis is almost entirely performed using a different approach than the one presented in this manuscript. Recent neurally inspired statistical approaches which show impressively better results on many language processing tasks have not yet been broadly used for information extraction and while the work presented in

this manuscript has different starting points, a different knowledge representation, and a different processing model, the analysis base and the evaluation can provide valuable inspiration for coming efforts.

1.1 Background and preliminaries

1.1.1 Information extraction

Information extraction (IE) is a technology used for locating and extracting pre-specified information from unstructured natural language text. The extracted pieces of information are also called *events* or *facts*. In information extraction the facts are extracted from a large text corpus, often from dynamic streams of news articles. Extracted events provide responses to questions such as who did what, where, when, and why. For a good overview of IE, with its history and development, see, e.g., Grishman (2019), Piskorski and Yangarber (2013), and Sarawagi (2008). Piskorski and Yangarber (2013) cover the advances up to the time when neural network methods broke onto the NLP scene and revolutionized many NLP tasks (in 2014).

The extracted information consists of structured objects, called *entities*, belonging to particular semantic classes, *relationships* between these entities, and *events* in which these entities participate. The extraction system places this information into a database for subsequent retrieval and processing.

Domains and portability

A *domain* is the broad topic or subject matter of the text from which we extract information, e.g., financial news, news about crimes, or medical reports. A *scenario* is the narrow type of events inside a domain that the IE user is interested in.¹ From 1987, IE was strongly influenced by two competitions, the Message Understanding Conference (MUC) (Hirschman, 1998) and Automatic Content Extraction (ACE) program, with support from DARPA, the US defense agency. These competitions covered scenarios such as naval operations messages, terrorism in Latin America, joint ventures in microelectronics, management succession, and satellite launches.

The scenarios used for this study are listed below in Table 1.1, grouped into three domains. These scenarios are presented in Sections 2.1.2, 2.1.1 and 2.1.3, with examples of relevant passages from news articles, and the sought information listed as events in tabular form.

¹In this work, *scenario* is sometimes also called *topic* when it is more appropriate for clarity.

Table 1.1: The main domains and scenarios included in this study

<i>Nature Domain</i>	<i>Business Domain</i>	<i>Cross-Border Security Domain</i>
<ul style="list-style-type: none"> • Natural Disasters • Infectious Disease Outbreaks 	<ul style="list-style-type: none"> • Management Succession • Corporate Acquisitions • Product Launches • Investments 	<ul style="list-style-type: none"> • Smuggling • Human Trafficking • Illegal Migration

Mergers between companies, for example, are events of interest in the Corporate Acquisitions scenario of the Business domain. Arrests related to human trafficking are to be extracted in the world in the Human Trafficking scenario in the Security domain, and infectious disease epidemics are to be extracted in the Infectious Disease Outbreaks scenario.

A typical pattern-based IE system has customizable *knowledge bases*. The knowledge bases contain a large set of linguistic *patterns* (see Section 2.4.3). The patterns are often regular expressions, matching a syntactically and semantically typical construction in text that states a sought fact. A sentence (or clause) where a pattern has matched is a *trigger sentence* (or clause). The patterns utilize general and domain-specific *lexicons*, and *ontologies* relevant to a given topic, where a set of *concepts* are arranged by their properties and relations to each other. One example of relation is hypernymy,² but the ontologies typically contain various other relations as well (see Section 1.3.2). The extracted facts are organized in scenario-specific output *templates* (see Section 2.4.2). Each *predicate* defines which slots belong to the template and the type of objects required for the slots.³ The knowledge bases have to be customized separately for each new domain as defined by the guidelines for the task. The guidelines specify user needs and what kind of fragments of information may constitute an event for a specific domain or scenario (Appelt et al., 1995; Yangarber and Grishman, 1998b).

All of these knowledge bases need to be customized for each scenario. Resolving the issues relating to portability is a key research question in this thesis.

²A hypernym is a more generic term that includes the meaning of a more specific term. For example, for Infectious Disease Outbreaks scenario, the term “disease” includes “yellow fever”; for the Natural Disaster scenario, “storm” includes “heavy rains”.

³For example, a “disease outbreak” predicate may call for slots like *disease name*, *number of victims*, *location*, etc.

Methods and systems

IE systems are either rule-based, or based on statistical methods, or a combination of both, depending on the nature of the extraction task. For this study, I used two IE systems which are based on pattern matching, Proteus (Yangarber and Grishman, 1997; Grishman, 1997) and PULS⁴ (Yangarber et al., 2005), see Section 2.3. For each scenario, I used annotated training data (see Table 2.11 in Section 2) for pattern building and linguistic analysis, and separate data for testing. My aim was to improve IE results by applying linguistic analysis. In the first part of this study, I analysed the language of the scenarios as part of adapting the IE system to new domains. To build the knowledge bases, I observed the linguistic and structural characteristics and the style of reporting in different scenarios. For each scenario, I attempt to analyze and quantify a wide range of features of the text. These include part-of-speech constructions relating to events or event attributes: verb constructions, temporal and locative constructions, and prepositions related to them; document length, the distance of attributes from the trigger sentence, the lexical devices that contributed to linking event to each other, types of inclusion relationships between events and incidents (in those scenarios that have inclusion relationships). I focus on issues that intuitively, based on my extensive hands-on experience with the data, seem to cause challenges for the IE system to process. The aim is to find systematic differences.

A selection of these features, which we were able to implement, were then used for the *relevance* analysis. which is the focus in the second part of the study. The notion of relevance is a measure of how well the extracted information answers the user’s interest. In order to build classifiers for relevance, I identified useful features based on the linguistic analysis and the customization experiences in the first part of the study.

1.1.2 Evaluation

The extraction results of an IE system are traditionally evaluated in terms of precision, the proportion of correctly extracted events of all extracted events, and recall, the proportion of correctly extracted events of all events. The notions of recall, precision and F-measure are inherited for Information retrieval. For IE the evaluation is harder to compute. The procedure to compute precision and recall is the following.

⁴Proteus is the name of the NLP research group developing information extraction systems at New York University’s Department of Computer Science. PULS (Pattern-based Understanding and Learning System) is a NLP project at the University of Helsinki.

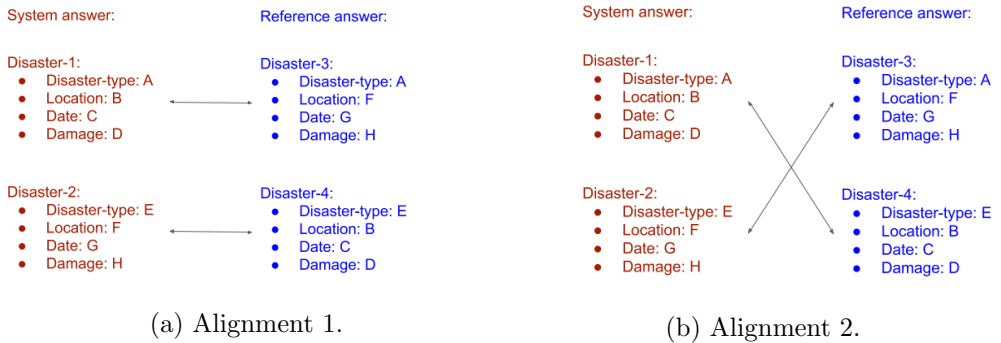


Figure 1.1: Possible alignments of two templates for a document about natural disasters.

Each event is represented as a template with multiple slots. The IE system returns zero or more templates for each document in the evaluation corpus. For each document in the evaluation corpus, human annotators have filled manually zero or more correct templates with slots. Those are the reference answers. The evaluation system is comparing the templates found by the extraction system to those made by human annotators, see a simplified example in Figure 1.1. The evaluation system compares the slots of the extracted templates to match the slots in the reference templates.

The evaluation is a two-step process. In the first step, it finds an optimal correspondence (mapping) between IE system output objects (templates) and reference (true) objects. This mapping is chosen so that the performance measure (F-measure) used for system evaluation is maximized. The slots may be weighted differently, which is taken in consideration in computing the score.

Once we have found the optimal mapping, the scoring program counts how many slots were filled correctly, given that mapping. For example, Figures 1.1a and 1.1b show two alternative mappings for two output templates produced by the system (on the left side) and two reference templates (on the right side). In the mapping in Figure 1.1a, only 25% of the slots are correctly filled. In the mapping in Figure 1.1b, 75% of the slots are filled correctly. Therefore the second mapping should be chosen by the evaluation procedure.

For the event template to be correct, all the slots have to be correctly filled. This type of evaluation was used, e.g., in the series of MUC initiatives for evaluating of IE results produced by the participants. For all scenarios, the same evaluation program is used, which was provided by the MUC organizers. Similar

evaluation for ACE (Automatic Content Extraction) is described in Doddington et al. (2004). In ACE, the objective was in extracting entities and the links between them. In this work as in the MUC tasks the objective is to find entire events. Entity recognition and coreference are built in the IE system, and not within the scope of this study.

1.1.3 Focus on user needs: relevance of IE results

The *relevance* of the IE results is the second principal theme of my work. It is essentially related to the practical results of the IE system and the IE user's needs. A key practical application of IE systems is in news surveillance. The focus now shifts from tuning the system and the knowledge bases to find all possible events, rather toward assessing the relevance of the IE results for the end-user, as described in Section 3.1.

News surveillance is an essential part of the work for analysts in various knowledge- and information-intensive fields. Tools for mining open sources for gathering strategic information are indispensable. However, not all extracted facts are equally relevant to users and certain types of information may be more crucial in one scenario than in another. To understand what the relevant information is, PULS utilizes user feedback to the surveillance system. The end-users of the PULS IE system at the time of this study are analysts in the fields of business, disease epidemics, and cross-border security surveillance. The relevance of (an event in) a document is scored by the end-users through the user interface. The relevant pieces of information in one article are what is relevant to the users, who has defined what kind of attributes of a given domain they want to know. An analyst is usually interested only in the relevant information of a narrow area in one scenario, and only in specific types of facts. In disease epidemics, for example, information such as the name of the disease and the number of casualties are essential. A healthcare professional may be less interested in vast outbreaks of an infectious disease that took place earlier or over the years, than in a very recent single case of an uncommon disease. So the latter case has *higher value* for that user.

Once the attributes of an event have been extracted, a relevance classifier is invoked to assign a relevance score to each extracted event, see Section 3.1. The event relevance scores also appear on the on-line server. In addition to editing the event fills, the users can also assign or edit the relevance labels on the extracted events. The set of events for which the relevance was manually labeled by human users are used for training and testing the relevance classifiers. For the second

part of the study, our IE system, customized for the given scenario, then marks events of high relevance from a stream of articles (Yangarber et al., 2007). I take a set of features identified from the analysis of the first part of the study, and we use statistical methods to quantify the differences.

1.2 Objectives and scope

Not all scenarios and texts are equally responsive to pattern-based IE. The problem is how to improve the results produced by “traditional” IE and understand the reasons for the better performance of one scenario as opposed to that of another. In this work my focus is on the linguistic aspects of this problem.

I examine the linguistic variation in texts in the same genre (see Section 1.3.1). Information extraction is a tool for this observation, as it enables us to measure how linguistic variation affects the IE results and their usability. The analysis of the scenario-specific and text-specific characteristics is done by applying concepts from linguistics, and in particular from text analysis (see Section 1.3).

The research questions evolved with time, so the latter questions are based on the earlier ones:

1. What accounts for the variation in the performance of an IE system on different scenarios?
2. How do the scenarios differ on the lexical level and on the structural level, and why?
3. For a given scenario, how could these differences be traced, measured and ultimately leveraged
 - a. to improve the IE system in general?
 - b. to improve the IE system in relation to user needs?

Questions 1, 2, and 3a are addressed in Chapter 2, and the question 3b in the next Chapter 3. I focus on those specific parts of the texts that are predefined as essential information of a given scenario. By analyzing the scenario- and text-specific characteristics, I obtain new insights to the applicability of IE systems to new topics (or *scenarios* in IE), and into better usability of the IE end-results. The usability is further improved by extracting information that has a high relevance to the user.

1.3 Linguistic background

This section presents the linguistic background of the concepts and ideas applied for this work. They are derived from the field of general linguistics and text linguistics to serve the practical function of improving the IE customization process and the usability of the IE results.

1.3.1 Genre and text types

Harris (1988) proposed using linguistic procedures to extract certain relations from scientific articles within a given domain, and to use the extracted relations for document access. He worked with *sublanguage* of science and restricted word co-occurrence. According to Zellig Harris, “the language of technical domains,” such as those found in genomics and medicine, has a structure and a regularity that can be observed by examining the corpora of those domains. Based on Harris’s linguistic theory, the Linguistic String Project (LSP) at New York University was one of the earlier research projects on computational processing of natural language. LSP was adapted, e.g., for analysis of medical patient records (Sager et al., 2002).

In our IE systems, extraction patterns are also built based on observing texts in the given domain. However, our IE system was applied to a wider range of scenarios, and our situation is more challenging. Whereas the domain of patient records has highly restricted, formulaic language, in some of our domains the language is less technical, is rather closer to general language of everyday communication, and contains more ambiguity.

As we show in Chapter 2, in the process of customizing our IE system, we need to take into account the domain, the register, and the text type. In addition to the notion of *domain*, introduced in Section 1.1.1, this Section clarifies the usage of these terms as used in this thesis: register, text type, and genre.

Register refers to language modality, written or spoken. We considered IE from both registers, but in this thesis we concentrate on written data.

Another key dimension defining the linguistic characteristics of texts, which affects the extraction process is the *text type*. For example, we find that dealing with news articles, as opposed to short news abstracts, strongly affects the IE process and the results.

The notion of *genre* and text type from linguistics is relevant for our work in information extraction, as these terms are often used interchangeably in the linguistic literature. In some cases they refer to different textual phenomena.

Genre is often defined by culturally conventional grouping of texts, or regarded as a culture-specific type of activity to which the text belongs because of its schematic structure, e.g., Gregory (1988), Eggins (1994) and Martin (1992). Examples are literary genres, such as short stories, romantic novels and sitcoms, or common written genres, such as instructional manuals, news articles and recipes.

In this work, I use the term genre to refer to a collection of texts that is defined by the purpose it was created for, a culture-specific type of activity. One collection of texts does not necessarily share the same linguistic characteristics. According to Halliday (1985), in the popular written genre of newspaper articles, as defined by its function, a news article is expected to properly report an actual event to the readers. For our work the concept of genre is less useful, as it covers all our corpora as the genre of news reports, regardless of the topic. Our IE system, originally tailored for business news, was not performing well on the Nature scenarios, even after lexical customization. Treating the corpus of news reports as linguistically uniform in the customization process, kept the performance of the IE system low. Rather, there are several types of texts in the corpus of news reports. Here I use the term *text type* following Biber (1988), who uses it in his statistical corpus-linguistic work to refer to groupings of texts⁵ that are similar with respect to their *linguistic form*, irrespective of genre categories. Biber (1988) analyzes spoken and written corpora⁶ using linguistic features combined with statistical methods. He uses features for which specific discourse functions have been claimed in the relevant linguistic literature. They include, for instance, tense and aspect markers, pronouns, discourse particles, place and time adverbials and word length.

The text type is defined by internal linguistic criteria, e.g., Lee (2001). The content of the text has an influence on its form (Saukkonen, 2001). Text type reflects the purpose that the writer had in mind when she wrote the text, and chose the linguistic characteristics and devices (semantic, structural, lexical, stylistic) realized in the given text. Style is an individual choice.

Given the vast amount of texts being processed by our IE systems, the variation in style was not initially considered to be important and was not specifically addressed. Further, generally, in IE the notion of text type has not been considered important on its own. However, with the rising interest in extracting sentence-level information from *any* kind of texts on the Web, the need to categorize texts into

⁵Biber suggests the following textual dimensions: involved production, informational production, narrative concern, explicit reference, situation-dependent reference, overt expression of persuasion, abstract information, and online informational elaboration.

⁶Sections of the London-Lund corpus of spoken English and the Lancaster, Oslo and Bergen between 1970 and 1978 (LOB) corpus of written English, compiled by Svartvik (1990).

text types has emerged (Santini, 2006). Web pages do not follow the traditional division into genre (Sahran and Imran, 2009; Kilgarriff and Grefenstette, 2003), and the text typologies already available might not cover them (Santini, 2006). On the Internet, the same scenario may be reported in individual styles in blogs, newswire, the Wikipedia, discussion groups, official reports, and so on. Attempts have been made to categorize texts. For example, Lijffijt and Nevalainen (2017) claim to be able to recognize the core genres⁷ in the BNC corpus using pairs of surface features, such as counts of nouns and pronouns, or average word lengths and type/token ratios. Biber and Egbert (2016) have applied multi-dimensional analysis based on linguistic features to a representative sample of the entire searchable Web, to recognize genre—or register—categories.

Link between scenario and text type

Observations on our corpus suggested that there is a tendency for certain scenarios to invoke a specific style of expression of the facts. News reporters change their style of reporting according to the scenario. Even if most of the analysed texts are news reports by content and mainly from the same or similar sources, the reporting style varies along a continuum, which ranges from scenarios like Nominations in the Business domain, with highly formulaic expressions at one end, toward scenarios like Natural Disasters with rich descriptive narrative at the opposite end.⁸ The variation in style in the texts reflects in the results obtained by our IE systems.

In addition being affected by the scenario, the style of reporting is influenced by the *purpose* of the text as well. The form varies according to the interpersonal functions of the text, that is, according to the role relationships and who is talking to whom and why (Halliday, 1985).

The linguistic literature analyzes the characteristics of different texts on several levels. According to Halliday (1985), language is structured to make three kinds of

⁷The British National Corpus genres: face-to-face conversation, prose fiction, broadsheet newspapers, and academic prose.

⁸In our applications, some of the articles are news summaries, which form a class of their own in the style of reporting. Summary articles are used as additional evidence, but are not the focus of this study, and are not considered in detail in the comparisons between scenarios. Such reports can include several scenarios within one news report, described briefly. Such summary reports are linguistically very challenging for reference resolution and inference rules (see Chapter 2), since the topic switches abruptly from one to another. The individual arguments of the events may be extracted correctly, but it is very challenging to group them correctly into events. The cohesion between the items and the coherence of the whole article is hard to track correctly (see Section 1.3.2). Summary articles with fragmentary coherence are problematic for IE, and require a different approach.

meanings at the same time, on three semantic levels: *experiential*, *interpersonal*, and *textual*. Experiential meanings (or *metafunction*; also *ideational metafunction*) are about the world and experience, about language as an informing medium: it answers questions like “who does what to whom.” It is about how we represent experience in language, what facts or opinions are stated in the text, what the text is about and what is said. Interpersonal meanings are about role relationships and attitudes toward the subject matter or the receiver of the message, that is, “who is talking to whom and why.” Textual meanings⁹ refer to the order of constituents, how the constituents are organized within and among clauses to achieve different purposes, and how the parts relate to each other, to the text, and to the purpose of the text.

On the experiential level, a piece of news about a hurricane and a piece about a company acquisition are events in the world that include participants, processes and circumstances. They provide an answer to the question what has happened and where. On the interpersonal level, in case of a piece of news about a hurricane, the attitude of the writer toward the subject matter, and the attitude of the reader, is more intense. Reporters need to catch their readers’ attention and meet their needs.¹⁰ On the textual level the structure of disaster news differs from that of business news, as shown in later chapters. The lexical choices differ as well. These levels materialize in our analysis in Section 2.4.

Bakhtin and Ghāsemipour (2011) suggest that there exists a very large number of genres. Various everyday genres are, e.g., greetings, farewells, congratulations, information about health, business, etc. These genres are so diverse because they depend on the situation, social position, and interpersonal relations of the participants in the communication. These genres have high, strictly official, respectful forms, as well as familiar ones. This is a wider sense for genre. Each area of speech communication has its own typical addressee and the assumed typical conception of the addressee defines the genre. Based on that and the expected reaction of the addressee to the message, the speaker chooses the style with which to communicate.

Understanding possible interactions between scenarios and text types is useful for IE, as the needs in IE have evolved in recent years. The initial IE systems were customized to only one scenario at a time. Currently, the need is to cover many scenarios simultaneously, and the goal is eliminating or reducing the time-consuming and expensive manual customization. It is very challenging to create

⁹Jeffries (2015) calls the ideational metafunction *linguistic meaning*.

¹⁰This is another aspect that Harris did not need to handle in his more limited scenario.

computational models of the different levels of language. The pragmatic or world knowledge is especially hard to encode in such a way that it would be usable in AI.

1.3.2 Cohesion, coherence and discourse structure

Studies in cohesion and discourse structure form the background for the analysis of linguistic and structural differences of our texts. To establish the relevance¹¹ ratings for events I explored linguistic devices that create coherence and help in developing a set of features that indicate the relevance of the events.

See Section 3.1 how we use the features to predict relevance, and Section 3.4 for a description of the individual features.

The stylistic variation in scenarios is manifested not only on the lexical level, but also on the structural level.

Discourse structure is a term used to describe the way in which an entire text is organised, e.g., how language is used in instructional manuals, poems, news articles, or recipes, see e.g. Halliday (1985) and Gregory (1988). Discourse analysts have been investigating the organization of information in different genres or text types, e.g., Biber et al. (2007); Marcu (2000); Mann et al. (1989). From the perspective of this study, an understanding how the text is structured (e.g., as in systemic functional approach) gives useful insight for finding the clues that signal where the relevant information is situated. Analyzing the discourse structure of the entire text is too challenging. For the purposes of IE, I limit my focus to observing how a set of pre-defined facts are expressed in the scenarios.

Cohesion is generally defined as the linguistic means that create the structure in the text by means of explicit linguistic devices. They signal relations between sentences, clauses, paragraphs and other parts of texts. Cohesive devices are lexical, such as phrases or words, and/or grammatical, providing the reader with hints to link prior statements with subsequent ones, e.g., van Dijk (1977) and Connor (1996).

Some of the domains are structured in such a way that the event and its attributes are far apart from each other. The cohesive devices work as glue between the event and its attributes and give the reader a hint how these attributes may be connected. We call this spreading of the facts in the text *scattering*. Bagga

¹¹ *Relevance* here does not refer to the term relevance as it used in the linguistic literature, e.g., in Sperber and Wilson (1986). It is related to it, though, in the sense that the information with high relevance value meets the user's expectations as an illuminating answer to a question that has been evoked in the user's mind.

and Biermann (1997) introduce a similar concept and call it a “level” of a fact. The level of a fact is defined as the number of arcs that connect the nodes of one event in a semantic network build for that text. Each node denotes an object, which is an item that fills a slot in an event template. Each arc represents a binary relation between the objects. The more arcs there are in an event, the more complex the event is. The scattering of a domain is the average scattering of the event in that domain. We take a simpler approach, and take the difference (here, in characters) between the first and last attribute of an event found in the text, see Table 1.2. This distance is considerably longer in the Nature domain and the Security domain, compared to the Business scenarios.

Table 1.2: Average scattering in different scenarios (in characters).

<i>Nature domain</i>			<i>Business domain</i>		
Disasters	Diseases	Security	Investments	Product Launches	Contracts
1725.5	1195.1	795.3	279.1	153.1	197.5

Grammatical devices signalling cohesion include:

G1. conjunctions that mark transitions, e.g., *but, so*:

“*The disease is generally not fatal, **but** some patients have been hospitalized due to dehydration*”

“*Deutsche Telekom already has a dominant on-line business in Germany, called T-Online, with almost 900,000 subscribers, **but** it is a business-oriented service.*”

G2. reference relations: the use of deictic, anaphoric and cataphoric elements or a logical tense structure, e.g., *she/he, this, they*

“**Health officials** said Tuesday that about 120 people in the New York City area had been stricken by an intestinal ailment caused by an exotic microbe, and **they** said the illness had been reported in at least 10 states and in Ontario, Canada”¹²

¹²Reference resolution is computationally challenging as semantic or pragmatic knowledge is often needed to understand the target of a referring expression, e.g., a pronoun. In this example, *they* could refer to the victims as well.

G3. ellipsis, elliptical constructions, e.g., in “*and [0] infected 300 [1]*”, where the ellipsis [0] marks a “missing” (elided) item of category DISEASE NAME, and [1] marks a missing item VICTIM, e.g., *people*

G4. substitution (e.g., *one, some, no*)

Lexical devices include:

L1. repetition, simple or complex, e.g.,

*What is significant about this **storm** is that it really is several **storms** one on top of the other,...*

L2. antonymy (*public—private*)

L3. hypernymy (superordinate) and hyponymy (subordinate) (*disease—Swine flu*)

L4. meronymy (*roof—building*)

L5. synonymy, e.g.,

*Authorities in the region said 50,000 homes **were without power**, and more were going dark as conditions deteriorated. In Kingston, upward of 80 percent of residents **lost electricity**.*

L6. collocations (Sinclair, 1991) (*disease—spread*) i.e., the tendency for certain words to co-occur and to associate strongly with each other in the language, e.g., “strong tea” or “powerful computer”, as opposed to “powerful tea” and “strong computer” which are not usually preferred by native English speakers.

Examples of cohesion-building devices are discussed in, e.g., Halliday and Hasan (1976), Hoey (1991), Halliday and Matthiessen (2004) and Tanskanen (2006). The cohesive profile, that is, the types of cohesion devices utilized by the text, varies across discourse types. A cohesion-creating item often appears in different collocations according to text variety (Halliday and Matthiessen, 2004). In our IE system, the semantic fields are defined in scenario-specific ontologies. For example, the nouns that appear with the verb *launch* or its synonyms, vary according to the scenario (Pivovarova et al., 2013). In the Rocket Launch scenario, the most typical noun to appear with the verb *launch* is *spacecraft* or *rocket*. In the Product Launches scenario in Business Domain, the most frequent object for the verb *launch* is the noun *product* or *line*, often preceded by the adjective *new*, see Section 2.4.3.

There is some interest in the textual function of collocation and how it varies across genres in corpus research (Gledhill, 2000). For example, certain collocations that create cohesion may characterize particular genres (Gledhill (2000) and Williams (2002)). The interaction of cohesion and discourse structure is linked with text type/genre, e.g., Gledhill (2000), Williams (2002) and Berzlánovich et al. (2009). For example, expository and descriptive (i.e., thematically organized) texts show higher lexical cohesiveness and closer alignment between discourse structure and cohesive structure than persuasive (more intentionally structured) texts (Berzlánovich et al., 2009). According to Tanskanen (2006), in academic writing (as a type of discourse) the use of lexical devices (e.g., repetition, synonymy) is less frequent than in casual face-to-face conversation and mailing-list texts, due to different communicative conditions.

Most of the patterns in our IE system could be seen as generalized collocations, based on an association between items and similar to “activity-related collocations” in Tanskanen (2006), defined as a relation between items based on an activity (e.g., eat—meals, drive—car).

The cohesive means define the overall coherence structure of the text. However, a cohesive text is not necessarily coherent since coherence is based on semantic relationships and pragmatics. Coherence in linguistics is what makes a text semantically meaningful and sensible to the reader, by relations between discourse segments, such as clauses, sentences, and chapters. Coherence is also dependent on the reader, how the reader perceives the text and what connections the reader makes. There are three components of coherence (Redeker, 2000), which help to define the structural relations of the text:

- R1. content relations** (additive, causal, temporal, contrastive, etc.), which in this work are reflected in the hierarchical relationship between events, see Section 2.4.2.
- R2. pragmatic or intentional relations** (evidence, justification, concession, etc.), which are harder to track, and are not analyzed or shown as such in our study.
- R3. sequential or textual relations** (summary, restatement, segment boundary, etc). In our study, these are exemplified by the division between *headline* and *header* vs. the body of the text, as a possible location for an event or a part of an event.

Tracing coherence in news articles is computationally challenging, because coherence in text is partially based on pragmatic knowledge. It is challenging for an IE

system to decide whether two or more events in the Disease Outbreaks scenario belong to same or separate outbreaks. The reader receives hints from the cohesive devices in the text, and from the reader’s world knowledge: the pragmatic knowledge. A set of cohesive devices were established for simple relationships between events in Publication IV.

The centrality of concepts in cohesive networks reflects their importance: lexical chains have been used as measures of centrality (Hoey, 1991; Tanskanen, 2006). The *repetition* features described in Section 3.4 are based on this observation and indicate high relevance of the event in the Infectious Disease Outbreaks scenario, as described in Section 3.4.1.

Cohesion and coherence are studied extensively in the linguistic literature, and computational models of cohesion have been attempted (Marcu, 2000). However, it is “not a fully understood issue in discourse organization” (Berzlánovich et al., 2009).

1.4 Related work

The scalability of a traditional non-customized IE system is in general very poor for different domains, e.g., (Etzioni et al., 2004). In Cardie (1997) it was stated that an IE system will work better if its linguistic knowledge bases are tuned to the particular domain. This had also been our experience in tuning IE systems for different scenarios.

The gap between results obtained in evaluation of IE systems for different domains and scenarios is the starting point for my work, and has led to the realization that the lexical and grammatical levels have an impact in how information should be extracted from the text.

The slow process of customization of a pattern-based IE system has been a principal bottleneck for wider application of IE. In pattern-based IE, the F-measure remains well below 80%. Each domain and scenario requires adding domain- and scenario-specific linguistic knowledge to the IE system. Manual modification of the knowledge bases is slow and error-prone. Finding a sufficient number of good patterns manually is a complicated process, and the system will under-extract if patterns are missing. On the other hand, precision remains low if the patterns are too general or brief (Grishman, 2019).

Training a language model on one corpus and testing it on another dissimilar corpus does not give good results (Kilgariff and Grefenstette, 2003; Sekine, 1997; Gildea, 2001). Even if the “simple” entities, such as locations or person names,

are correctly extracted, the performance of the system degrades when the training corpus is newswire and test corpus is, e.g., e-mails, blogs, social media or other colloquial style (Yates, 2009). Performance is better if the test corpus is similar to the training corpus. Traditional IE has been limited to a few different sources of texts, typically newswire, and a limited set of domains and no specific attention to text type.

Supervised machine learning has been used by many IE system developers. For example, the ACE evaluations (Automatic Content Extraction), introduced in Section 1.1.1, from 2001 onward, encouraged supervised-learning approaches, by providing manually annotated data to train machine learning algorithms, in some cases quite large datasets.

Cardie (1997) suggested that more attention to unsupervised methods is needed, as quick customizability to a new domain by end-users would be a desirable feature. She also advocates the insertion of pragmatic knowledge, such as temporal, causal, or other complex relationships among events. The relationship between the events is one of the main considerations in my work as well, discussed in Section 2.4.2. Unsupervised, or weakly-supervised learning methods are used in our system as well to produce candidates for domain-specific extraction rules, as described in Yangarber et al. (2000b). However, even with the unsupervised pre-selection, a considerable amount of manual work is still required.

The lack of scalability in IE is problematic, given the great variability in sources and styles on the Web, and especially in social media. The heterogeneous nature of language is hard to model even if limited to only one scenario. The interest in the Web as corpus increased the motivation to find alternatives to the slow and expensive customization processes in IE. Another reason for the increase of developing semi-supervised methods was the decrease of annotated training data from MUC competitions (Grishman, 2019). For example, in 2009 NIST (the US National Institute of Standards and Technology) organized the annual “Text Analysis Conference” and provided a large volume of *unannotated* data to encourage participants to experiment with semi-supervised methods (Ji and Grishman, 2011). The data consisted of news articles and blogs.

Certain linguistic phenomena, such as coordinated structures, coreference, passive constructions and named-entity recognition are difficult to model computationally. The “higher” the linguistic level, the more difficulties arise. Elliptical constructions, idiomatic expressions or sarcasm on the pragmatic level of language are challenging or even impossible to recognize, and those phenomena are very common in spoken language or in the language that is used in social media. Statistical models have been used to model linguistic phenomena, but such mod-

els have difficulties in modeling the more rare phenomena on the higher levels of language.

Also, in current research there is less emphasis on linguistic formalism (Grishman, 2019), and more reliance on data-driven approaches. More recently, the trend has been to extract in a language-independent fashion, and from practically any domain and genre, quickly and with minimal human effort. This is done by utilizing unsupervised or weakly supervised machine learning techniques in the hope of avoiding the scalability difficulties of traditional IE. These difficulties arise due to the diversity of text styles and genres and the considerable amount of work that creating a balanced test corpus for all domains would require (Yates, 2009; Etzioni et al., 2004). Extracting information with weakly supervised or unsupervised methods from all possible sources seems a very ambitious goal, based on my experience with IE customization.

It is easier to train an IE system on homogeneous data. However, in general, the data on the Web does not follow traditional divisions into text types (see Section 1.3.1), e.g., those proposed by Beaugrande and Dressler (1981): descriptive, narrative, argumentative, scientific, didactic, literary and poetic (Santini, 2006). This is a challenge for an IE system that extracts from any source in the Web. Open Information Extraction (OIE) tries to extract relations in a domain-independent way. E.g., Knowitall (Etzioni et al., 2004) tries to automate the process of customization, to be domain-independent, and extracts facts from the Web. This is done by using lighter-weight techniques without producing a logical representation of the text after parsing the sentences. Knowitall and Textrunner (Banko et al., 2007) generate new pattern candidates automatically or semi-automatically. This relates to research in ontology-based IE, e.g., Saggion et al. (2007), both in terms of using pre-existing ontologies to help customize IE systems, and in using IE to populate new ontologies.

Some work in IE aims to automatically recognize the domain to which the text belongs. Patwardhan and Riloff (2006) extracts domain-specific patterns from documents on the Web. The documents are chosen by applying patterns learned from a small training corpus (MUC-4). Semantic affinity to already existing patterns is computed for each pattern to automatically infer what type of information the pattern extracts. Riloff (1995) proposes an approach to recognize a domain through information extraction, using automatic pattern-finding and collocation-finding trained on annotated data. She observes that including features, such as a set of *stop words* is crucial for enhancing the results of determining the domain based on automatic pattern finding, as well as the voice of the verb (active or passive), etc.

The intuition of lexical associations across sentences is modeled in Soricut and Marcu (2006). Lexical similarity creates textual cohesion. For example, certain words (e.g., a disease name) used in a sentence tend to evoke the use of certain other words (e.g., *infected*) in the subsequent sentence. Cohesion exists between event types and event arguments: for example, launching a product is very different from launching a missile, so modeling interactions between triggers and arguments is important. Further, cohesion exists between adjacent events, for example, arrests often follow criminal activities (Li et al., 2013).

Related work for news surveillance is done, for instance, by Health Map (Freifeld et al., 2008). After potentially relevant documents have been selected by using keyword queries, experts have to check the relevance of the extracted events. Work related to the relevance of an event (or a topic, or a paragraph) has been done in summarization, e.g., by Strzalkowski et al. (1998). IE is extensively applied for business intelligence, e.g., by Cvitas (2010).

Maedche et al. (2005) describes how automatic intelligent Web exploration with the help of shallow IE techniques works for many domains. The suggested bootstrapping approach allows for the fast creation of an ontology-based IE system. Attempts have been made to extract a large number of scenarios and languages from the Web, e.g., Shinyama and Sekine (2006); Lucas (2005).

During the last 5 years since the publications in this work, neural networks (NN), especially deep and attention-based neural networks, have revolutionized the field of NLP. Performance obtained by NNs on many NLP tasks now surpasses the performance of other, previously popular methods. NNs are also used in both supervised and unsupervised settings (LeCun et al., 2015).

According to Grishman (2019), after more than 25 years of development and a wide range of new approaches in IE, the F score has only advanced from the low 60s to the low 70s on standard event classification benchmarks, so the problems addressed in this thesis remain very topical.

1.5 List of original publications

This thesis is based on the publications listed here and included at the end of this work. We present a short description of the content, the author’s contribution, and the additional value each publication contributes to this thesis. The author has been involved in the development of the Proteus and PULS IE systems for several years, participating in development and maintenance of multiple scenarios.

IE customization is a complex task involving building the knowledge bases, lexical resources, conducting linguistic analysis, data annotation and experiments.

1.5.1 Publication I

Ralph Grishman, Silja Huttunen, and Roman Yangarber. “*Information extraction for enhanced access to disease outbreak reports.*” In *Journal of Biomedical Informatics*, volume 35(4), 2003, pages 236–246

This paper presents the Infectious Disease Outbreaks scenario, (a.k.a. Medical scenario) with a detailed description. It contains a non-technical introduction to the Proteus-BIO IE system for the Medical scenario and the end-to-end work flow, from raw text input to the structured user-friendly database of extracted facts. My contribution to the paper is analyzing the texts and the features that are specific to this scenario, and customizing all knowledge bases for the experiments presented in the paper. My analysis exposed how this scenario differs from scenarios studied in earlier projects and publications.

I proposed the principles of splitting the representation, since in the Medical scenario, all the event arguments (such as entities, numbers, etc.) do not appear near the trigger sentence, but must be recovered from multiple sentences in the document. Therefore, I introduced the two-level representation, with “incidents” on the lower level, and “outbreaks” on the higher level. I conducted the reported experiments, error analysis, and participated in the process of writing the paper.

1.5.2 Publication II

Silja Huttunen, Roman Yangarber, and Ralph Grishman. (2002b). “*Diversity of Scenarios in Information Extraction.*” In *Proceedings of LREC: the 3rd International Conference on Language Resources and Evaluation*, 2002, pages 1442-1450

I am the primary author of this paper, and contributed most of the experimental design, most of the analysis, and most of the writing. The paper contrasts traditional style MUC scenarios, such as Management Successions and Terrorist Attacks, with the more complex post-MUC scenarios: Natural Disasters and Infectious Disease Outbreaks. (We refer to these and similar scenarios collectively as “Nature” scenarios) I observe differences in how these scenarios organize the sought information in the text. I describe the differences in the scope of a single event in the Nature scenarios vs. the Management Succession scenarios, and other traditional MUC-style scenarios. Events that take place across space and time form hierarchical relations, unlike the events that occur at one point in time (as

usually reported with business events, e.g., an acquisition of a company). These factors cause problems with the traditional template structure. A new modular template structure is proposed and inclusion relationships between incidents are defined.

1.5.3 Publication III

Ralph Grishman, Silja Huttunen, and Roman Yangarber. “*Real-time event extraction for infectious disease outbreaks.*” In *Proceedings of HLT 2002: Human Language Technology Conference, 2002*, pages 366-369

In this paper, I am one of the two primary authors. I contributed a major part of the text. The paper describes “Proteus BIO”, the end-to-end IE system for extracting information about infectious disease outbreaks, and how the process works in practice. I describe the structure of event data base (section 4), and the extraction engine (section 5) relating to pattern matching and event recognition. Normalization of disease names and time expressions, countries and cities is presented as well.

1.5.4 Publication IV

Silja Huttunen, Roman Yangarber, and Ralph Grishman. (2002a). “*Complexity of Event Structure in IE Scenarios.*” In *Proceedings of COLING 2002: The 19th International Conference on Computational Linguistics, 2002*

In this paper, I am the primary author. I contributed to the experimental design, analysis of the results, the structure and the text of the paper. I compare the Nature vs. the Business domains, describe the “scattering” phenomena, and propose the way to recover inclusion relationships between incidents.

1.5.5 Publication V

Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen, and Roman Yangarber. “*Assessment of Utility in Web Mining for the Domain of Public Health.*” In *Proceedings of the NAACL HLT 2010, Second Louhi Workshop on Text and Data Mining of Health Documents, 2010*, pages 29-37

In this paper I am one of the two primary authors. I proposed the idea of the discourse vs. lexical features, based on experiences from IE customization for real-world scenarios and based on linguistic literature. We introduce the system for providing decision support to the Public Health professionals in epidemic surveillance. The paper proposes the notion of considering IE results from the subjective

point of view of quality, rather than the traditional criteria of F-measure and correctness. I participated in the design of a classifier that rates the content of a document according to *relevance*: the utility of the extracted event to the user (on a Likert scale). The classifier uses discourse and lexical features as independent variables. The paper describes the PULS IE system, and user-centric issues form the basis and motivation for further research into relevance. The actual relevance features are not presented in this paper. I contributed to the error analysis and to writing the paper.

1.5.6 Publication VI

Piskorski, J., Atkinson, M., Belyaeva, J., Zavarella, V., Huttunen, S., and Yangarber, R. “*Real-time text mining in multilingual news for the creation of a pre-frontier intelligence picture.*” In *Proceedings of ISI-KDD: ACM SIGKDD Workshop on Intelligence and Security Informatics, at KDD-2010: 16th Conference on Knowledge Discovery and Data Mining*, 2010

This paper focuses on the domain of Cross-Border Security. My contribution is the customization of the IE system for all scenarios related to the Security domain focusing on cross-border crime, e.g., illegal migration, smuggling, human trafficking, etc. I contributed the text describing the event extraction processing chain, and the specifics of the Security domain. The event schema used in epidemic surveillance is similar to that used in the Security domain. However, in the latter domain, the schema is significantly more complex, and requires covering many similar and partially overlapping events. In this way it is similar to the Natural Disaster scenario. This is joint work with the JRC team, described in Section 1.6. It presents the combination of two systems, PULS and JRC’s own system, MedISys. The two systems are evaluated together.

1.5.7 Publication VII

Huttunen, S., Vihavainen, A., von Etter, P., and Yangarber, R. (2011). “*Relevance prediction in information extraction using discourse and lexical features.*” In *Proceedings of NODALIDA: the 18th Nordic Conference of Computational Linguistics*, 2011, pages 114-121

I am the primary author of this paper. I contributed to the experimental design, analysis of the results, the structure and the text of this paper. I describe the discourse and lexical features used for relevance prediction in detail. Lexical features are bag-of-words features. The discourse features include layout, posi-

tioning, event compactness, time and recency. In addition, “blacklist” features signal low relevance. We aim to identify the features that affect the importance of an event to the user.

The experiments indicate that relevance is a tractable measure of quality in the domains we studied. We presented prediction accuracies for discourse and lexical features separately, and in combination, using Naive Bayes and SVM classifiers.

1.6 Structure of the research

Each article provides a partial solution to the research problem. The first set of publications included in this work, 1.5.1, 1.5.2, 1.5.3, 1.5.4, forms the qualitative part of this study. In the papers, we describe the scenario-specific characteristics on the lexical level and on the structural level. While I customized the scenarios for the Business domain, the Nature domain, and the Security domain, I found differences between scenarios in how the text is organized in news reporting, and proposed a hierarchical template and cue words to improve recall.

The linguistic groundwork, the customization process and its scenario-specific characteristics are not described in detail in the papers because of the space limits. The customization process is similar to that described in Yangarber and Grishman (1997).

All customizations for post-MUC scenarios were done by me or under my supervision (apart from the first stage of the Natural Disaster scenario).

The second set of publications, 1.5.5, 1.5.6 1.5.7, describe machine learning approaches that aim at improving precision of the IE results. The methods described in these papers are embodied in systems that are now in continuous operation, used by real-world users in international partner organizations:

- JRC (Joint Research Center of EU) MedISys project since 2008
- ECDC (European Center for Disease Control) since 2008
- Frontex (European Border and Coast Guard Agency) since 2009,
- a large Finnish business news media company, 2008–2016,
- Agentum, 2008–2016.

PULS has been adapted to analyse texts in the epidemiological domain.¹³ PULS provides information to MedISys, a project of the JRC, for epidemiologists

¹³puls.cs.helsinki.fi

to track infectious disease epidemics around the world. Appendix C Figure C.1 shows a screenshot of the MedISys webpage,¹⁴ continually updated in real-time with the Infectious Disease Outbreaks scenario. Frontex¹⁵ analysts track border-security related events for security situation awareness at the EU external borders and in related third countries (Atkinson et al., 2010). Agentum feeds are varied news articles, company websites, product review websites.

¹⁴<https://medisys.newsbrief.eu/medisys/helsinkiedition/en/home.html>

¹⁵Frontex (www.frontex.europa.eu) is European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union.

Chapter 2

Analysis of cross-domain issues in IE

In this chapter I describe the process of finding, analysing and measuring the differences between the texts that report about the topic and the influence of the topic on the text. The chapter presents the methods, data, tools and the results. The process involves qualitative and quantitative methods. In part this is a description of the system development process, and in part it is a retrospective analysis of results of the development process.

First, I observe the linguistic and structural characteristics and the style of reporting for the particular scenario, for which the IE system is being customized. For the linguistic analysis I used a small training corpus, described in 2.11. The training corpus was analyzed both manually for inclusion relations (in Section 3) and for examining the lexicon on the semantic level, and computationally for word counts, and the Connexor FDG parser, described in Järvinen et al. (2004), for part-of-speech comparison between scenarios. The larger set of documents retrieved by the partners were used for testing and for additional information, e.g., synonyms and constructs to be added to the knowledge bases.

On one hand, we aim at improving recall. This is done by adding patterns and concepts for the scenario. The qualitative approach (described in more detail in Section 2.4) was partially deployed during the initial customization process, and consists of detailed lexical and structural scenario-specific analysis of the texts.

In the initial phase of customization, we used automated pattern-finding tools for finding possible candidate patterns.

After the customization phase, the texts are analyzed by the IE system. The compilation of a test and a training corpus, containing relevant and non-relevant documents, is crucial for the next phase: measuring the proportion of correct events extracted. These phases are repeated several times.

On the other hand, we are concerned with precision, the proportion of extracted events that are correct. In fact, we go further to evaluate which events are not only correct, but most interesting for the user, in Chapter 3. The work flow of the IE system is described, e.g., in Publication I.

2.1 Scenario characteristics and customization

In this thesis I focus on four domains of interest that are listed in Table 1.1 in Section 1.1.1. The four domains are here in the order in which they were studied and the results were published:

1. Natural Disasters,
2. Infectious Disease Outbreaks,
3. Business Domain:
 - Acquisitions,
 - Corporate Mergers,
 - Investments,
 - Product Launches,
 - Posts,
 - Layoffs,
 - Contracts and Orders,
 - Ownership,
4. in Cross-Border Security,
 - Smuggling,
 - Illegal Migration,
 - Human Trafficking.

Prior to these, the IE system had been customized for the MUC scenarios, such as Management Succession (Grishman, 1995) and Terrorist Attacks (Yangarber and Grishman, 1998a). The experience from customization of the earlier MUC tasks guided the later tasks.

2.1.1 Nature domain

The Nature domain in this study consists of scenarios such as Natural Disasters and Infectious Disease outbreaks. The first customization task was extracting facts about Natural Disasters from spoken and written news reports.

Natural Disasters

For this scenario, the task is to find occurrences of disasters around the world, as reported in newspaper articles. For example, we might need the following attributes for Natural Disasters: what type of disaster occurred (earthquake, storm, flood, etc.), where it occurred, when, how many people were killed or injured and what damages were caused (in quantified terms, if possible). The disaster may have several manifestations, e.g., storm, mud slide, or flooding. When this happens, we consider the overall event as consisting of several component *sub-events*, which are linked together to indicate that they are subordinate descriptions of the same overall natural event. An example of a disaster event is shown in Table 2.1, with two sub-events.

- “*Severe thunderstorms raked the Southeast with rain and golf ball-size hail and produced tornadoes that destroyed a Georgia motel and killed one person in a mobile home Sunday night.*”

Table 2.1: Example of a Natural Disaster event

<i>Disaster</i>	<i>Date</i>	<i>Location</i>	<i>NumberDead</i>	<i>AmountDamage</i>
thunderstorm	Sunday night	Southeast	–	–
tornado	Sunday night	Georgia	one person	motel

Following thorough customization, I compared the performance of scenarios in the Business domain, the Nature domains, and the MUC scenarios (Management Successions and Terrorist Attacks).

The automatic unsupervised pattern discovery system ExDisco (Yangarber et al., 2000a), applied for the scenarios performed worse on Nature domains than the Business domain. Automated pattern finding takes into consideration only the literal word sequences. If the text contains “noise,” e.g., appositional elements, or long temporal or locative adverbial phrases, or other explanative sequences, they will break the patterns and the automated process will not detect them.

Infectious Disease Outbreaks

For each Infectious Disease Outbreak we need to find the name of the disease, the number of victims, dead and sick, and the type of victim. In our implementation, the victims of infectious diseases could be people, animals, as well as plants; news reports often carry information about infestation of forests and plants important for agriculture. In addition, we track the location of the outbreak, the date, and possibly other information. As an example, the following text fragment is from a message posted on the ProMED list:¹

- “*Nine people have been killed and 630 hospitalized in Zimbabwe following an outbreak of the cattle disease, anthrax, which started a month ago [...].*”

Table 2.2: Example of Infectious Disease Outbreak

<i>Disease</i>	<i>Date</i>	<i>Location</i>	<i>VictimSpecies</i>	<i>VictimDead</i>	<i>VictimSick</i>
anthrax	started a month ago	Zimbabwe	humans	9	630

The customization process includes taking account of domain-specific knowledge and based on that creating concepts, an ontology, patterns, and inference rules for the scenario.

For example, a domain-specific knowledge base for Infectious Disease Outbreaks scenario includes a set of customized patterns and a sub-ontology of names of diseases, viruses, drugs, etc., organized in a conceptual hierarchy. From the sentence:

- “*More than 500 cases of dengue hemorrhagic fever were reported in Mexico last year, with 30 deaths*”

the system extracts an event in Table 2.3, by two patterns, seen in Table 2.4, **disease-reported** and **number-of-casualties**. The latter pattern is conditional, it is applied only if the first pattern fires. The status of the casualties is *infected*, *sick*, or *dead*. The argument **disease** is a name of a disease, or a disease-related

¹ProMED (<http://www.promedmail.org>) is a mailing list where medical professionals in the field contribute updates on epidemics and outbreaks of diseases around the world.

term, *disease*, *epidemic*; it may include the cases of diseases (with or without a number). The additional event arguments **location** and **date** are picked up by the sub-pattern “SSA” (scenario-specific adjunct), which can occur in several places in the sentence. The **number-of-casualties** pattern is used only if a **disease** is found nearby in the context.

Table 2.3: Example of Infectious Disease Epidemic event

<i>Disease</i>	<i>Date</i>	<i>Location</i>	<i>Victim Species</i>	<i>Victim Dead</i>	<i>Victim Sick</i>
dengue hemorrhagic fever	last year	Mexico	cases	30	500

Table 2.4: Example patterns for Infectious Disease Epidemic scenario

<i>Pattern</i>	<i>Pattern Syntax</i>	<i>Extracts</i>
disease-reported	SSA* np-head(disease_activity ,below) verb-group(C-report) SSA*	disease name, location, time
number-of-casualties	np-head(number ,only) 'of' np-head(case ,below)	number of casualties

2.1.2 Business domain

The business domain contains corporate related actions reported in the business news, e.g., changes in top-level corporate posts, companies acquiring other companies, companies launching new products, investing in new projects, etc.

Management Succession

The Management Succession scenario consists of reports about top-level corporate managers who have got new jobs or left their old jobs. This scenario emerged in MUC-6 in 1995 that focussed news articles on management changes (Section 1.1.1). The attributes of the event that we seek are: what position is changing hands,

who is leaving the post, who is assuming the post, why the change is occurring and what company it is. The following is an excerpt from a news report in Wall Street Journal (WSJ):

- *“In April, Drug Emporium retired its founding chairman and chief executive, Philip I. Wilber, and shunted Mr. Wilber’s successor, his son Gary Wilber, into the newly created post of vice chairman. The new chairman and chief executive, David L. Kriegle, assumed his post last week.”*

In Table 2.5 the extracted information is shown in structured form.

Table 2.5: Example of Management Succession events

<i>Person</i>	<i>Company</i>	<i>Post</i>	<i>Date</i>	<i>Status</i>
Philip I. Wilber	Drug Emporium	chairman	April	out
Philip I. Wilber	Drug Emporium	CEO	April	out
Gary Wilber	Drug Emporium	vice chairman	April	in
David L. Kriegle	Drug Emporium	chairman	last week	in
David L. Kriegle	Drug Emporium	CEO	last week	in

Corporate Acquisitions

The Corporate Acquisitions scenario covers reports about companies buying other companies. The attributes of the event that we seek to extract are: the seller, the buyer, the item that is sold, price and date. The first two are most commonly companies, although they can be persons as well. The item must be a company, not, e.g., a government organization. For example,

- *“Westinghouse Electric Corp. said Wednesday that it had agreed to sell its military and electronic systems businesses to Northrop Grumman Corp. for \$3 billion in cash. The announcement closely follows Westinghouse’s sale late last month of the Knoll Group, its office-furnishings unit, to Warburg Pincus Ventures L.P. for \$565 million. “*

produces events in Table 2.6.

Table 2.6: Example of Corporate Acquisitions events

<i>Seller</i>	<i>Buyer</i>	<i>Item</i>	<i>Price</i>	<i>Date</i>
Westinghouse	Grumman Corp.	military and electronic systems businesses	\$3 billion	Wednesday
Westinghouse	Warburg Pincus Ventures L.P.	Knoll Group	\$565million	late last month

Product Launches

The Product Launches scenario consists of reports about companies launching new products to the market. The attributes that we try to extract for each event are: the company, the new product, date and country. The product (the *item*) could be a wide variety of noun phrases. Certain noun phrases are automatically disqualified as products, due to their placement in the domain ontology (i.e., the concept hierarchy), such as, dates, company names, locations, etc. The examples in the following sentences produce the events in Table 2.7.

- “*Now that it is almost certain that Samsung will unveil its Galaxy S5 smartphone during the Mobile World Congress 2014, all eyes are on what’s actually going to emerge from Samsung’s stable.*”
- “*Samsung, which will unveil a high-end Galaxy phone next week, climbed the most since August.*”
- “*An executive at T-Mobile said the company was introducing its new DriveSmart service at the request of customers.*”

Note that *Samsung* is normalized into the full name of the entity *Samsung Electronics*. This is done to try to assure that the database of extracted facts is consistent and can be searched easily using unified, “canonical” names for all entities, when they can be determined.

2.1.3 Security and Cross-Border Crime

The Cross-Border Security domain in PULS contains several scenarios of border security that are developed and evaluated together, because they form parts of

Table 2.7: Example of Product Launches events

<i>Company</i>	<i>Products</i>	<i>Price</i>	<i>Date</i>
Samsung Electronics	Galaxy S5 smartphone	—	2014
Samsung Electronics	high-end Galaxy phone	—	2014.02.23–2014.03.01
T-Mobile	DriveSmart service	—	

the same task for the same end-user, Frontex—the European Border and Coast Guard Agency. The Cross-Border Security domain is described in Publication VI. Because the tasks share many linguistic characteristics and have much semantic overlap, they also share the knowledge bases. The differences lie in the lexicon and concepts. The scenario to which a given trigger sentence belongs depends on the type of the perpetrator, action (performed by a perpetrator) and/or other possible entities (e.g., goods) involved. A separate component in the IE system applies “inference rules,” which are logical rules for transforming some of the event attributes based on the values of other attributes. For example, if the text mentions drugs or weapons, then the event’s predicate might be changed from “illegal entry” to “smuggling.” The scenarios are

- illegal migration incidents (illegal entry attempts, illegal stay)
- refusal of entry
- uncontrolled exit
- cross-border criminal activities
- human trafficking (forced labour, prostitution, child trafficking)
- smuggling (excise goods, human organs, drugs, radioactive substances, weapons etc.)²

The following news extract is an example of a *smuggling* event and the attributes sought from it.

²The PULS system also extracts events about other crisis events, e.g., terrorist attacks and attempts, kidnappings, explosions, arrests and court sentences, but the evaluation of the performance of the system for those scenarios is beyond this study.

- “*DUSHANBE, January 27, 2012, Asia-Plus - Tajik law enforcement authorities have seized some 106 kilogram of narcotics in the southern Khatlon province.*”

Table 2.8: Example of Cross-Border Security: Smuggling event

<i>Crime</i>	<i>Date</i>	<i>Location</i>	<i>Suspect</i>	<i>Object</i>
smuggle-drugs	Jan 27 2012	southern Khatlon province, Tajikistan	—	narcotics

Terms and concepts have to be defined in the ontology, and the clause-level patterns have to reflect syntactically and semantically how the essential information about the topic is expressed. Modifications are needed also to the inference rules, which combine events, items, and relations between them.³ This generally involves detailed analysis and manual intervention.

The performance of the IE system, evaluated by F-measure, was poorer on the Natural Disaster scenario than on the MUC scenarios about joint ventures, management succession, or satellite launches. After the poorer results in the Natural Disaster scenario,⁴ I compared the pattern types and news reports in the Natural Disaster and the Infectious Disease Outbreaks scenarios against the Nomination scenario (a simplified version of the management succession scenario) to understand the differences in performance. For each scenario I compared the type of events that the IE system failed to find or overgenerated, on the training corpora. To improve the results, we changed the template structure (Section 2.4.2). I compared events and texts for linguistic and structural differences for all of these scenarios, discussed in Section 2.4. I studied the general and domain/scenario-specific characteristics of the texts in detail, as well as other text-type related characteristics that could have an impact on the way the facts are expressed and organized in the text.

I reported on the scenario-specific characteristics, and solutions to the problems of “event scattering” and sub-events in Publications II and IV. The aim of the analysis is to extract events and their related attributes, and to restore the cohesion and thus achieve more integrity of facts found in the text, and fill the database with fewer gaps and more complete events. We aim for cohesion, on one hand,

³These aspects of pattern-based IE are described in greater detail in the papers.

⁴Natural Disaster scenario was initially customized at New York University by the author, B. Megyesi and S. Rydin from the University of Stockholm in 1999.

between an event and its attributes, and, on the other hand, between different events.

The qualitative analysis of a text in a scenario focuses on the relevant event and parts of the relevant event. The aim is not to provide an exhaustive analysis of a scenario or a text, but rather to improve the performance of the IE system, by understanding and defining the characteristics that need to be taken in consideration in order to improve the quality of IE.

As an outcome of the analysis of the variation in performance between the scenarios and texts, the differences were classified on **the structural and the lexical** levels. On the structural level the major complication is that an event and its attributes do not necessarily appear in a single trigger sentence, but are more commonly spread over the text in a more or less systematic way. On the lexical level, the complication consists of the type and amount of variation in the linguistic expressions used in the different scenarios (see Section 1.3.2 L3, L4, L5). On the structural level, I detected the cues, which are the cohesive devices used by the writer to link the story together (see Section 1.3.2 R1,G1). The cues indicate relations between items. On the lexical level the links are in form of repetition (see in Section 1.3.2 L1) and semantic relationships between the expressions used. The reader sees the connections between the essential items in the text not only because of pragmatic knowledge about world, but also because the writer of the article provides clues to help guide the reader to link the pieces of information together properly. The idea of extracting not only sentences, but also smaller spans, and then identifying the linking elements between them is not new in IE. The ACE program's research objectives "are viewed as the detection and characterization of Entities, Relations, and Events,"⁵ that is, finding entities first, then the links between them. This is in contrast with the traditional MUC tasks, where the objective was to find entire events.

In summary, certain types of scenarios produce worse results than others in terms of the traditional MUC-type F-measure evaluation when the same IE system is applied. The results are worse for the Nature scenarios which include Natural Disasters and Infectious Disease Outbreaks.

2.2 Data

The data consist of general news articles, abstracts of new articles, transcribed broadcast news, web pages and e-mail messages from a mailing list.

⁵<http://projects.ldc.upenn.edu/ace/>

Table 2.9: Amount of data for initial customization

<i>Data Source</i>	<i>Amount of data</i>
Wall Street Journal	about 10000 articles
AP	about 10000 articles
NYT	about 200Mb
CNN, NPR	about 1Gb transcribed TV and radio

For the customization and discovery procedures we used data from various sources: general news articles reported by The New York Times and Associated Press; domain-specific news reported by The Wall Street Journal, Federal Register, World Health Organization and ProMED; Radio and TV news programs for the spoken register, broadcast by CNN (TV) and National Public Radio. With Proteus, we used material from the time periods of 1987–1995, which is divided as follows: WSJ, amount approximately 10 000 articles, AP, amount approximately 10 000 articles, NYT about 200Mb, CNN and NPR together about 1Gb of transcribed TV and radio broadcast news. Although the reports came from different sources, they seem to form a relatively unified category, namely *news reporting*.

Data for the Business domain consists of three corpora as is shown in Table 2.10. Before 2003, the corpus for management succession consists mostly of Wall Street Journal articles. The second corpus consists of **abstracts** of news articles, written in English,⁶ provided by a partner in the PULS research project (see Section 2.3). The third corpus consists of large amounts of increasingly varied news articles, company websites, reviews from product reviews websites, provided by another partner, Agentum. This is the Business Articles corpus.

The Natural Disaster scenario was part of the Event-99 Project (Hirschman et al., 1999). The data for the Natural Disaster scenario was the only one that included transcribed broadcast reports, from ABC, CNN, PRI, VOA (Voice of America). The written news were from the AP (the Associated Press) and The New York Times. For the Infectious Disease Outbreaks (also called Medical scenario) the material consisted of the ProMED mailing list, and the WHO (World Health Organization) web reports of disease epidemics; and articles from The New York Times. The data stream is provided by JRC, where PULS has been adapted

⁶Some of the writers are news analysts, native speakers of Finnish.

<i>Domain</i>	<i>Sub-corpus</i>	<i>Data Source</i>	<i>Type</i>
Natural Disasters		ABC, CNN, PRI, VOA; AP, NYT	transcribed broadcast news; news articles
Infectious Disease Outbreaks		ProMED, WHO, NYT (JRC)	mailing lists, news reports
Security		news feed from FRONTEX	mailing lists, news reports
Business (Management Succession)	1	WSJ	news articles
	2	PULS partner	Abstracts
	3	PULS partner, Agentum	news articles, websites, etc.

Table 2.10: Type of data for Domains

to analyse texts in the epidemiological domain, which has been in use since 2008 by European Union member states through the MedISys web-site.⁷

For the Security domain, the news feed originates from FRONTEX.

The type of data is listed in Table 2.10.

The *training corpus* for a linguistic analysis consists of documents provided by the MUC organizers for the Natural disasters scenario, and a blind selected sample of documents in the other scenarios (see 2.10). The IE system was continuously tested on live data.

2.3 Setup and tools

The Management Succession, Corporate Acquisitions and Natural Disasters in this study were developed using the Proteus IE system (Grishman, 1997) developed at the Department of Computer Science at New York University. Previously, the Proteus system has been customized for several news scenarios, as part of the MUC program.

The Medical scenario about the Infectious Disease Outbreaks, and the scenario of Acquisitions in the Business domain were partially developed under the Proteus

⁷<https://medisys.newsbrief.eu//medisys/helsinkiedition/en/home.html>

<i>Scenario</i>	<i>Number of documents</i>	<i>Words in total (min – max)</i>	<i>Median</i>
Natural Disasters	17	8415 (38–3113)	548.0
Infectious Disease Outbreaks	29	7433 (57–839)	195.5
Security	29	7777 (42–764)	182.0
Investments	19	1763 (30–346)	82.0
Product Launches	16	1167 (37–121)	70.0
Contracts	35	2680 (27–130)	69.5

Table 2.11: Training corpus for the linguistic analysis

project and partially under the PULS project. Manual and weakly supervised methods were employed for pattern-building.

The IE system receives a stream of potentially relevant documents from the system’s information retrieval (IR) component that continuously polls news sites, for example, for the Medical scenario (Atkinson et al., 2011). The news filtering is done using Boolean keyword-based queries.

Examples of scenarios for which systems have been developed using Proteus or PULS are Management Successions (MUC-6, 1995; Grishman, 1995), Rocket Launches (MUC-7, 1998; Yangarber and Grishman, 1998a), Airplane Crashes (MUC-7, 1998) and Terrorist Attacks (MUC3/4). Subsequently to the MUCs, the Proteus IE system was customized to extract Corporate Mergers, Corporate Acquisitions, Natural Disasters, see the Event-99 specifications in Hirschman et al. (1999); it was also customized for Infectious Disease Outbreaks, IFE-Bio (Publication I).⁸ The customization process for the latter scenario utilized unsupervised methods for pattern acquisition (Yangarber et al., 2000a,b) as well.

In our IE system, the knowledge bases (KBs) consist of a scenario-specific lexicon, supplementing the base English-language COMLEX lexicon (Macleod et al., 1994), and scenario-specific concept hierarchy, in which the lexical entries are organized into semantic classes, so that the patterns can refer to these concepts; scenario-specific predicates, which are the logical structures which are filled when patterns match. Domain-independent knowledge, such as proper nouns referring to names of organizations, countries and cities, are listed in a separate

⁸Initiated as part of DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program about Bio-Security (IFE-Bio).

sub-ontology. PET example based customization tool (Yangarber and Grishman, 1998a) provides a pattern editor for building syntactically and semantically generalized pattern sets (Yangarber and Grishman, 1998b).

2.4 Results of the linguistic analysis

Defining IE guidelines (see Section 2.1) for the Natural Disasters scenario was not straightforward. A typical event in the earlier MUC scenarios consists of attributes that occupy one slot each in the event template. By contrast, a typical Natural Disaster event could consist of multiple attributes per slot, and, on the other hand, multiple slots may not be filled at all. The amount of synonymy⁹ in the lexicon was greater, and the amount of patterns needed was larger. Although we applied the same customization process as in developing the MUC scenarios, and despite adding a greater number of concepts and patterns, the results did not improve. The F-measures remained lower, compared to the MUC Nominations scenario, as seen in Table 2.12. The evaluation for template filling and computation of precision, recall and F-measure was performed as explained in Section 1.1.2.

When P, R and F measures are compared for different scenarios it raises the question to what extent they are commensurate, in reference to the template structures being filled in each case. For example, one scenario’s template may contain twice as many slots as another.

If numbers were computed only for templates filled *completely* and correctly and then direct comparison would not be fair. But in our evaluation, Section 1.1.2, we count the total of correctly filled slots in all templates in the optimal mapping. Therefore, comparing the percentages of filled slots makes sense.

Table 2.12: Performance for a MUC scenario and Natural Disasters scenario

<i>Domain</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>
MUC/Nominations	72	84	77.36
Natural Disasters	36	58	44.23

⁹The term “synonym” is used in a domain-specific sense. In the concept hierarchy, we group terms that serve the same function in the domain-specific patterns. For example, in the Nominations scenario, various verbs may indicate that a person quit her position: verbs such as *quit*, *step down* and *die*. In the context of this scenario, we refer to these verbs as “synonyms” in the scenario-specific sense.

The corpora to which I’m referring here are manually annotated, and used during the customization process. This data set is small (on the order of 100 documents), and divided into “training” and “test” corpora. The training corpus is used continually during the customization and studied in detail. The test corpus is used only occasionally for more accurate evaluation, and was kept “blind” to avoid biasing the developers.¹⁰ Analysis of the texts in the Natural Disaster scenario showed that they appear to exhibit a more fragmentary reporting style and “scattered” events. As a result, the automatic acquisition of patterns and lexicon, which was applied to speed up the customization process (Yangarber et al., 2000a,b), also performed worse on this scenario.

The Natural Disaster scenario corpus is not uniform. There were lexical and structural variations between the texts. At first this variation seemed to depend on whether the documents were spoken or written. However, in the Infectious Disease Outbreaks scenario, where we used only written text, similar structural and lexical tendencies appeared as in the Natural Disaster scenario. When the IE system was applied to the Business domain, performance varied not only according to the scenario, but also according to text type, that is, business news articles vs. abstracts based on news articles (see Section 2.4.3). The two kinds of Business domain data should be treated separately.

In fact, for Natural Disasters the spoken data gave slightly better results than the written data, as Table 2.13 demonstrates. The results, however, are not very reliable since the amount of spoken documents in the sample corpus was relatively small, as seen in the table.

Table 2.13: Performance for Natural Disaster scenario

<i>Domain</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>	<i>No of Documents</i>
Disasters Spoken	81	41	54.64	18
Disasters Written	34	26	29.46	10
Disasters All	58	36	44.23	28

Thus the linguistic differences are attributable not only to the channel—written or spoken—but also to the scenario. Also, the data on different scenarios in written language was abundant, but (transcribed) spoken material was scarce,

¹⁰The automatic discovery procedures for domain-specific lexicons and patterns was done using much larger corpora, since they don’t require labeled data.

which prevented an exhaustive quantitative study of spoken language. Rather than analyzing the differences between the spoken vs. written news, focus shifted to the more subtle differences between the scenarios and the style of reporting (Publications II and IV).

2.4.1 Lexical analysis

On the lexical level the Nature scenarios use more metaphorical expressions, less predictable constructs, with less concentrated statement of facts, synonyms (especially verbs) that belong to a different semantic field in their core meaning, and personification of non-human participants. These factors make it harder for the IE system to capture the events of Nature scenarios.

The wording of a business event is more compact and less ambiguous. The writer of routine business news is often an editor, whereas the articles about catastrophes are written by a named journalist (Gregory, 2001). This has an impact on the style in which the article is written. The characteristics of the texts are analyzed on several levels, as was introduced in the Section 1.3.1 on linguistic background about genre and text types. On the experiential level, a piece of news about a hurricane and one about a corporate acquisition are events in the world that include participants, processes and circumstances. They provide an answer to a question what has happened and where. On the interpersonal level, however, the attitude of the writer toward the subject matter and the reader's expected reaction influences the lexical and structural choices the writer makes, that is, the style (Bakhtin and Ghāsemipour, 2011). The attitude of the writer of news about a hurricane toward the subject matter, and the attitude of the reader, is more intense. But the writer of a business news hands over a piece of information to the reader without showing her attitude toward the subject matter.

The writer takes the role of a storyteller when reporting about a natural disaster. She aims at capturing and holding the reader's attention by evoking emotions. The lexical choices differ as well. A hurricane is represented as an emotional, tragic and colourful event, with twists in the plot, whereas a corporate acquisition is a neutral, straightforward event.

The examples below are typical of the business news domain:

- “*CLINTON, N.J. – Foster Wheeler Corp., an engineering and power-plant company, said it will acquire Enserch Environmental Corp. in a deal valued at \$90 million to \$95 million.*”

- “*National Semiconductor Corp. has agreed to sell half its mainframe computer business to Memorex Telex NV for \$250 million in cash and 4 million shares of Memorex Telex stock, the companies said Monday.*”
- “*MIAMI, Fla. – Intercontinental Bank acquired Interstate Bank Holding Co. and its subsidiary, the Bank of Coral Gables, for \$15.9 million.*”

Regardless of whether the register is spoken or written, the lexical choices in the natural disaster domain are more often metaphorical than in business news, for example:

- “*Cars lay buried in the muck, and mobile homes were **tossed together like the cars of a derailed train.** The sides of a water slide were **chewed up like a dog toy.**”*
- “*Tidal surges occasionally flooded coastal areas, and **roofs were stripped off beachfront condos, exposing their furnishings like children’s doll-houses.**”*

In Natural Disasters, expressions using lexical terms that are often seen in reporting about war/military conflicts, especially the verbs, are often used to describe the occurrences and effects of a disaster as movements and actions of a personified agent, as in the following sentences from news reports:

- “*Canada’s devastating ice storm **loosened its grip today**”*
- “*Severe thunderstorms **churned up several tornadoes and prompted flooding [...]**”*
- “*Heavy thunderstorm **gave Texans a drenching Tuesday [...]**”*
- “*Storm **knocked power out in central Indiana [...]**”*
- “*Snowstorm **killed one person in a mobile home Sunday night**”*
- “*Storms and tornadoes **besieged [...]**”*

The disaster is personified and the verb phrase that usually requires an animate agent is used metaphorically to describe the actions of the disaster. Often disasters are also named (e.g., *El Niño*, *Hurricane George*), which adds to the personification.

The Disease Outbreaks scenario also personifies diseases, but since diseases are caused by bacteria or viruses, which are living organisms, the personification may seem more justified.

- “[...]public health and municipal workers battling a rising tide of a disease once thought eradicated from this part of the world **dengue hemorrhagic fever**. It is **the deadly first cousin** to the more common dengue fever”

For locative and temporal expressions, the location of the disaster often functions as a patient, but is sometimes perceived as an agent as well. It can function as a subject of a verb that usually requires an animate actor, as in the following news fragment:

- “*Biting wind, snow and bitter cold continued their assault on the northern Plains Saturday, while parts of **Tennessee and Ohio enjoyed** record highs.*”

Special problems for reference resolution appear when personal pronouns are used impersonally. In the following news report a period appears in the middle of the sentence when the reporter decided to continue the phrase after an unusually long pause (alternatively, this could be analyzed as a sentence fragment):

- “*in kentucky and ohio, **they** are cleaning up after a powerful storm that dropped record amounts of snow . **more than 22 inches in some places.***”

Locations and temporal expressions have vaguer extent in the Nature domain, and are thus expressed in a more descriptive and lengthy manner than in the Business scenarios. Table 2.14 shows the number of locative expressions for Nature Domain and Business scenarios, and Security Domain in the training corpus.

The locative expressions that occur in Nature Domain and Security Domain are much longer on average, compared to the Business Domain scenarios. The selection of prepositions indicate vaguer limits for the location in the Nature scenarios than in Business news.

In general, an occurrence of a natural disaster in time and space is not perceived as punctual. A punctual incident is one that covers a specific point in time and space. For example an occurrence of an acquisition of a company, or a change of post in a company in the Business domain is typically presented as punctual in the text. In the following example there are locative expressions from Natural Disaster reports in 1–2, and 3–4 are from Disease Outbreaks:

Table 2.14: Amount of locative expressions and types of prepositions in six scenarios. *Length* in the average number of words in a locative expression.

<i>Scenario</i>	<i>Locative expressions</i>		<i>Prepositions</i>	
	<i>#</i>	<i>length</i>	<i>#</i>	<i>most frequent</i>
Natural Disasters	122	5.1	15	in, across, from, at, toward, near, over
Infectious Disease Outbreaks	133	3.8	11	in, from, on, to, within, on to, near, throughout
Security	152	3.3	15	from, in, to, into, at off, through, via, for
Investments	49	2.9	4	in, to, at
Product Launches	24	1.8	4	in, of, onto
Contracts	71	2.0	2	in

1. “*the hills above Guerneville, California*”
2. “*The two top corners of the country*”
3. “*villages near Caspian sea*”
4. “*in Kauakaul and Rajaul blocks in Bihar’s Nawada district*”

Temporal grounding is expressed in a less punctual way in Disaster and Disease Outbreaks reports than in Business news. Table 2.15 shows the amount of temporal expressions and the related prepositions: for Nature domains, Security domain, and for three scenarios of the Business Domain. Comparing with the amount of data for these scenarios, Table 2.11, we see that temporal expressions are more rare in the Business scenarios. Temporal expressions are divided into two types: “simple” expressions, which can be captured by very simple patterns vs. “complex” expressions, which require much more complex temporal patterns. Examples of simple expressions are: “November 2010” or “2012”; these were common in Investments and Contracts scenarios. Examples of complex expressions are “from 1 July 2010 to the end of 2012,” “by Summer 2012” or “in the third quarter of 2010.”

The two occurrences of temporal expressions in the Product Launches scenario are “the year-end shopping season of 2010” and “before the end of 2008”, describing the time of the launch of a new production line to the market.

Table 2.15: Amount of temporal expressions and related prepositions in six scenarios. In parentheses are the numbers of “complex” temporal expressions. *Length* in the average number of words in a temporal expression.

<i>Scenario</i>	<i>Temporal expressions</i>			<i>Prepositions</i>	
	<i>total</i>	<i>complex</i>	<i>length</i>	<i>#</i>	<i>most frequent</i>
Natural Disasters	49	(47)	3.0	10	from, before, through, by, after on, over, since, at (by late)
Disease Outbreaks	53	(32)	3.4	8	on, since, between, from, after in (this, so far, less than)
Security	72	(49)	2.8	14	on, in, of, throughout, after, by since, during, for, from, within
Investments	24	(6)	2.6	5	in, by, during, around, through
Product Launches	2	(2)	6.0	3	of, before
Contracts	20	(8)	3.0	7	of, by, over, to, from, between

The lack of any preposition is also included in the preposition count, as well as the prepositions, for practical reasons.

The Nature and Security domains have more varied expressions of time, and the selection of prepositions used reflects the non-punctuality of the event.¹¹

Below are examples of temporal expressions: 1–2 are from Natural Disaster scenario, and 3–5 from Disease Outbreaks:

1. “*since the weekend*”
2. “*by Friday night*”
3. “*to Mid-October*”
4. “*in March, April, and May of 1995*”
5. “*between June and October in the years 1990–93*”

These temporal expressions refer to spans of time with vague limits. The differences in temporal and locative expressions in the Business news and Natural

¹¹Some temporal expressions themselves are hard to delineate, e.g., “Not *until their night of fear dissolved into a pearly dawn Friday* could Carolinians held hostage to Hurricane Fran know the depths of the disaster.”

Disaster reports reflect the different discourse worlds of the respective domains. The location and time may be more important to the reader of Natural Disaster reports than they are to the reader of Business domain.

Transmitting news verbally brings additional features to the text. In the speech of the news reporters, and the upset, excited or scared people who are interviewed after a disaster, we frequently find colloquial characteristics, like slips of the tongue, repetitions, corrections and less formal syntactic structures, as seen below:

- *“in california, it’s el niño causing trouble again. the second of three el nino fueled storms hit northern california today, destroying property and dampening spirits.”*

In written register the previous example would be *“el niño is causing trouble again in california”* as the written register prefers standard syntax and prefers unmarked word order as in the following:

- *“A brutal northeaster thrashed the Eastern Seaboard again Thursday with cold, slicing rain and strong winds.”*

Elliptic phrases are common, as in the following example, and are problematic for information extraction:

- *“no electricity, and there is is it is very cold at home”¹²*

Lexical and structural characteristics of the texts had an impact on how well the pattern-based IE system performed on the texts. False starts, repetitions, and corrections are frequent in spoken language. IE patterns are easily “broken” because of the noisy data. Constructing the patterns to match short expressions is one way of preventing the breakage, as described in Section 2.4.2

Generally, it is easier to make patterns with a closed set of terms. If the text contains a vast number of terms, it is necessary to create elaborate ontologies. For unrecognized proper nouns, it is safer in Business news to assume that they are company names, because the domain is more formulaic, and only the highly relevant information is given, not unnecessary information, which could confuse the reader.

Table 2.16 shows a breakdown in the corpus for the four scenarios, namely, Natural Disasters, Infectious Disease Outbreaks, Product Launches and Investments:

¹²This follows the original transcription.

the number of connectives, adjectives, verbs (with the number of past tense verbs in parentheses), pronouns, and their percentages. In the Infectious Disease scenario, 212 of the 361 adjectives are distinct (†). In the Product Launches scenario, of 45 distinct adjectives, 17 are the token *new* (††). This corpus is slightly different from a corpus used in an earlier experiment, shown in Table 2.20: some documents were added for the Natural Disaster scenario, and some were removed for the Medical scenario.

Table 2.16: Parts of speech per scenario

	Natural Disasters		Infectious Disease Outbreaks		Product Launches		Investments	
Connectives	3.1%	286	2.47%	209	3.1%	45	2.5%	62
Adjectives	3.9%	367	4.2%	361 [†]	7.1%	104 ^{††}	4.0%	101
Verbs	11%	1063	11.0%	925	9.7%	143	9.0%	227
(past tense)	(55%)	584	(58.5%)	541	(33.6%)	48	(33.9%)	77
Adverbs	5.4%	504	4.2%	358	1.9%	28	2.4%	60
Pronouns	5.6%	525	4.3%	370	2.9%	43	2.4%	60
(1.person)	(8%)		(8.4%)		(0%)		(0%)	
Words		9374		8434		1471		2510

There are more verbs in Nature scenarios than in the selected Business scenarios. Especially the amount of past tense forms is considerably larger in Nature scenarios. Past tense verb forms indicate narrative text type, e.g., Schiffrin (1981).

The verbs are used in patterns to describe the actions of the main actor, that is, a disease, a disaster, or a company, depending on the scenario at hand. For Natural disasters there are 34 verbs that describe the actions of a disaster, e.g., *assail*, *attack*, *blight*, *lash*, *plague*, *trash*, *topple*. In the corpus there were 374 types of verbs, of which 85 had relate to Disasters. For Infectious Disease Outbreaks, 51 verbs were used in the patterns, e.g. *hit*, *infect*, *paralyze*, *strike*. In the corpus there were 317 different verb types. For Investments, 14 verbs were needed to describe the action of a company, and for Product Launches 18. The number of verb types in the corpus was 104 for Investments and 92 for Product Launches. For Security, the patterns were constructed from three perspectives (the perpetrator,

the authority, and the victim), so the numbers are hard to compare. The number of different verb types for Security was 329.

The largest amount of pronouns is found in Natural Disasters and Infectious Disease Outbreaks scenarios. The type of pronouns differ as well. In both Nature scenarios the first person pronouns cover about 8% of all pronouns but in the Business scenarios they do not appear at all. The first person pronouns generally indicate interpersonal focus and are used for comparison of spoken and written registers, see, e.g., Biber (1988):225.

Adverbs appear more commonly in informal texts and speech than in formal elaborate texts (Biber, 1988). The Nature domain contains twice as much adverbs as the Business domain. The variation in adverbs ending *-ly* (Biber calls them “total adverbs”) differs in Natural Disasters and Investments so that in Natural Disasters there are 77 tokens and 45 types. The most common adverb is *neatly*. In Investments scenario there were 16 tokens and 15 types.

2.4.2 Structural analysis

In this section I present the structural analysis. I demonstrate the domain-specific organization of events in the text and how the text is structured to convey the facts in Business and Nature domains. This is reported in Publications II and IV.

Events comprised of components that do not appear together in one sentence or paragraph are problematic for pattern-based IE. The scenarios show variation in how events, or components of events, are organized in the text. The *scattering* phenomenon means that the parts of the events may appear far apart from each other. The scattering is calculated from event templates’ pointers into the source texts from the first character of an event attribute to the beginning of a last event attribute in one text, as seen in Table 1.2.

In the Nature scenarios, scattering was more frequent than in the MUC tasks and the Business domain.

Scattering is partly due to editorial structure. The scattering of the components of events can be understood in light of the fact that the order of presentation of instances of an event in a news article is dictated by their presumed *relevance* to the reader, rather than by the causal relationships that may take place between them. This is very hard to model computationally. Often, cause and effect are in separate clauses, sentences, or even paragraphs. The “inverted pyramid” model (Bell, 1991), used for writing in news articles, is a story structure that places the most important details at the beginning, and introduces details of perceived lesser

importance in successive paragraphs (Scollon, 1998). This is also evident in this study (see Section 3.4, and Figure 3.2 in Section 3.4.1).

Scenarios also vary in the hierarchical relationships between concepts in the trigger sentences. Nature texts frequently contain more related events than Business texts. The information in the various mentions is overlapping, forming *inclusion* and *causality* relations (Huttunen et al., 2002). Inclusion means that one incident may include others, which give further detailed information. The inclusion may be e.g., temporal, geographical, or numerical – amount of damage or number of victims.

Business news structure

In the Business scenarios, the focus is on the actual change of post, or ownership of a company, as opposed to, e.g., the atmosphere in the company when the news was received, or the possible consequences. The main fact usually appears at the beginning of the report, without much introduction. All components of the fact are typically in one paragraph, or even in one sentence. Location in Business news is rarely mentioned, since it is not of importance in this scenario. If the date is not mentioned in the article, it can be inferred from the document date, since it is assumed that the date of the incident is close to the publication date.

The following news extracts present typical news fragments (WSJ):

1. *“The parent of RKO General announced Thursday it would sell Los Angeles radio stations KRTH-AM and FM, the latest step in the FCC-ordered dismantling of the RKO network.”*
2. *“First Interstate Bancorp gave in to Wells Fargo & Co., agreeing to be purchased for \$11.6 billion.”*

The remainder of the example 1 (originally a 13-sentence-long article), describes the background of the seller and the buyer in a few sentences, and the history of the previous purchases and sales.¹³ The second example is a very typical way to express a business event. It has all the necessary components in one sentence in a typical order. However, the facts to be extracted are not always in the same sentence, or not even in the same paragraph even in the Business scenario. For example, the price is sometimes found in the following paragraph, as below (WSJ):

¹³The purchases and sales were part of the information extraction task, but, typically, only the recent ones are relevant for the user. The background for purchases and sales is less relevant.

- “*WASHINGTON – The Justice Department signaled it is within days of completing an antitrust agreement that could clear the way for **AT&T Corp. to acquire McCaw Cellular Communications Inc.** The department asked a federal judge for an extension until Monday to file its recommendation in **the planned \$12.6 billion acquisition.**”*

Where the acquisition event itself is considered as background information, in a longer news article, the price may be missing. The event in such cases usually appears in the middle section of the article, rather than in the beginning (see Section 3.4.1). The following is an example of Corporate Acquisition (Associate Press):

- “*A spokesman for Lehman Brothers said the award will be paid from a litigation reserve established when **Smith Barney, a unit of Travelers Inc., acquired Shearson last year from American Express.** Lehman is now an independent company. American Express spun off its Lehman stock to shareholders earlier this year.”*

In general, while the foreground facts are scattered, the background information is frequently presented in a more compact form, since the facts are stated in shortened form for quick explanatory purposes.

Nature domain structure

In the Natural Disasters and Disease Outbreaks, I observed more intra-event scattering than in Business news. In Natural Disasters the amounts for damages, counts for the dead and the injured, and descriptions of the victims may appear relatively far from the first mention of the main disaster. Consider the following fragment of a news report of more than 600 words (names of disasters are underlined, damages are in boldface):

- “*SEA BRIGHT, N.J. – A brutal northeaster thrashed the Eastern Seaboard again Thursday with cold, slicing rain and strong winds that caused flooding in coastal areas of New Jersey and Long Island. The storm also prompted fears of widespread sand erosion as violent, churning waves – some up to 20 feet high – pounded beaches throughout the region. But even as the pelting rain and gusting wind made for nasty conditions across the New York metropolitan region, damage seemed to be limited to*

*flooding, scattered power failures and temporary road closings in coastal areas. **No injuries or deaths** were reported. Elsewhere along the East Coast, **19 deaths** have been attributed to the storm since it began on Monday, bringing tornadoes to southern Florida before churning to the north and hammering the coast from Georgia to New Jersey with heavy rain and winds as high as 75 miles per hour.*

*The **19 deaths** include **five** in accidents on snowy roads in Kentucky and **two** in Indiana. **Two men** died when the roof caved in under 11 inches of snow at a recycling plant in Princeton, W.Va. And in South Carolina, **a pregnant woman** drowned when her car plunged into a swollen creek.[...]*

In view of the editorial organization, an article may continue to list more and more remote and lesser damages and conditions caused by the storm, so the editor is able to cut the article at any point, depending on the space available. The Sea Bright report later on returns to describe the main disaster in more detail, the damages and their exact causes, more precise routes and locations of the disaster, and actions taken by the authorities. While shifting focus on various locations, the current disaster is compared with earlier ones, as to their effects and intensity, backed up by statements from the authorities, and descriptions of actions, such as evacuations.

In the Disease Outbreak reports the sought facts are as frequent as in the Natural Disaster reporting, and intra-event scattering is heavy in both. The following is an update report from ProMED. The locations and victims are highlighted in the text to show the scattering:

- *“KAMPALA: The Ugandan Health Ministry announced here on Tue 12 Dec 2000 that another **7 Ebola hemorrhagic fever cases** have been reported in Uganda since last Friday [8 Dec 2000], bringing **the total number of cases to 413**. ‘**A total of 5 cases** in the northern Ugandan district of Gulu and **2** in Masindi in the west have been admitted to hospitals [during this] period,’ Assistant Commissioner for National Disease Control, Alex Opio, told a news briefing here at the Ministry’s headquarters. ‘Things are getting better with fewer cases reported in the past 5 days, and it is really very encouraging.’*
Opia said that the Ministry is still intensifying control measures and supplying necessary protective materials to the affected districts, adding that social mobilization is going on. Control efforts will not end until 42 days after the

last patient is discharged from hospital.

*Meanwhile, the Health Ministry has announced that the Mbarara district in the southwest is now Ebola-free with no new cases reported in the past 6 weeks. **A total of 5 cases** had been confirmed with **4 deaths** in Mbarara district in October.[...]*

In this example, the cases in different locations cause event scattering. The number of new victims is seven (first sentence), which is further analyzed by location in the second sentence. Mapping the places and numbers into one incident is challenging. The expressions are less metaphorical than in Natural Disaster news, but the scattering of the components of facts is very similar.

This 272-word article further describes the actions taken by the authorities (*intensifying control measures...*), and the statements concerning other affected locations. More numbers of victims are mentioned at other locations and on other dates, and the total numbers for the entire outbreak.

The definition of a fact is complicated in Natural Disaster scenario because the boundaries of disasters are harder to define. It is less clear how far one disaster can extend geographically, how many metamorphoses it can undergo, and what kind of damages it can cause. Further, the events interlocked and related to each other in complex ways, raising important questions about how best to organize the extracted events, and how to define the structure and the extent of the events. To present hierarchically arranged information, a new solution is explored in Huttunen et al. (2002).

As a consequence of these structural factors, the Nature scenarios presented problems with the definition of facts that did not arise in the Business scenarios. This included, in particular, delimiting the scope of a single event, and organizing the events into templates.

Filling the IE template

In Natural Disasters and Disease Outbreaks the organization of the components of facts and the relations between them is more complex than in the Business scenarios. This problem of intra- and inter-event *scattering* was introduced in Sections 2.4.2 and 2.4.2. Because of the scattering phenomena, it is difficult to organize and define extracted events in the Nature domain using the traditional template structure (which suits the Business domain). Due to intra-event scattering, a new, modular template structure is proposed and inclusion relationships between modules are defined. To be able to extract events correctly in the Nature

Table 2.17: Template: Disease Incident

<i>Slot Name</i>	<i>possible fillers</i>
Disease Name Date Location	
Case Descriptor Case Number Case Status Victim Type	NP (e.g., people, cows, ...) { infected, sick, dead } { human, animal, plant }
Hierarchical Event Link	

domain requires the division of the event into separate *modules*. One set of patterns is built to capture the description of the disease or the disaster. A separate set of patterns is built to capture the material damages and the casualties, etc. These modules are then merged, but with a restriction: if deaths and damages are found an, event is created only if fills for disaster and location have also been found somewhere nearby in the same article. In this way deaths and damages caused by, e.g., a war, sporting accidents, or car accidents are not included. The modules (the pieces of information relating to the same event) may become attached to the wrong events.

In the Nature domain, scattering also occurs on the inter-event level. For example, in Infectious Disease Outbreaks, grouping separate events into outbreaks based on whether the name of the disease and the location are compatible in the events, and dates associated to each sub-event are near each other (e.g., within a month). When those requirements are fulfilled, the events are assumed to form an *outbreak*, a super-event, comprised of the component sub-events. Publication II, (1.5.2, “Diversity of Scenarios”) describes the template structure of these scenarios further.

The template for Infectious Disease Outbreaks is in Table 2.17. The template has slots for the disease name, date and location. Four slots describe the victim(s) of the disease. The *case descriptor* slot is filled with a noun phrase that describes the victim, as found in the article. The number of the victims fills the *case number* slot. The *case status* and *victim type* are set to one of the three possible values for these slots in Table 2.17, based on expressions found in the text. A separate

slot indicates the parent incident, the *hierarchical event link*, see the following subsection.

Each extraction pattern operates only on one sentence at a time. Due to intra-event scattering, we may fill several such templates, and not all the slots in each template are necessarily filled. Also, typically there are several templates per document. For example, in one sentence the extraction system may find only the disease name, the location and the number of diagnosed cases for one event, and only those three slots are then filled. Another template may be filled only with number and case status. The rest of the slots are allowed to be empty.

Inclusion relation among events

The templates often have overlapping information, if they belong to the same event, i.e., the same outbreak, and one or more templates may be partially included in one another.

To handle the scattering problem in the Nature scenarios, I make a distinction between *outbreaks* and *incidents*. An outbreak consists of several reported incidents. An incident is a mention of a specific occurrence relating to some area or region of space and some span of time, that is part of a larger outbreak. An outbreak is a continuous chain of incidents, with no time gaps in the chain longer than some window (we use a window of one month). In this way it was possible to define the template as consisting of only one incident. The incident templates can be grouped into outbreaks by *hierarchical links*, i.e., event pointers, based on time and location they occur. The result is a simple template, but many of them in one document, and they are interlinked.

Table 2.18 shows an extract of a news about meningococcal disease spreading in Sudan. Figure 2.1 is a graph of the inclusion relationships among incidents extracted from this Disease report. The figure shows the main incident with several sub-incidents. Two of the sub-incidents have further sub-incidents. The different types of inclusions for each scenario are explained in the next section.

Figure 2.2 shows a graphical representation of inclusion by causation in Natural Disaster scenario. The incidents are extracted from Table 2.19.¹⁴

There is a causation relationship between the incidents. It is important to recover the long causation chains from the text.

¹⁴Note that the northeaster is not in causation relationship with storm, which began on Monday. The damages that the synonymous northeaster caused, are from the following Thursday.

Table 2.18: Example of a Infectious Disease Outbreak Report

- (0) Meningococcal in *Sudan*
- (1) A total of **2 549 cases** of meningococcal disease, of which **186** were fatal, was reported to the national health authorities between 1 January and 31 March 2000.
- (2) *Bahar aj-Jabal State* has been most affected to date, with **1 437 cases** (including **99 deaths**) reported in the *Juba city area*.
- (3) Other States affected include *White Nile* (**197 cases, 15 deaths**), [...]

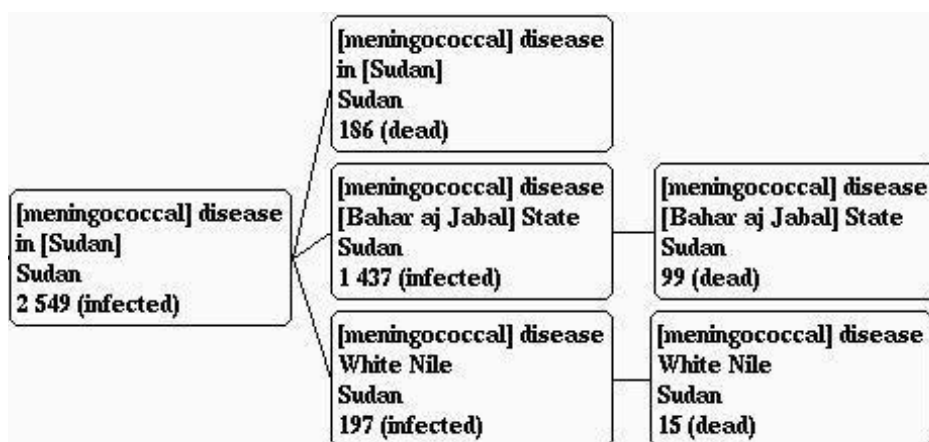


Figure 2.1: Inclusion relationship between incidents in Infectious Disease Outbreaks

Note that the northeaster is not in causation relationship with storm, which began on Monday. The damages that the synonymous northeaster caused, are from the following Thursday.

Since the lexical and structural differences were considerable between the scenario-specific expressions of facts, new methods were employed to catch those events. Text analysis methods were used to analyze the location of the events in the text in different text types, and the cohesive devices were explored to see how the different components of an event relate to each other. The distance between the facts (Bagga and Biermann, 1997) varied along the scenarios.

Table 2.19: Example of a Natural Disaster Report

- (1) A brutal northeaster thrashed the Eastern Seaboard again Thursday with cold, slicing rain and strong winds that caused flooding in coastal areas of New Jersey and Long Island. [...]
- (2) Elsewhere along the East Coast, 19 deaths have been attributed to the storm since it began on Monday.
- (3) The 19 deaths include five in accidents on snowy roads in Kentucky and two in Indiana. [...]

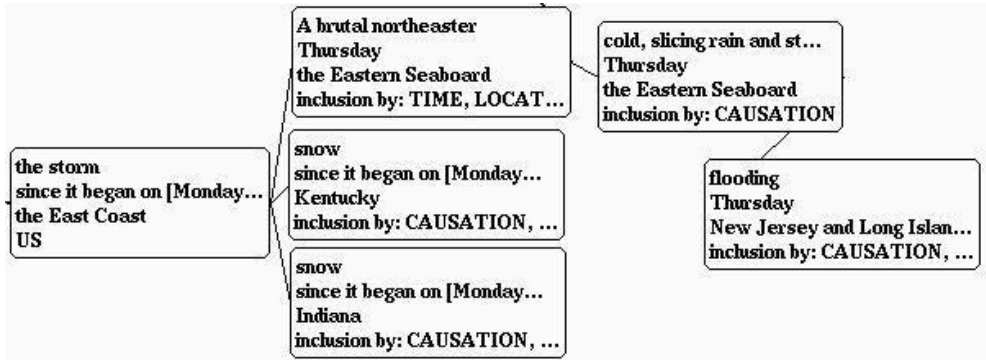


Figure 2.2: Inclusion relationship between incidents in Natural Disasters

The nature and direction of *the inclusion and causality* relationships between facts was analyzed using cue words and scenario-specific ontologies, which are built on the basis of the lexical analysis.

Cues that signal the type of inclusion

Next, I define cues that point the relationship between templates, to try to capture both intra- and inter-event scattering. The cues are linguistic expressions that give the reader hints of how the pieces of information are linked to each other. This is presented in Publication II and Publication IV.

To study scattering, I analyzed the texts with the IE system and defined whether the events were separate or a part of another event, and how far apart the parts were. To investigate the extent of inclusions and their distribution by type,

I analyzed 40 documents related to the Nature domain,¹⁵ and manually tagged the inclusion relationships present in these documents to confirm the feasibility and applicability of this approach.

On the discourse-semantic level, following the classification of levels in systemic-functional linguistics, as explained in Section 1.3.2, we try to identify those cohesive devices which carry information about how the different components of an event relate to each other. Based on extensive work with our corpus, we observe a strong tendency of certain scenarios to invoke a specific style of expression of facts, even if the source of the text is the same for all scenarios. The style varies according to the scenario on a continuum: Nominations (in the Business domain) having highly formulaic expressions at one end, and highly descriptive narrative of Natural Disasters at the other end. To link the scattered events, it is important to take into account the organization of the parts of the event in the text, lexical cohesive devices, such as repetition, elliptic constructions, and cues.

I analysed the characteristics and direction of *the inclusion and causality relations* between incidents using cues and scenario-specific ontologies, constructed on the basis of the lexical analysis (Publication IV).

The following are examples from short news reports about infectious disease outbreaks. The cue is underlined. (For additional examples, see Appendix A)

1. “A total of 8 Guatemalans have died from the effects of dengue virus infection, 6 of whom were children [resident in] the sweltering and [untidy] town [Coatepeque] of 200 000 people located 130 miles west of Guatemala City.”
2. “The number of people sickened by *E. coli* bacteria linked to a Green Bay meatpacker has increased to 35, including a 6-year-old Houston County girl who has developed a life-threatening complication.”

The first victim count (8 Guatemalans) in the first example includes the second victim count (6 children), and is thus inclusion by victim count.¹⁶ There are also multiple inclusions.

¹⁵The training corpus was used to evaluate the performance of our IE system on these tasks. For the Disaster scenario we analyzed a total of 14 reports from NYT, ABC, APW, CNN, VOA and WSJ. For Infectious Disease Outbreaks, a total of 26 documents from NYT, ProMED, WHO, and ABC.

¹⁶The hypernym-hyponym relation is defined in the ontology. The victim count in the sub-events, allows us to determine the direction of inclusion. Without the number, the inclusion relation between NATIONALITY (that is, Guatemalans) and CHILDREN could hold in either direction.

In the second example we see inclusion by victim count, and also the disease name (*E.coli bacteria*) includes the *life-threatening complication*.

Table 2.20 shows the number of incidents found in the documents, as well as the number and the types of inclusion. The following is a summary of the inclusion relationships found in the two Nature scenarios.

- location: e.g., victim count in one city contributes to the victim count in the whole country.
- time: e.g., victim count for an update report contributes to the overall victim count since the beginning of the outbreak.
- status: dead or sick count is included in the infected count.
- victim type or descriptor: e.g., “people” includes “health workers,” and “children.”
- disease name (Disease scenario): e.g., the number of Hepatitis C cases may be included in the number of Hepatitis cases.
- disaster (Disaster scenario): e.g., damages caused by *rain* may be included in the damages caused by *rain and winds*.
- causation (Disaster scenario): a disaster can trigger derivative disasters.

We manually tagged the inclusion relationships present in these documents. Typically each document has one high-level event, with no parent events. Some documents refer to other events (e.g., earlier in time), e.g., to compare a current event with a precedent. These get a separate template and do not cause an inclusion. Table 2.20 shows the number of incidents found in the documents, as well as the number and the types of inclusions. There are also multiple inclusions, marked with “+” in the table: e.g., infected *health workers* in a town in Uganda are included in the total number of infected *people* in the whole country: this inclusion by *case-descriptor* and *location*.

Multiple inheritance also occurs: in table 2.18, the deaths in Bahar aj Jabal State contribute to the infected count in that state, as well as to the total number of deaths in Sudan. However, in table 2.20, we show only the inclusion in the immediately preceding parent.

The *distance* and the nature of *the relationship* between the incidents is important to the scattering phenomena. The organization of discourse has been studied from the point of view of IE in Bagga and Biermann (1997), where the focus is on

Table 2.20: Type and Number of Inclusions

	<i>Disaster Scenario</i>	<i>Disease Scenario</i>
Documents	14	26
Words	6500	9500
Incidents	112	125
Inclusions	81	57
Inclusions by		
time	6	6
location	20	19
status	1	19
case-descriptor	1	6
case-descriptor+location	–	3
disease	–	1
causation	19	–
causation+location	11	–
causation+time	3	–
time+location	7	–
disaster	5	–
disaster+location	2	–
damage	4	–
others	2	3

counting the distances between the sub-events or parts of the event. Rhetorical structure theory (RST) (Mann et al., 1989; Taboada and Mann, 2006b,a) provided an elaborate analysis of the relationships between text segments or parts of discourse situated close to each other.

Table 2.18 shows an example of news, reporting a Sudanese outbreak, and Table 2.21 demonstrate the scattering problem and how to recover the relationships between the sub-events.

In our IE system, events belong to separate outbreaks when the date and/or geographical location are sufficiently far apart in the respective events extracted by the system. Usually the date and location are found quite close to the event in the text. If the dates are clearly months/years apart from each other, the events are separate. But if there is no locative/temporal indication of the event, then it is trickier. A pattern-based IE-system automatically creates separate entries for

Table 2.21: Incidents from Infectious Disease Outbreak Report

<i>Disease</i>	<i>Location</i>	<i>Infected</i>	<i>Dead</i>
Meningococcal	Sudan	2549	186
	Bahar aj-Jabal State	1437	99
	White Nile	197	15

the events, and the events are treated as separate from each other. This causes problems if the database does not provide any links between those events, since for example the number of victims and damages caused by a natural disaster or disease outbreak are then multiplied. There is a need to point out the existing relationship between those events.

2.4.3 General and scenario-specific pattern types

In this section I list typical patterns built for each scenario, the number of the patterns and the F-measure for a current set-up. Collocational differences between two types of text in the same business scenario vs. general news are shown to illuminate the point. There is no separate publication about the patterns but as it is the basis for this study it is presented here.

Natural Disasters

For the Natural Disaster scenario, a typical pattern is

- DISASTER – [VERB] – LOCATION.

The DISASTER slot in this collocation is filled either with the noun *disaster* or one of its synonyms, or a type of disaster or a named disaster (e.g., *flood*, *storm*, *El Niño*, *hurricane Katrina*). LOCATION is a named location, either specific or vague. VERB is a verb describing the specific type of impact that the disaster may have on the location, or some other possible relationship between the two. The VERB is a type of action in the appropriate concept hierarchy describing how disasters affect locations. Several synonyms (in this context) in the concept hierarchy may be used: *hit*, *enter*, *reach* etc., and disaster-specific verbs such as *rock* with earthquakes and *denude* for wildfires as in

1. *Earthquake rocks Mexico's Pacific coast.*

2. *Wildfires denude Colorado foothills.*

Each scenario has its own ontology where the relationship between the terms are defined and hierarchically arranged. For example, *disaster* is a hypernym or superordinate for *flood*, *storm*. If a slot may be filled with a word *disaster* or any other subordinate term, it is marked with “DISASTER,below” in a pattern so the type-of subdisasters are applicable as well.

For the pattern of type

- DISASTER – [VERB] – HUMAN

the slot for HUMAN is filled with anything that refers to a human being. The synonyms for the verb in VERB are numerous, e.g., *affect*, *kill* since disasters may evoke so many more consequences on a human being than a company on another company. For disasters, the division between the verbs into final categories was based on the consequences on humans: whether the person was hurt, killed or affected. The affected option was chosen if it was not clear whether the victims were hurt or killed. Also the economical consequences were extracted separately: how many houses/roads/powerlines or other property the disaster affected/broke.

- *Drought in Thailand’s northern province Nan has caused more than \$1.2 million damage and left more than 200,000 without sufficient water, authorities said.*

The F-measure for Natural Disasters was 44.23 (precision 36, recall 58) for the test corpus. For the training corpus the F-measure was 72.69 (precision 71, recall 75). The F-measures and number of patterns for all scenarios¹⁷ are listed in Table 2.22:

In the disasters and medical scenarios we evaluated all the event and sub-event level templates. We did not evaluate the inclusion relations because it was unclear how to do that at the time.

This table suggests several observations. The longest documents are in Nature scenarios with the Natural disasters being the far longest. The shortest documents, by far, are the Business documents. The amount of work put in one scenarios can be estimated by the number of patterns. With that, Nominations scenario with one of the smallest pattern sets give the highest scores. Disaster scenario with

¹⁷ Pattern information for Nominations scenario is not available, because it was customized before my work. All business scenarios the same sub-patterns, for which reason I don’t include them separately in the table.

Table 2.22: F-measures and numbers of patterns for all scenarios. Document length is given in the number of words.

<i>Domain</i>	<i>F-measure,</i> <i>(prec/recall)</i>		# patterns (subpatterns)		Avg doc length (shortest/longest)	
Disasters	44.23	(36/58)	58	(16)	781.5	(38/3113)
Diseases	71.25	(65/79)	133	(28)	249.2	(57/839)
Security	48.83	(47/41)	279	(29)	268.6	(42/764)
Business/all scenarios	58.93	(61/57)	327	(35)	—	—
Nomination (MUC-6)	77.36	(72/84)	34	—	—	—
Investments	—	—	50	—	92.8	(30/346)
Product Launches	—	—	16	—	73.0	(37/121)

considerably more patterns, have very low scores despite having almost twice as many patterns. Security has a slightly higher performance, but that came with the expense of many more patterns.

Business domain

A typical pattern in Business domain is

- COMPANY – [VERB] – COMPANY

The pattern reflects interactions between companies. A common type of interaction is buying or selling. The most common VERB is BUY or SELL, with very few synonyms. SELL contains only the verb *sell*. BUY contains three verbs, *acquire*, *sell* and *purchase*.

For the PRODUCT-LAUNCH scenario in Business domain one of the most frequent patterns is

- COMPANY – LAUNCH – ITEM

where ITEM may be anything from the noun *product* to any product name, or line of products that is usually caught by the named-entity pattern of the system.

The number of all the clause-level patterns for all Business domain scenarios together is 327 (with 395 subpatterns).

The F-measure for all the Business domain scenarios (without Nominations scenario) is 58.93, with 61 precision and 57 recall.

For comparison, for the concepts in the Nominations scenario the class of synonyms is restricted: for COMPANY – VERB – PERSON where VERB is HIRE/FIRE, the only synonym (in this context) for *hire* is *appoint* (and *elect*, but only in passive voice); synonyms for *fire* are *dismiss*, *oust* and, *remove*. For the pattern PERSON – VERB, (where VERB describes the move that a post-holder him/herself initiated), there are more synonyms (and paraphrases) due to the variety of reasons that the act has taken place: the most general synonyms are *depart* and *quit*, followed by the more restricted *resign*, *step-down*, *retire* and *die*.

The F-measure for Nominations is 77.36 (precision 72/recall 84).

The lack of synonyms for the Business domain (See Table 2.26) is due to the strictly fact-oriented style of the reporting and the scarcity of expressions with which e.g., purchases between companies are announced in the news genre. In comparison, there is much more variation in the descriptions when it comes to a person who has contracted an infectious disease, or even how a disease appears in a location, especially in the general news as opposed to more scientific genres such as medical reports written by medical experts. The general news needs to attract readers' attention.

There are also lexical differences between the Business Abstracts targeted to business specialists, and the Business Articles targeted to general public. For example, in the Business Abstracts corpus the VERB *launch* appears frequently, and always related to company actions and expressing a relevant fact, as seen in Table 2.23.

In general news, however, *launch* is common but appears in wide range of collocations (see Section 1.3.2 L6), such as in Table 2.24

In the Business Articles, the VERB *launch* appears rarely and even then not necessarily in the company action. In the corpus the VERB *launch* appeared only once, as seen in Table 2.25.

In the Product Launches scenario, the verbs forming the verb class LAUNCH are in collocation with the *launchable* class of object PRODUCT, especially in the Business Abstracts corpus. In Appendix B, Table B.1 shows all possible objects that are launched in Business Articles, general news, and Business Abstracts. The objects that do not qualify as products to be launched are highlighted. The greatest number of non-qualifying objects was found in general news.

This suggests that in the more formulaic and restricted articles the patterns could be more loosely defined, e.g., it might suffice to have a pattern COMPANY-

Table 2.23: Example of *launch* in Business Abstracts

<i>Actor</i>	<i>Verb</i>	<i>Patient</i>
Ryanair	launch	service/flight/route
Blue/it	launch	flights
Air-Easy/Jet-airline	launch	flight
it/company	launch	service
company	launch	product
it/company/firm	launch	range
it/company	launch	products
it/company	launch	products/production
company	launch	tender
North	launch	artillery barrage
European comission	launch	investigation
TAP	launch	route
European comission	launch	public consultation
company	launch	line
Apple	launch	version

LAUNCH-X instead of defining the object X further (or, defining the object but not the verb).

Another difference between two kinds of business data is that only in the Agentum data was there the problem that the pronouns ‘we’ and ‘they’ refer to COMPANIES, which are inanimate non-human objects. In Business abstracts this problem did not appear.

Security domain

For the Security scenarios such as Smuggling, Human Trafficking or Illegal Migration a basic pattern is of type

- PERSON – SMUGGLE – ILLEGAL-SUBSTANCE

In addition to the large and growing number of new drug-names, the complication with Security domain arises from the wide range of potentially smugglable items, and names referring to them, e.g., animals or parts of animals, alcohol

Table 2.24: Example of *launch* in General News

<i>Actor</i>	<i>Verb</i>	<i>Patient</i>
it	launch	uprising
Obama	launch	WhiteHouse run
Romney	launch	unsuccessful 2008 White House bid
Bahrain	launch	sweeping crackdown
candidates	launch	final stretch run
Scotland Yard	launch	murder investigation
NCAA	launch	own inquiry
British Liver Trust	launch	national awareness campaign
North	launch	artillery barrage
British police	launch	investigation
Peters/Ms. Peters	launch	website
QPR	launch	Barton appeal
...	launch	...investigation/inquiry/crackdown/ comeback bid/military manouvre/ campaign/Obama/hungerstrike/ fully revamped version/ rebellion, deadly attacks...

with many referring terms such as *booze* and *liquor*, and ARCHAEOLOGICAL-ARTEFACT, arms, chemicals, cars, human organs etc.

For the Security scenario, the F-measure on the test corpus has been around 43.83, with 47 precision and 41 recall.

One aspect that helps in locating events for Security from news reports is the change of perspective. A person smuggling something illegal from one country to another, either directly or through a third country, is rarely reported in news as an isolated event (unless somehow speculative or imaginary as in novels or special news analysis), but usually in relation to an authority act that already has taken place. For example the following news report includes an action by the police (underlined).

- *An alleged Malaysian drug dealing ring has been broken up by the Indonesian National Police drug division, which arrested seven suspects in Medan, North Sumatra, and seized 5 kilograms of crystal methamphetamine on Monday.*

Table 2.25: Example of *launch* in Business Articles

<i>Actor</i>	<i>Verb</i>	<i>Patient</i>
government	launch	“comprehensive” review

So some patterns that otherwise would be too general, were attached to an authority-statement in the same sentence, which helped with the precision. For example, a pattern

- AUTHORITY – ACCUSE – PERSON

where verb class ACCUSE includes verbs such as *prosecute*, *charge* and *arrest*. The authority-statement, AUTHORITY, may be a government organization such as *Immigration and Customs Enforcement agent* or *Bosnian police, patrol boat, federal grand jury* etc, or the noun *authority*.¹⁸ This pattern will then fill the slot for a suspect. If the suspect slot is filled, we may then fill in the found crime. Many of the patterns have the authority action and the crime in the same pattern. If not, the rest of the slots are filled with entities that additional sub-patterns have found in later/previous sentences.

A variety of person names may resolve with suspects when they are not actually suspects at all. Several sub-patterns capture non-suspects, such as

- PERSON – TITLE-AUTHORITY

where the person name is attached to a title of a certain authority.

Many news reports are reporting the suspects penalized in a passive voice, without an exact mention of the authority. A pattern of type

- PERSON – pass-vg CHARGE – VING SECURITY – NP

takes care of a lot of cases where a named person (e.g., *Nicolaides, Thanksin*) or a human being (*five men, 41-year-old man*) are charged with any action listed in the Security scenario, with any noun that follows. Or, if the verb is not listed as a Security domain but the noun is, then this is also a valid Security event. Depending on the verb/noun, the Security event is listed as either Human Trafficking, Smuggling or Illegal Migration.

¹⁸There are actually two versions of this pattern, because AUTHORITY covers a set of authorities, the GOVERNMENT-ORG covers a set of pre-listed government organizations.

Infectious Disease Outbreaks

For the infectious disease epidemics an exhaustive list of disease names is crucial for the IE to be successful. New disease names are discovered with patterns that require e.g., a specific verb for getting a disease but leave the slot for an object open, for example,

- PERSON – INFECTED – ‘with’ ANYTHING

where a PERSON is anything referring to a human being, INFECTED includes passive voice of verbs such as *infect*, *affect*, *diagnose*, *sicken*, *hospitalize*. ANYTHING is filling up a disease name -slot. A large number of the non-contagious diseases are listed and thus prevented from creating a medical event, but a new disease name referring to a non-infectious disease may fill the slot and create a false event. Another type of common event is

- CASE – DISEASE

in which CASE includes the noun ‘case’, which is very typical, and the name of a disease. The verb group REPORT in the beginning of the pattern or the passive voice of the verb class CONFIRM in the end of the pattern is limiting the amount of false positives. The pattern [CASE – DISEASE] pattern finds the following events:

- *Egypt on Sunday reported two more death cases of A/H1N1 flu*
- *The minister revealed to Radio Dabanga that three cases of yellow fever have been confirmed in the locality*

For the Natural Disasters, the role of time and place is very crucial. Several sub-patterns capture the time and place of the incident, and a complex combination pattern for various expressions referring to time and place is inserted inside the clause patterns by rules.

The F-measure for Infectious Disease Outbreaks scenario is 71.25, with precision 65 and recall 79.

Examples of the patterns in all scenarios are shown in Table 2.26 on the facing page and in Table 2.22 the number of patterns and the evaluation numbers for each domain.

To be able to include in the IE results the items that were outside the trigger sentence, I created *inference rules*. The inference rules define the category of

has to be part of the class as C-MEANS (means of transportation of smuggle-item, for example the trunk of a car or an airplane) and C-DRUG (if the candidate concept is a name of a drug such as marijuana or ecstasy) in the SMUGGLE-scenario, in case there is a trigger sentence representing a smuggling between two countries. In some cases other conditional states are required, e.g., the smuggling-event is conditioned to the appearance of at least two country names and an item belonging to the class C-MEANS, indicating transportation such as a van, suitcase, or airport. Then the smuggled item slot may be filled with a variety of entities. That is a very domain-specific class of events.

Chapter 3

Assessing relevance of IE results for the user

In the second part of my thesis the IE results are analysed from the user's point of view. This is reported in Publication VII.

To measure utility, we combined methods from text mining and linguistic analysis to identify features that predict the relevance of an event or a document to a given user. This way it is possible to offer the user not only correct but also highly relevant information.

We observe scenario-specific and text type-related differences. The focus is on general and scenario-specific features that have predictive power about the relevance of the extracted event. We gathered a set of features and applied statistical methods used to investigate whether there is a set of features that correlates with the event relevance, in different domains or scenarios. We study how the individual discourse features predict relevance by themselves, and look for differences between the domains.

The chapter presents the methods, data, tools and the results for the relevance study.

3.1 User-centric relevance of IE results

The news extraction and relevance prediction works in three phases: first, the potentially relevant articles for the target domain are identified using a broad Web search, based on queries that are Boolean combinations of keywords. The second phase employs IE to extract events from the acquired articles. The third phase determines the relevance of the extracted events to the end-user.

Publications V and VII describe this process. The information caught by the combined efforts of IR and the PULS IE-system covers a wide range of relevant events varying from moderately to extremely important to the user, including also events that are irrelevant, even if correctly extracted. A substantial amount of irrelevant, even if correct, events were extracted, especially in the Medical scenario. Two ways have been used in this study to evaluate the IE results for a given scenario: first, the traditional F-measure, which is in relation to the text, and measures the objective correctness of the sought events. But the correctly extracted event is not necessarily relevant.

The second way to evaluate results, developed here, is user-oriented in the sense that it gives each event a measure according to its relevance to the user along a 5-step scale described in Table 3.1 to define the relative relevance of the event for Infectious disease epidemics. These scores are reduced for simplicity, into

Table 3.1: Guidelines for relevance scores

Criteria	Score
New information; highly relevant	5
Important updates; on-going developments	4
Review of current events; hypothetical, predictions	3
Historical/non-current, background information	2
Non-specific, <i>non-factive</i> events; secondary topics	1
Unrelated to target domain; useless	0

either a three-way classification—*high* (4–5), *low* (1–3) and *irrelevant* (0) events,—or a binary classification—where events with a score of 4–5 are considered high relevance, and those with a score of 0–3 are considered low-relevance.

To find out what the user needs are, we built an environment that allows users to interact with the system by rating the analyzed content of an IE system, as seen in Figure 3.1. Each event is converted to a feature vector, to which a classifier (see Section 3.3) assigns a relevance score. The event’s relevance score appears on the on-line server in an environment. The end-users were allowed to review and rate events extracted from the news using the user interface of the PULS news surveillance system, indicating how relevant the extracted information was for them. We present machine learning methods to predict the relevance of events to users. The score is shown in the on-line server with the event, and the user

Lang	Published	Source	Disease	Country	Date	Total	↑	Descriptor	Note	Reviewed	Rel
en	2010-04-29	newsmedical	Avian Influenza	India	2010-04-29	7		seven new cases		ecdc	2
en	2010-04-29	hpa	Measles	UK	2009	39		39 confirmed cases		ecdc	3
en	2010-04-29	hpa	Norovirus	UK	2010-04-28	--		--	ship	ecdc	4
en	2010-04-29	hpa	Meningococcal Meningitis	UK	2010-04-26	1		one suspected case	official info	ecdc	5
en	2010-04-29	promed	Measles	Spain	--	65		16 cases	linked to the bulgarian outbreak (Roma pop)	ecdc	5
en	2010-04-29	promed	Measles	Spain	2010	65		65 cases		ecdc	5
en	2010-04-29	crnn	--	Germany	2009-05-04	--		her husband			
en	2010-04-29	reliefWeb_latestUpdates	Shigella	Haiti	2010-04-21	100		100 children			
en	2010-04-29	reliefWeb_latestUpdates	Cholera	Haiti	2010-04-21	100		100 severe cases			
	2010-04-29	africaFM	Malaria	Africa	2010-04	--		--		ecdc	2
en	2010-04-29	YorkshirePost	--	UK	2010-04-28	--		Mrs Bradshaw		ecdc	0
en	2010-04-29	GulfDailyNews	--	--	2004	16	↑	16		puls	0
en	2010-04-29	AsiaOne	Dengue Hemorrhagic Fever	Brunei	--	2		two cases		ecdc	2
en	2010-04-29	AsiaOne	Dengue	Brunei	2010	77		only 53 cases		ecdc	3
	2010-04-29	ChinaPost	Typhus	Germany	1942-06-12-2010-04-29	--	↑	Anne Frank		puls	0
	2010-04-29	AP	H1N1	Puerto Rico	2009-12	--		whose infant daughter		ecdc	2
	2010-04-29	chosun	--	South Korea	2008	more		more patients	Health tourism	ecdc	2
en	2010-04-29	sciencedaily	Influenza	--	2005-2008	13%		the women	pregnance/drugs (study)	ecdc	3
en	2010-04-29	sciencedaily	Tuberculosis	Peru	--	1	↑	a contemporary leader		ecdc	0
en	2010-04-29	sciencedaily	Unknown Disease	South America	2010-04-23	--	↑	Simon Bolivar		puls	0

Figure 3.1: Infectious disease outbreaks: table view with marked relevance on the rightmost column.

may check whether the relevance score is correct, and whether the attributes of the event are extracted correctly. In case of errors in the automatic extraction, the UI allows the user to correct erroneous fills, e.g., if a company name, country, or a disease name was extracted incorrectly.

The screenshot in Figure 3.1 shows relevance predictions highlighted with colours, which allows easy notification of high relevance events.

The users' feedback with corrected relevance was used as training data for building relevance classifiers to better serve the end-users, see Section 3.2. The relevance estimations were based on linguistic and meta-linguistic features of the texts. The features are engineered through detailed and multifaceted linguistic analysis. Identifying relevant pieces of information in a scenario is a combination of qualitative (described in Section 2.4) and quantitative methods (see Section 2.4.2). Quantitative analysis uses general features of the extracted events and statistical methods, applied for rating the importance of the events to the user, in the Business, Medical, and Security domains.

We used statistical packages (see Section 3.3) to measure how well lexical and discourse features, described in Section 3.4, are able to predict the relevance of an event for the user.¹

For rating the relevance, we defined the characteristics of the relevant vs. non-relevant events listed in Publication V. The features are based on the studies of cohesion and coherence, on interpretation from the data. We also must take into account the technical limitations of the NLP tools, for example, pragmatic features may be too difficult to implement. In the statistical experiments for each scenario, a graph demonstrating correlation between the rated relevance and every feature was built. These experiments are described in Publication VII, Publication V, Huttunen et al. (2013), and Vihavainen (2011). The outcomes of the experiments for the scenarios were compared, in relation to the effect of groups of features. The relevance studies were applied in the following domains:

- Infectious Disease Outbreaks
- Business Domain: Corporate Acquisitions and Product Launches
- Cross-Border Security: Human Trafficking, Smuggling (arms, drugs, goods, CBRN,² etc.) and Illegal Migration

I compared the scenarios to each other in how they organize the high-relevance vs. low-relevance or background information. Statistical methods were used to compare scenarios and the appearance of events rated by relevance.

The features were tested first in the Medical scenario, because for this scenario we had the greatest amount of user interaction and feedback, the best recorded user experience.³ Then the obtained results on Medical scenario were compared with Business scenario. The experiment with Medical and Business news is described in detail in Publications V and VII. Eventually, Medical, Business and Security domains were compared in relation to the features to see whether there is variation in how the relevant information is expressed. Security domain is described in Publication VI.

¹The PULS IE-system also computes confidence using discourse-level cues, such as: confidence decreases as the distance between the sentence containing the event and event attributes increases; confidence increases if a document mentions only one country.

²Chemical, biological, radiological and nuclear materials.

³The tests are performed only if the corpus includes articles from all categories, relevant, non-relevant and moderately relevant. In the earlier phase of the study, a problem arose with the Business scenario as all the articles from the feed were relevant. This was fixed by extending the news feed also to non-relevant articles.

3.2 Data

For the relevance study, the data consists of hand-labeled data from the PULS user interface (UI). The UI also allows the users to correct **erroneous fills**, e.g., if a company name, a country or a disease name is extracted incorrectly by the automatic extraction, the IE error can be corrected by the user. The manually checked/corrected data are used for training and testing the relevance classifiers. One document never contributes events to both training and test sets, to avoid training and testing on the same data. Since some of the labeled events are also corrected by the end-users (in case of erroneous extractions), we have two parallel sets of events with relevance labels. The “raw” events, with fills extracted by the system, and the “cleaned” events, i.e., the same events with corrected fills. The “raw” set is more noisy, since it contains errors introduced by the IE system. The relevance classifiers were trained on the cleaned labeled data. For evaluation, the classifier performance was tested against both the cleaned and the raw events, although the focus was on classification performance on the *raw* events: the goal is to build a classifier that is applicable to the real extracted event stream. The “raw” scores in evaluation give us an indication of what performance can be expected in a real-world setting. In the Business domain, we used 213 user-labeled events, in 127 documents. The division of the hand-labeled data is 45% high-relevance and 55% low-relevance documents. For evaluation in the Business domain we use 220 event level events and 160 document-level events. For the Medical scenario we use about 1000 evaluated events and 600 document-level events. About 80% of examples are labeled with lower relevance.

3.3 Setup and tools for relevance

The statistical experiments were done with R and the Java-based Weka toolkit (M.Hall et al., 2009), which provides a suite of machine learning algorithms, such as Naive Bayes and SVM (Support Vector Machine), see Vihavainen (2011). Evaluations are done using ten-fold cross-validation. The results were evaluated using precision, recall, F-measure, and accuracy for high vs. low-relevance classification. We compare the different classifiers to see which approach gives the best results using our features.

The comparison of Naive Bayes and SVM and how they performed is reported in Publications V and VII.

Table 3.2: Classes of discourse features

Feature	Relevance	Comparison
Layout features		
1. <i>event-trigger-is-in-header</i>	3.3	3.10
2. <i>some-trigger-found-in-header</i>		
3. <i>trigger's-relative-location-in-document</i>	3.2	
4. <i>actor-in-trigger-or-header</i>		
4a. <i>actor-in-trigger</i>		
4b. <i>actor-in-header</i>	3.4	
4c. <i>actor-in-headline</i>		
5. <i>country-in-trigger-or-header</i>		
6. <i>document-length</i>		3.11
Compactness		
7. <i>trigger-actor-distance</i>		3.14
7a. <i>trigger-disease-distance</i>	3.5	
7b. <i>trigger-country-distance</i>	3.6	
8. <i>is-actor-found-before-trigger-sentence-end</i>		
9. <i>num-uniq-countries/actors-in-trigger-sentence</i>		
9a. <i>num-uniq-countries-in-trigger-sentence</i>		
9b. <i>num-uniq-actors-in-trigger-sentence</i>		
10. <i>num-uniq-countries/actors-until-trigger-end</i>		
11. <i>num-uniq-countries/actors-in-document-events</i>		
12. <i>contains-valid-country/actor</i>		
13. <i>content-repeated-in-header-or-document</i>	3.7	
13a. <i>content-repeated-in-header</i>		3.12
13b. <i>content-repeated-in-document</i>		3.13
14. <i>num-of-events-in-document</i>		
Time		
15. <i>event-has-time-of-occurrence</i>		3.15
16. <i>distance-trigger-sentence-to-date-span</i>		
17. <i>time-diff-pubdate-event-start/end</i>	3.8	
Low relevance indicators		
18. <i>is-blacklisted-data-in-header/headline/document</i>		
19. <i>num-of-harm-events-in-document</i>		
Domain-specific		
20. <i>is-named-victim</i>	3.9	
21. <i>is-unspecified-illness</i>		

3.4 Results: domain features

For predicting the relevance of the extracted information, two sets of features were gathered, *lexical features* and *discourse features*. Lexical features are simpler, low-level features based on bags of words. Lexical features capture local information, while discourse features capture longer-range relationships within the document.

Discourse features are based on properties of the article text and of the events extracted from it. The features are listed in Table 3.2. Some of the discourse features were chosen following an intuitive importance based on experiences gained from the customization process. The guidance and justification for the feature choices came from the text linguistic literature about cohesion and coherence in text (Hoey, 2001, 1991; Berzlanovich et al., 2009; Tanskanen, 2006; Halliday and Matthiessen, 2004; Halliday and Hasan, 1976). Some of the chosen features are based either on their claimed power of signalling e.g., important passages in text, that is, something that meets the readers' expectations on what is supposed to be there, or, on signalling another kind of importance in the text.

A set of discourse features used for the study express position of the event in the document (having to do with structural intra-document qualities). Another set of the features express compactness, e.g., the distance of the event attributes from the event as opposed to scattering (of the event's attributes), which is structural intra-event quality. Recency in time or the recency of the events' occurrence is one feature. For example, 'last week' refers to a period of a week earlier from two years ago, if the article was written two years ago. That is a pragmatic quality.

The features are handled in the form of a rule and observations on how often these features, or rules, succeed, or indicate that an item might be relevant or non-relevant, and how often the features fail or the rules are violated. There are features that e.g., signal importance, like position or repetition. The writer is using repetition that signals that (reader's) expectation has been met (about what the content or topic of the text and what questions it is answering) (Hoey, 1991). Cohesive ties, e.g., repetition, connect sentences close by and far away from each other.

The aim is to see whether a feature correlates with the event being highly relevant. For example, *actor-in-headline* indicates whether the main actor or target of an event was also mentioned in the headline of the text. We test whether it is true that the event where this feature appears is relevant or less relevant, with counts of true/false positives/negatives. This tells how reliable the feature is. For example, a feature states that "for the article to be relevant, the noun phrase of major importance must appear in the headline." A noun phrase of major impor-

tance includes a scenario dependent central predefined fact such as a name of a disease, disaster, or name of a company as actor/target. By training the classifier on the tagged data, we measure whether it is true or false, that is, whether this is useful information, how often disease in headline predicts relevance.

The features can be domain-specific or domain-independent, but the approach in general is domain-independent. The features aim at being as general as possible. A possible feature could be “the disease must be mentioned in the headline and in the first sentence.” But having the disease name as a part of a feature is a bad generalization if it is for all scenarios. For domain-independence, the same features should be applicable, if possible, to all scenarios. Parallel to disease name, in Security-scenario there is the suspect/item that is/is not in header (etc); in Business-scenario it is the name of a company (although the act of acquiring or merging might be more proper). The active participant is not necessarily a human agent but could be either disease, company, suspect, or item, and is defined in the domain specific knowledge bases and patterns in the IE system. The participants are here called ‘actors’. The feature mentioned above is thus called *actor-in-trigger-or-header*.

In Section 3.4.1 show how the individual discourse features predict relevance by themselves in Medical domain. In Section 3.4.2 we evaluated the predictive power of lexical and discourse features and how they perform with either lexical or discourse features, or both combined. Section 3.4.3 examines whether there is variation between the scenarios.

3.4.1 Discourse features

In this section the focus is on how the individual discourse features predict relevance by themselves and, in Section 3.4.3 whether there is variation between the scenarios, or the types of text.

The discourse features include information about the number of events, positioning of the event in the document, the compactness of the placement of the event’s attributes (Bagga and Biermann, 1997; Agichtein and Cucerzan, 2005), and the recency of event occurrence, and are listed in Table 3.2. The first set of features, the **layout features** reflect the position of the trigger sentence in the document and predict the event’s relevance for the end-user. The layout features allow us to quantify the assumption that important details of news topics are placed in the beginning of the article, whereas less important details are stated later according to the *inverted pyramid* principle (Bell, 1991).

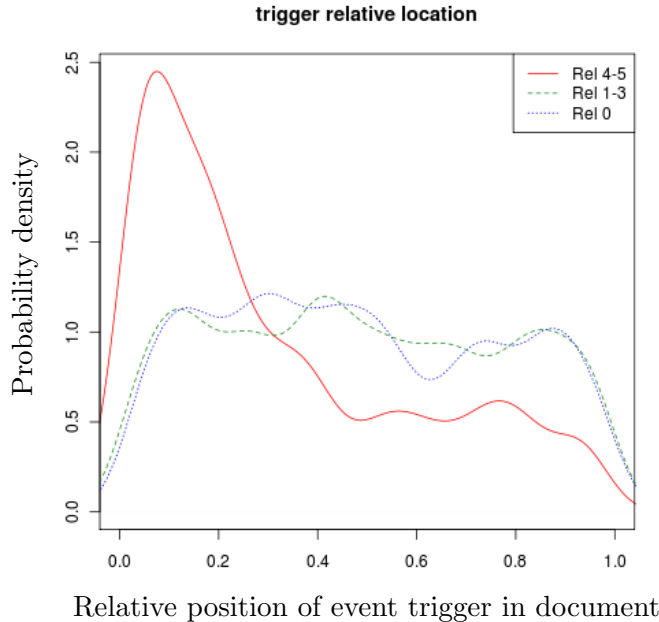


Figure 3.2: Medical scenario: Conditional distribution of *trigger's relative location-in-document*: the relative position of the trigger sentence in the document, given the relevance class/rating: high vs. low vs. zero. The relative position is the x-axis, the y-axis shows the probability density.

For the Medical scenario, Figure 3.2 shows the feature *trigger's relative location-in-document*. Sentences with very high relevance events tend to appear in the beginning of the document. Low relevance (1-3) or irrelevant (score 0) events were distributed randomly.

A high relevance trigger sentence seemed to appear most likely in the header of the document.⁴ We also examined whether the trigger sentence for that specific event exists in the header (feature *event-trigger-is-in-header*) and a closer look reveals that indeed the events with the highest relevance score are most probably, compared to other less relevant events, found in the header of the article, as shown in Figure 3.3.

⁴The term *header* included the headline and the first two sentences of the news article for space saving reasons. But generally, *header* is only the first paragraph, not headline.

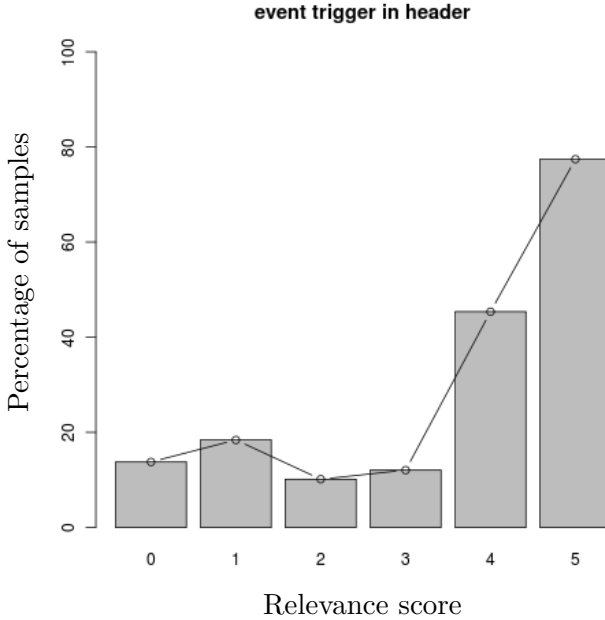


Figure 3.3: Medical scenario: Percentage of samples where *event-trigger-is-in-header* is present: the trigger of the event is in the header, given the relevance score.

The feature *actor-in-trigger-or-header* means, for the Medical scenario, whether the name of the disease appears in the header (headline or the following two sentences). Figure 3.4 shows that the presence of a disease name in header indicates high relevance event. This is parallel to the intuitive expectation that breaking news about epidemics mention the disease name relatively early in the article.⁵

In short, the Medical scenario shows a strong tendency for the initial position of an important event (including the actor, disease name).

The second set of features reflects the scattering of a single event in the text. The **event compactness** is measured by the distance of the event attributes from the trigger in the text. In a compact event, the event attributes are situated close to the trigger. The compactness features track the distance of mentions of the

⁵Note that the IE system extracts *unknown illnesses* and *mysterious illnesses* as diseases, as such occurrences are of particular importance for the end-user.

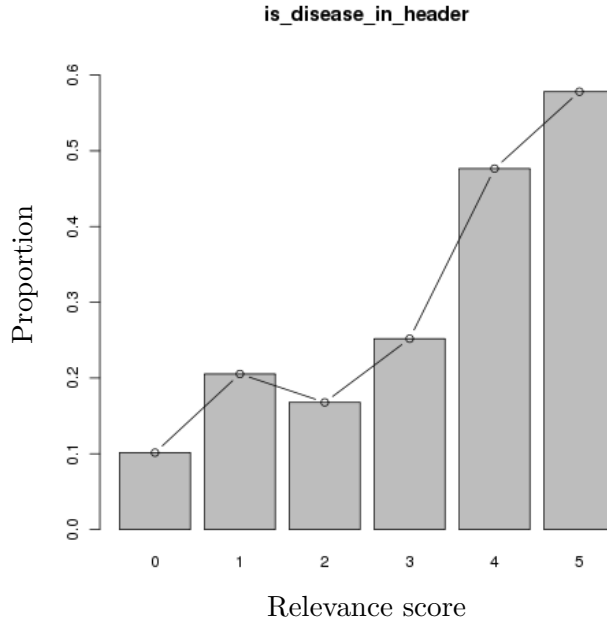


Figure 3.4: Proportion of the events where the feature *actor-in-header* is present: the actor (i.e., disease name) appears in header, given the relevance score for the event.

attributes from the event’s trigger. The effect of compactness of an event on its relevance is here modelled by, e.g., measuring the distance between the trigger and the actor. The *actors* are ‘active’ participating attributes of an event. The actor for the Medical scenario is the name of the disease, and for the Business domain the name of a company. The distance may be measured as the number of characters, words, or sentences.

Figure 3.5 with the feature *trigger-disease-distance* shows the distribution of the distance between the disease name and the trigger in Medical scenario. The horizontal axis shows the distance in sentences. We grouped the instances where the distance was greater than 10, or the disease name was not mentioned. This group is plotted at x-axis with the value 15 (for visualization).

The Figure compares between relevance zero vs. one to three vs. four to five relevance events. For events that contain no actor (i.e., a disease) at all, the

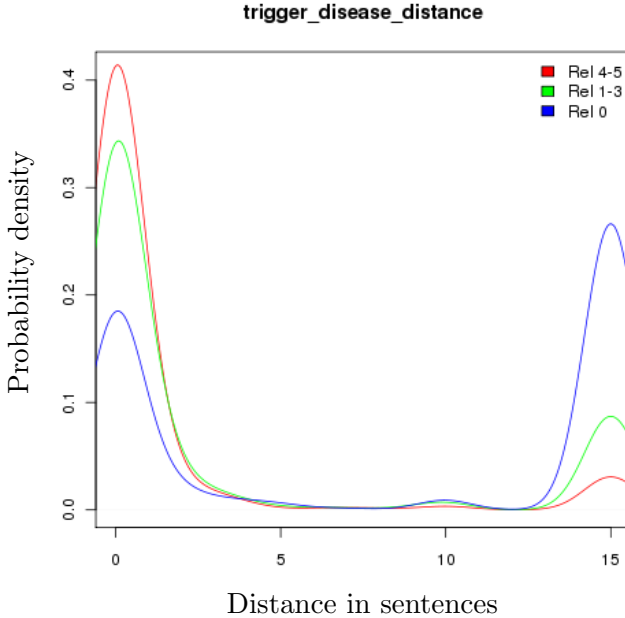


Figure 3.5: Medical scenario: feature *trigger-disease-distance*: Distance of actor (namely, disease name) from event trigger. Note: instances where no disease is mentioned in the document (or distance > 10) are plotted at 15 for visualization.

feature receives a special value NA. The disease names of the least relevant events are furthest from the trigger, and the disease names of the most relevant events are closest to the trigger. That is, the name of the disease is more likely to appear in a trigger sentence of a high-relevance event.

The distance of the geographic location from the event trigger is also a compactness feature. There is variation in the importance of the location in the three scenarios. In Business scenarios, the acquisition events often do not even mention the location in the article text, and also for the Product Launches scenario it is not mentioned often. But for the Medical scenario the location is essential. The probability density of the feature *trigger-location-distance*⁶ is shown in Figure 3.6 for the Medical scenario. The location strongly tends to appear close to the trig-

⁶This is similar to *trigger-actor-distance* signalling the distance between the location and the trigger sentence.

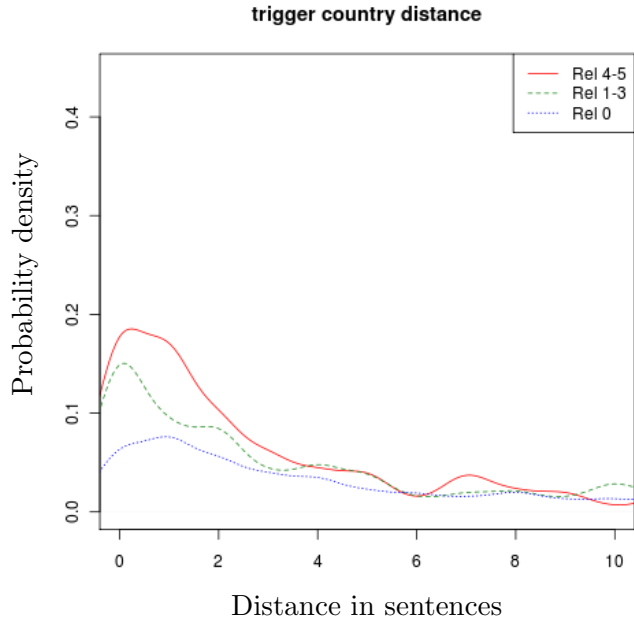


Figure 3.6: Medical scenario, feature *trigger-country-distance*: Distance from country to trigger sentence.

ger sentence, if the event is of high relevance. However, the name of the disease appears closer still.

A subset of the compactness features are the *content repetition* features. They measure whether an important slot filler, such as an actor (e.g., disease name), is repeated inside the document. These features are chosen based on the assumption that repeated mentions of a key actor should positively affect the relevance. Conversely, features that count the number *distinct* actors mentioned in the text may serve as good indicators of *lower* relevance. For example, an article mentioning several diseases or companies is less likely to be of high relevance, containing very specific and topical news. Rather it may be more likely to be a broader review or overview article.

We observe repetition of a part of an event, that is, one or more slots of an event template are repeated in the header or the rest of the text body (feature *content-repeated-in-header-or-document*). Figure 3.7 shows that an attribute is more likely

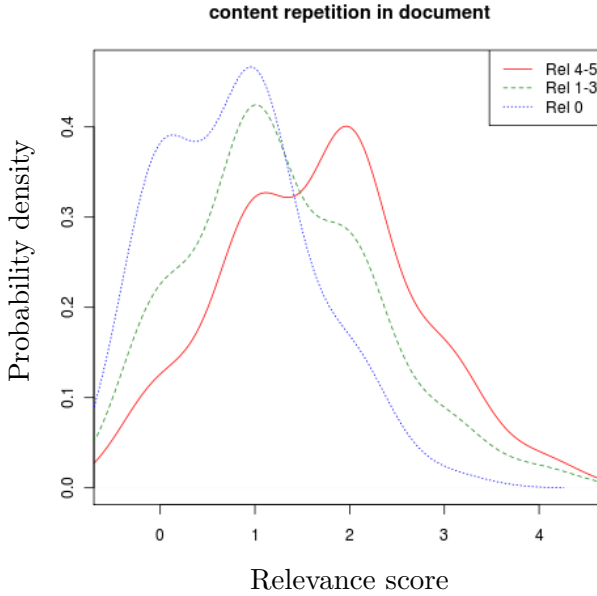


Figure 3.7: Feature *content-repeated-in-header-or-document*: Repetition of content in the Medical scenario

to be repeated if the event is rated high-relevance. In the lower-relevance events, an attribute is less likely to be repeated.

Time and recency features relate to the recency of an event, comparing the time attributes of an event with the publication date of the news article saved in the metadata, i.e., the difference between publication date and the reported event date (feature *time-diff-pubdate-event-start/end*), as seen in Figure 3.8. The PULS system may also extract hypothetical events and events which have a projected or expected date in the future. These type of events are rarely relevant for the end-user monitoring real-time activities. So recency is a good indicator for relevance of news since highly relevant articles usually describe more recent events. In Medical scenario, for example, events that occurred too long ago are of very low relevance, such as “*King Tut died of sickle cell disease, not malaria.*”⁷

⁷The cause of death of King Tut was breaking news that was reported in many newspapers in 2010, over 3000 years after it occurred.

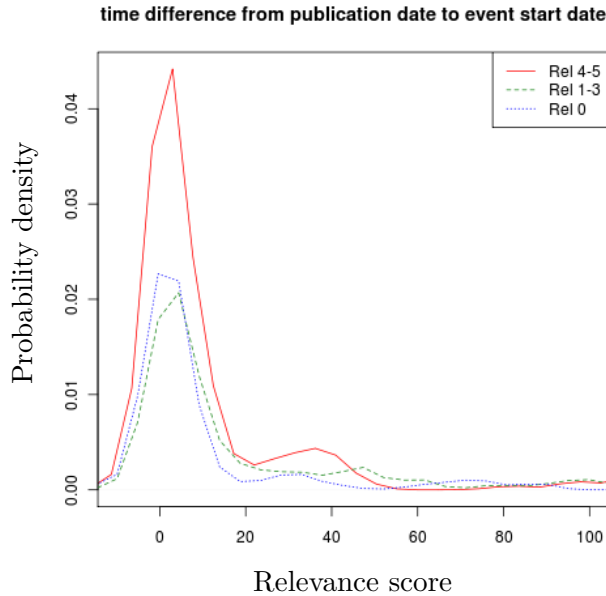


Figure 3.8: Feature *time-diff-pubdate-event-start/end*: Distribution of the difference days from publication to event date in the Medical scenario. (Negative values indicate events in the future. Instances where the feature value is > 120 or where event date is missing are plotted at position 120, outside the figure.)

We examined a number of **domain-specific features**. For example, for the Medical scenario, the system extracts names of victims where possible. One example is the *is-named-victim* feature. Some news articles about genuine epidemic outbreaks name the victim as well, in an attempt to personalize them for the reader and make the experience more immediate. But obituaries, articles about public figures, and other irrelevant items from the epidemiological perspective, name the victims as well. A better way to predict the relevance is to count the named victims. If there are named victims mentioned in an article assumed to report about an infectious epidemic, it is correlated with possible irrelevance, as shown in Figure 3.9.

This feature is an example of a domain specific feature. Most features, however, are applicable directly to all domains. For example, the number of unique actors preceding the trigger sentences, is applicable in all domains. This feature

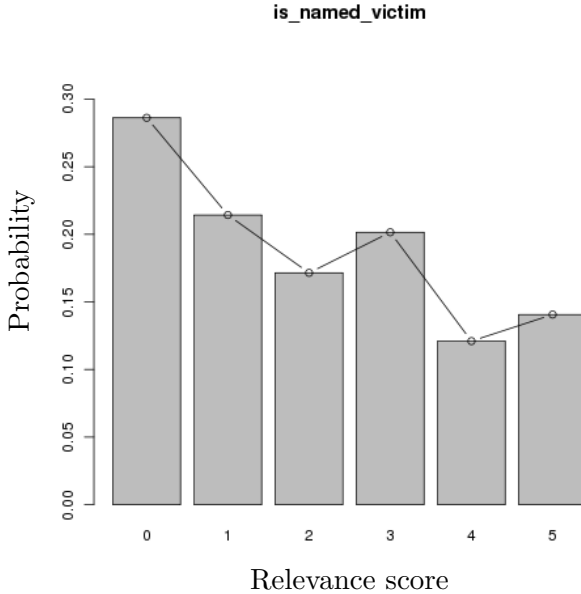


Figure 3.9: Feature *is-named-victim* and the relevance score in Medical scenario

is negatively correlated with relevance: the more named actors are found in the preceding sentence, the less relevant the article is. For the feature *actor-found-before-end-of-trigger sentence*, in the Medical scenario, if no disease names exist before the trigger sentence, then the document is likely irrelevant. In relevant events there is often only a single “actor”: in the Business domain the actor is a company name; in the disease domain, the actor is the disease. For the Security domain, defining and finding the right actor from the text is trickier. The actor role may be a perpetrator or a victim, but is not limited to that. Depending on the focus or point of view, the actor may be the target slot as well. The representatives of law are often the actors. They arrest criminals, which function more like targets, similar to the victims of human trafficking, forced labor, smuggled animals, and so on.

Three other kinds of features were considered for the experiment: *blacklisted events*, *negative events*, and *missing attributes*. The first two features are used in patterns as well. We devise a set of *blacklist features* that signal low relevance in a given domain. The blacklisted features are manually selected terms that

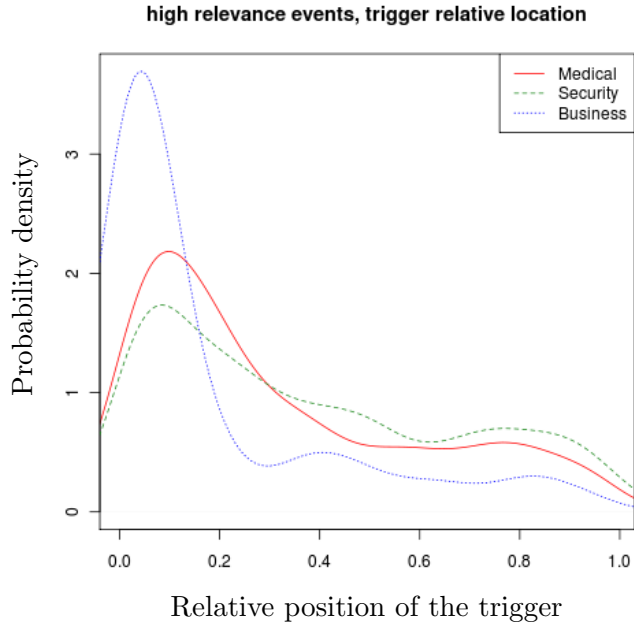


Figure 3.10: Feature *trigger’s-relative-location-in-document*: Relative location of the trigger sentence in three domains

commonly appear in irrelevant documents. If a blacklisted feature appears in an extracted event, it signals that the event is irrelevant, since it either belongs to a wrong topic or the event is irrelevant for other reasons. For example, in epidemic surveillance, terms such as *vaccination campaign* and *obituary* are strong indicators of low relevance for the entire news report. Obituaries in the Medical scenario are a common source of false positives, since the extraction patterns often fail to distinguish based on *local* cues alone whether a reported death is caused by an infectious disease epidemic or by another illness. For the Business domain, a blacklisted item is e.g., *President*, since it mostly refers to a head of state, rather than head of a company. Also *investigation* and *inquiry* are blacklisted, since they often appear in the same environment as a new product in the Product Launches scenario. It is easier to list items that are not new products, than all the possible new products in the world. If a blacklisted feature appears in a template slot, the event type is marked “blacklisted” and not shown to the user. However, it is taken

Table 3.3: Relevance classification (high vs. low) on Business and Medical scenarios. Accuracy for discourse, lexical, and combined features.

Business Domain						
Classifier	All events			First events only		
	Lexical	Discourse	Combined	Lexical	Discourse	Combined
SVM	72.2	84.6	85.3	70.4	81.8	82.2
Naive Bayes	74.3	75.7	82.5	70.3	81.7	82.2
Medical Scenario						
Classifier	All events			First events only		
	Lexical	Discourse	Combined	Lexical	Discourse	Combined
SVM	82.2	85.1	84.2	87.2	88.5	89.6
Naive Bayes	79.7	80.7	84.6	85.8	85.0	89.2

into account as a *negative event* by the classification tools for relevance prediction: it lowers the probability of being relevant for the other extracted events in that document.

PULS extracts *negative events* to capture the events that frequently interfere with events of interest. For example, in the business domain, satellite/rocket launches may trigger patterns for finding new product launches, since they are syntactically similar. Natural disasters (flooding, earthquake, etc.) with casualties often interfere with patterns of the Medical scenario. The number of negative events found in a document is a discourse feature. A *missing attribute* may also be an indication of a negative event. For example, our end-users, experts in the field of infectious disease epidemics, rejected or marked irrelevant events that were missing the name of the disease. The absent disease name was not the cause of rejection, but was merely a symptom of irrelevance.

3.4.2 Combining lexical and discourse features

As *lexical features* we use the bag of words in the trigger sentence and in the sentences immediately preceding and following the trigger sentence. The surrounding sentences provide additional context for disambiguation. For example, the trigger sentence picked up by the system may include deaths and injuries, but in principle the article could be about any kind of casualties.

To reduce data sparsity, the sentences are pre-processed by a lemmatizer, and passed through a named entity (NE) recognizer, which replaces persons, organiza-

tions, locations and disease names with a special token indicating the NE’s class. Stop words were dropped—prepositions, conjunctions, and articles.

We built relevance classifiers to test how well the lexical and discourse features indicate relevance in Corporate Acquisitions and Product Launches scenarios of the Business domain and Infectious Disease Outbreak scenarios (Medical scenario).

We evaluate the predictive power of the discourse and lexical features using two classifiers: Naive Bayes (John and Langley, 1995) and SVM (Platt, 1999). The evaluation results presented here are accuracies. The full experimental evaluation is reported in Publication V (for Medical scenario only) and Publication VII (for Medical and Business scenarios), and in Huttunen et al. (2013).

Since parts of the labeled data are *corrected* by the user, we have *two* parallel sets of events with relevance labels: the “raw” events, as extracted by the system, and the “cleaned” events, i.e., the same events with corrections. The raw set is more noisy, since it contains the errors that were introduced by the system.

The relevance classifiers are trained using cleaned labeled data. But for evaluation, the classifiers are tested against both the cleaned and the raw events, although we are most interested in the performance on the raw events: ultimately, the goal is to build a classifier that is applicable to the real extracted event stream. The raw scores in evaluation give us an indication of what performance we can expect in the real-world setting.

These classifications were obtained on approximately 530 Medical documents, containing 900 events, (see Section 3.2 for the data). In the Business domain, we used 127 documents with 213 user-labeled events.

For each classifier, Table 3.3 shows the performance using discourse features only, lexical features only, and the combined set of features, and compares performance on *all* events vs. the *first* event in the document. In the table, the bold score indicates the best score achieved for the given classifier and the type of events considered. We can see that in the Business and Medical scenarios, the discourse features alone perform better than lexical features most of the time. Also, in most cases, for both domains, combining discourse and lexical features performs better than using either discourse or lexical features alone.

3.4.3 How scenarios reflect discourse features

In this section we examine how scenarios differ with respect to features strongly signaling high relevance, which we investigated above. A comparison of the scenarios—Infectious Disease Outbreaks (Medical), Business and Security—shows differences related to features predicting high relevance. For this comparison we

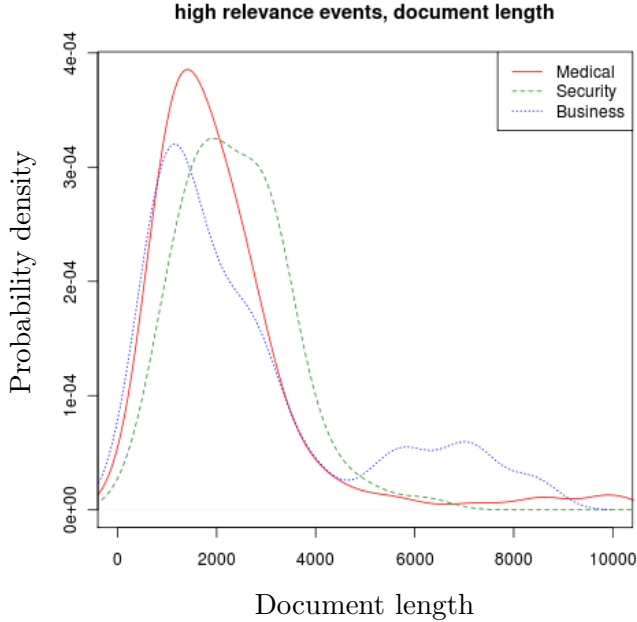


Figure 3.11: Feature *document-length*: Document length in the three domains for high-relevance events

use a subset of the features in Table 3.2. The features in Table 3.2 were tested on the Medical scenario, and were shown in Section 3.4.1 to be useful for predicting relevance of events.

In the Medical scenario, the **layout features** show a strong tendency for the initial position of an important event (including the actor, disease name) as seen in Figure 3.3 for the feature *event-trigger-is-in-header*, and in Figure 3.4 for the feature *actor-in-header*.

However, comparing Business, Medical and Security scenarios with respect to the initial position feature *trigger's-relative-location-in-document*, shows a difference between the scenarios, seen in Figure 3.10 on page 85. In the Business articles this tendency is strongest. Medical articles also strongly favour the initial position of a relevant event. In the Security domain the high-relevance facts seem to be most spread across the article, even if the preference for initial position is still strong there as well.

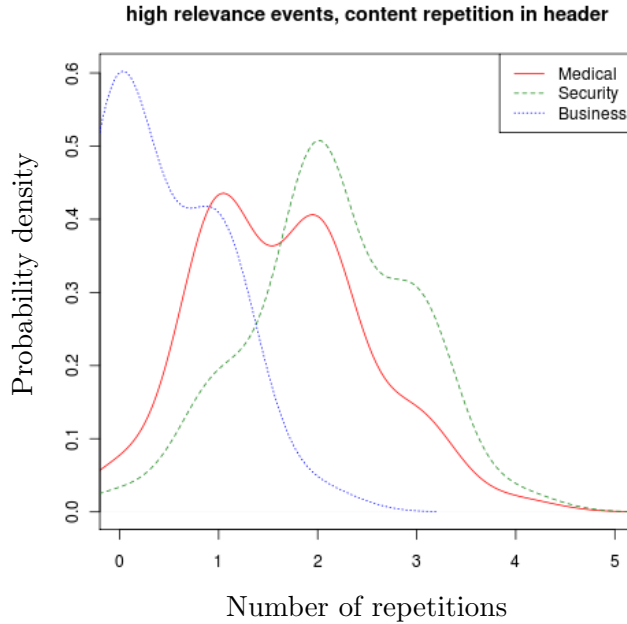


Figure 3.12: Feature: *content-repeated-in-header*: Event attributes repetition in header in the three domains.

Layout features also include *document-length* and the position of the trigger sentence in the document (*trigger's-relative-location-in-document*). The length of the document in the Medical, Business, and Security scenarios is shown in Figure 3.11 for high relevance events. There seems to be no big difference between the three domains in the length of the articles containing most relevant events, although the Medical scenario seems to have the least variation in length.⁸

For the **event compactness** features, the location tends to appear close to the trigger (*trigger-country-distance*) in high relevance events in the Medical scenario, as shown in Figure 3.6. Some of the compactness features test whether an important fill is repeated in the document. For example, if an actor is repeated, it is likely to affect relevance positively (*content-repeated-in-header-or-document*). Figure 3.7 showed that an event attribute is more likely to be repeated twice in

⁸The *document-length* feature would be interesting to study in relation to relevance, but that is left for future work.

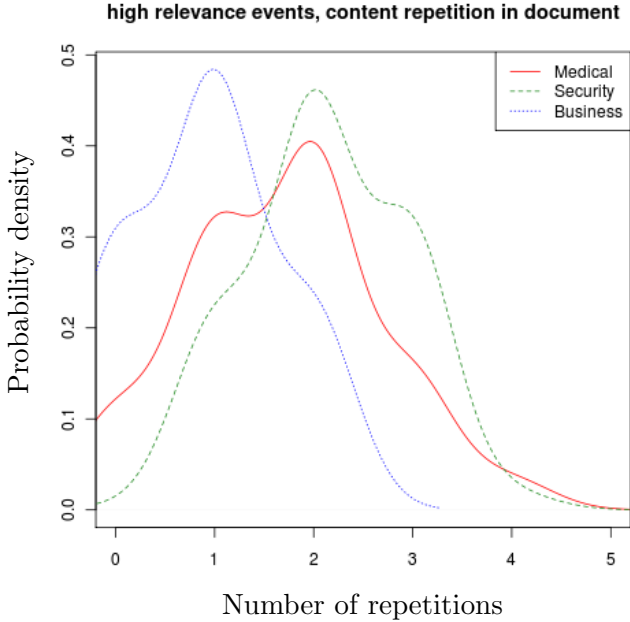


Figure 3.13: Feature *content-repeated-in-header-or-document*: Content repetition in entire document in the three domains

documents that are rated high-relevance in the Medical scenario. The actors that are included here are the name of the disease for the Medical scenario, the name of the company for the Business domain, and the name or description of the suspect or item for the Security domain, and for all three domains, the location, time, and descriptor (actor). All of these occupy an event slot at least twice in the headline, header, or the rest of the document. Figure 3.12 shows the probability density of the repetition feature *content-repeated-in-header*, the repetition of a mention of possible actors in the three domains. The Figure shows the repetition in the document header.

According to Figure 3.12, the Business domain has the least repetition of high relevance events, whereas the Security domain has the most repetition on average. The Medical scenario appears in between.

In Figure 3.13 Medical, Business and Security domain are compared for the content repetition feature *content-repeated-in-header-or-document* in the entire

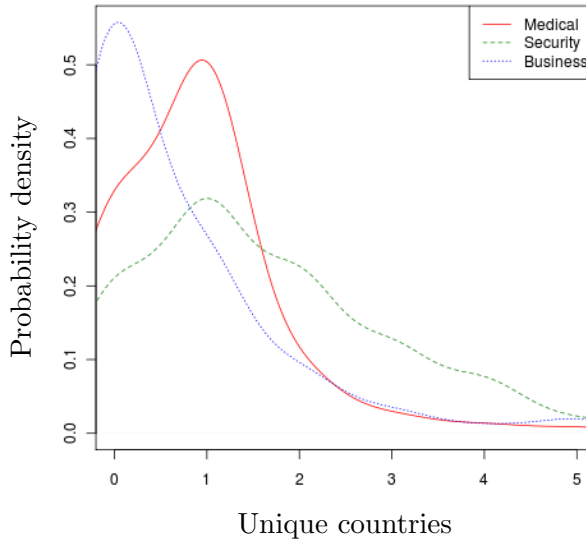


Figure 3.14: Feature *num-uniq-countries/actors-until-trigger-end*: Number of unique countries mentioned from the beginning of the document until the event trigger in high-relevance documents, in the 3 domains.

document. The Security domain repeats events most often, the Business domain the least. This is probably due to the narrative nature of Security domain, and the changes in perspective (perpetrators, victims and officials).

All the repetition seem to be relatively the same in all three domains. The Business domain articles repeat, if at all, attributes more in the rest of the document than specifically in the header, at least compared to the two other domains, especially the Security domain that prefers the repetition in header.

The feature *num-uniq-countries/actors-until-trigger-end* is shown in Figure 3.14 across the three domains. The locations are counted only until (and including) the trigger sentence, not for the entire document (unless the trigger appears in the last sentence). The figure shows that location appears more repeatedly in Medical documents than in Business documents.⁹ In the Business domain the country is not mentioned at all in 50% of the documents. The Security domain has a more

⁹One exception in the Business domain may be Product launches, where the target area for the new line of product is less punctual.

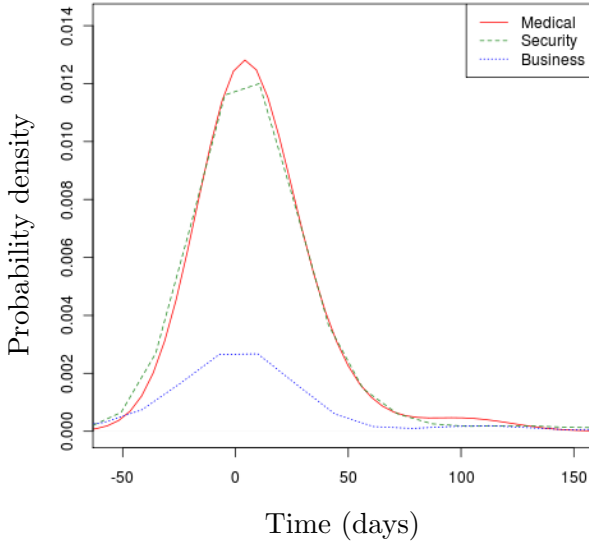


Figure 3.15: Feature *time-diff-pubdate-event-start/end*: Time difference between event date and publication date in three domains

even distribution, between one and two mentions of the countries in a document, over 30% have one country and over 25% have two countries. The Security domain has a higher rate of three, four and five countries than other domains; this is unsurprising, since the Security domain focuses on *cross-border crime*.

Since an infectious disease epidemic is geographically less punctual than a Business deal, as discussed in Section 2.4.1, it is not surprising that there are more mentions of locations in Disease news articles than in Business articles.

Comparing the three domains and **recency in time** feature (*time-diff-pubdate-event-start/end*) appears in high-relevance documents in all three domains the same way, see Figure 3.15. In Security and Medical scenario event time coincides clearly with the publication time when the event is rated highly relevant. For the Business domain, the time appear less important overall, as seen also above in the discussion about temporal expressions, Table 2.15 in Section 2.4.1.

Chapter 4

Summary and conclusion

My objectives in this study, first presented in Section 1.2, is to understand the reasons for the difference in performance for different scenarios and to find ways to improve the results produced by “traditional” IE methods. In this Section I summarize the results in relation to the objectives.

4.1 Summary of results and discussion

1. What accounts for the variation in the performance of an IE system on different scenarios?

There are systematic differences in how news topics are reported, (Chapter 2 and, Publication II). Both lexical and structural characteristics of the texts have an impact on how well the pattern-based IE system performs. Patterns built for Nature scenarios fail to capture events as easily as patterns built for Business scenarios, so the extraction results remain poorer for the Nature scenarios (Section 2.4.3 and 1.5.1, Publication I). The syntactic-semantic patterns for Nature scenario reports are harder to construct because the language is more emotionally loaded, linguistically less compact, have more variation in linguistic expressions and in the organizing of the facts (Section 2.4.2 and Section 2.4.1).

2. How do the scenarios differ on the lexical level and on the structural level, and why?

The scenarios and text types are interlinked. The differences between scenarios are at both the lexical and the structural level (Section 2.4, Section 1.5.3, Pub-

lication III, and 1.5.4, Publication IV). At the lexical level there are a greater number of lexical choices in Nature scenarios (Section 2.4.1). At the structural level the differences are in organization of the arguments belonging to the same event, and how the relationship between the arguments are marked, that is, the cohesive relations between the arguments (Section 2.4.2). The author of a news report tries to make the text interesting for the reader or meet the reader's needs (Section 1.3.1). The business news are mostly stated as facts, with factual style of writing. These facts need to be unambiguously understood and quickly processed.

Natural disaster news have some characteristics of fictional writing and are often written by named contributing journalists, with a more dramatic style, as catastrophes and human suffering touch emotions deeply, and attract readers. The Natural Disaster news reports are the longest ones on average (Section 2.4.3). The amount of characteristics of fictional writing depend on the source of the article, that is, whether it is meant for a wide audience with the intention of affecting emotions of the readers vs. weather reports.

The Medical scenario has both fictional and factual characteristics, depending on the source. There may be differences in the writer's intent or goals, as a reader of a medical post about a possible disease outbreak may need to take action and needs to have all the facts at his or her disposal, and preferably nothing extra.

News about infectious diseases is more vague than business news because of the uncertainty and vagueness that relates to the spread of a disease (Section 2.4). The infectious disease data in this study is mostly written by medical professionals, so the reports are more informative and less descriptive than news about natural disasters (Section 2.2).

Even if the topic is the same, the style of writing may vary as with business news reports vs. abstracts based on those reports (Section 2.4.3). In Business news, there are lexical differences between *business abstracts*, targeted to business specialists, and *business articles*, targeted to the general public. The two kinds of Business domain data should be treated separately (Section 2.4).

On the lexical level, more varied language and colourful vocabulary used in expressing the Nature events, especially with Natural Disasters, similar to reports about sports or war. The characteristics of natural disaster reporting include (Section 2.4.1):

- metaphorical/idiomatic expressions,
- verbs/nouns/appositions that are harder to anticipate,
- vivid verbs, synonyms, metaphors,

- a greater number of adjectives,
- personification of non-human participants.

The events are scattered in text, combined together mostly in the reader's mind, with the help of the cues given by the writer and the reader's pragmatic knowledge (Section 2.4).

As we discuss in Huttunen et al. (2002), business news use:

- standard, formulaic, and more rigid style, to minimize confusion or distraction for the reader,
- familiar, predictable terminology, which makes pattern-based IE more effective.

The sparing use of synonyms in the Business domain is due to the strictly fact-oriented style of reporting, and the formulaic ways of expressing the events, e.g., corporate acquisitions. By comparison, different ways can be used to verbalize how people contract a disease, or how a disaster strikes a location. This is especially true of general news, compared to the more scientific text types, such as medical reports written by experts (Section 2.4).

Business abstracts/summaries contain even less ambiguous terms and strictly scenario-specific language. News articles have a more varied vocabulary.

The differences on the structural and lexical level between the Nature domain and the Business domain are due to the function of the text and the nature of the scenario. Events in the Nature scenarios move through time and space. In the Business domain the activities are typically "punctual" (Section 2.4.2).

As we discuss in detail in Publication IV, for the Nature scenarios this causes:

- the use of evocative, unpredictable constructs,
- scattering of events, e.g., cause and effect are in separate clauses/sentences or even paragraphs,
- vagueness within the scope of an event,
- hierarchical organization between events, marked by grammatical, lexical and discursive means.

For the Business scenarios:

- a relatively compact, SVO sentence structure,

- shorter sentences,
- a flatter event structure.

3. For a given scenario, how could these differences be traced, measured and ultimately leveraged?

This consists of two sub-questions:

- a. to improve the IE system in general?
- b. to improve the IE system in relation to user needs?

To address Question 3a, the approach chosen here is first, to focus on few scenarios through information extraction.

In some of the more complex IE scenarios, events appear far apart from each other across the text and the relations between those events may be difficult to establish. In the Publications 1–4 this problem is examined, and some solutions in the form of cues are discussed. To capture the Nature events better, the patterns should tend toward modular extraction, that is, combination of event-level multi-slot patterns and relevant NPs that fill individual slots, but only in case the extraction system has found sufficient evidence of an event in the vicinity of the trigger and the discovered arguments. This evidence can be gathered by inference rules. For example, in the smuggling scenario of cross-border crime, it is not sufficient to find a trigger sentence and objects in text that belong to concept classes such as *MEANS* or *DRUG*. In order to conclude that smuggling has occurred between two countries, we require the appearance of at least two location names in the vicinity of the event.

To address Question 3b, we focus on issues that are of high relevance to a user, and narrow down the features that may carry impact. We demonstrate in Publications V and VII, that by means of detailed linguistic analysis of the scenario, we can identify the features which impact the quality of extracted events. In particular, we show that the relevance and usefulness of the events can be modeled effectively by means of such features. We found strong correlations between the presence of such features and how useful the events are for the end-user.

4.2 Conclusions and implications for future work

The analysis presented in this thesis and in the attached publications is relevant for the approaches to the various established sub-problems of pattern-based IE.

I believe that this analysis is relevant for approaches that have emerged more recently as well.

Recent research in neural networks (NNs) has several views. One of them is using only the input from the lowest level, e.g., only the tokens in the text, or even only the individual characters in the text. The claim is that utilizing any higher-level features is not necessary, since the network will learn them automatically and feature-engineering is not considered essential. In some settings and some tasks, this has been shown to be achievable to some extent (Feng et al., 2018; Vaswani et al., 2017).

For these reasons, the features presented in this thesis are relevant in today's NLP environment, including in IE, and may be directly usable in many complex, high-level analysis tasks, including NN-based approaches.

These features may easily be captured by simple methods, as presented here, and may be significant for the task, and therefore may enhance performance, (Senrich and Haddow, 2016; Yin et al., 2016). The high-level, hand-crafted features, such as those that I presented here, are easy to compute from the text, for example, how many times is a disease mentioned in the text, is the disease in the header, etc.

Then they can be given as additional inputs to a neural network designed to perform various IE-related tasks.

If take such an approach, it will add to the transparency of the overall model. Further, useful linguistic information will not be going to waste and will have a chance to help the model. We have already shown that these features help with machine learning, in several papers in this thesis.

References

- Agichtein, E. and Cucerzan, S. (2005). Predicting accuracy of extracting information from unstructured text collections. In *Proceedings of the 14th ACM international conference on Information and knowledge management*.
- Appelt, D., Hobbs, J., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Meyers, K., and Tyson, M. (1995). SRI intl. FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, Columbia, MD. Morgan Kaufmann.
- Atkinson, M., Belayeva, J., Zavarella, V., Piskorski, J., Huttunen, S., and Yangarber, R. (2010). Real-time text mining in multilingual news for the creation of prefrontier intelligence picture. In *Proceedings of the 6th Conference on Knowledge Discovery and Data Mining (KDD-2010) ACM SIGKDD Workshop on Intelligence and Security Informatics*. IEEE ISI-2010: Intelligence and Security Informatics.
- Atkinson, M., Piskorski, J., der Goot, E. V., and Yangarber, R. (2011). *Counterterrorism and Open Source Intelligence*, volume Vol. 2 of *Lecture Notes in Social Networks*, chapter Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering, pages 355–390. Springer-Verlag.
- Bagga, A. and Biermann, A. W. (1997). Analyzing the complexity of a domain with respect to an information extraction task. In *Proceedings of the 10th International Conference on Research on Computational Linguistics (ROCLING X)*.
- Bakhtin, M. and Ghāsemipour, G. (2011). The problem of speech genres. *Literary Criticism*, 4(15):114–136.

- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *In Proceedings of the 20th international joint conference on Artificial intelligence. IJCAI*.
- Beaugrande, R.-A. and Dressler, W. (1981). *Introduction to text linguistics*. Longman.
- Bell, A. (1991). *The Language of News Media*. Springer-Verlag.
- Berzlánovich, I., Egg, M., and Redeker, G. (2009). The interaction of coherence and lexical cohesion across genres. In Backus, A., Keijzer, M., Vedder, I., and Weltens, B., editors, *Artikelen van de Zesde Anéla-conferentie*, pages 33–42. Delft.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.
- Biber, D., Connor, U., and Upton, T. A. (2007). *Discourse on the move: using corpus analysis to describe discourse structure*, volume 28 of *Studies in Corpus Linguistics*. John Benjamins B.V.
- Biber, D. and Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44.
- Cardie, C. (1997). Empirical methods in information extraction. In *AI Magazine*, volume 18, pages 65–79.
- Connor, U. (1996). *Contrastive Rhetoric: Cross-Cultural Aspects of Second-Language Writing*. Cambridge University Press.
- Cvitas, A. (2010). Information extraction in business intelligence systems. In *The 33rd International Convention MIPRO*.
- Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. In *LREC*, volume 2(1), pages 837–840. Lisbon.
- Eggs, S. (1994). *An Introduction to Systemic-Functional Linguistics*. Pinter.

- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Webscale information extraction in knowitall (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*.
- Feng, X., Qin, B., and Liu, T. (2018). A language-independent neural network for event detection. *Science China Information Sciences*, 61(9):092106.
- Freifeld, C., Mandl, K., Reis, B., and Brownstein, J. (2008). Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. In *Journal of American Medical Informatics Association*, volume 15.
- Gildea, D. (2001). Corpus variation and parser performance. In *Proc. Conf Empirical Methods in NLP*.
- Gledhill, C. (2000). *Collocations in Science Writing*. Günter Narr Verlag, Tübingen.
- Gregory, M. (1988). Generic situation and register: A functional view of communication. In J.D. Benson, M. C. and Greaves, W., editors, *Systemic Perspectives in Discourse*, volume 1 of *selected theoretical papers from the 9th Intl. Systemic Workshop*. Ablex, Norwood, NJ.
- Gregory, M. (2001). Personal communication.
- Grishman, R. (1995). The NYU system for MUC-6, or where's the syntax? In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, Columbia, MD. Morgan Kaufmann.
- Grishman, R. (1997). Information extraction: Techniques and challenges. In Pazienza, M. T., editor, *Information Extraction*, Lecture Notes in Artificial Intelligence. Springer-Verlag.
- Grishman, R. (2019). Twenty-five years of information extraction. *Natural Language Engineering*, 25:677–692.
- Halliday, M. (1985). *Introduction to Functional Grammar*. Edward Arnold, London.
- Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

- Halliday, M. and Matthiessen, C. (2004). *An introduction to functional grammar*. Arnold, 3d edition.
- Harris, Z. S. (1988). *Language and information*. Number 28 in Bampton lectures in America. New York : Columbia University Press.
- Hirschman, L. (1998). Language understanding evaluations: Lessons learned from MUC and ATIS. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*.
- Hirschman, L., Brown, E., Chinchor, N., Douthat, A., Ferro, L., Grishman, R., Robinson, P., and Sundheim, B. (1999). Event99: A proposed event indexing task for broadcast news. In *Proc. DARPA Broadcast News Workshop*, Herndon, VA.
- Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.
- Hoey, M. (2001). *Textual interaction. An introduction to written discourse analysis*. Routledge.
- Huttunen, S., Vihavainen, A., Du, M., and Yangarber, R. (2013). *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing*, chapter Predicting Relevance of Event Extraction for the End User. Springer-Verlag.
- Huttunen, S., Yangarber, R., and Grishman, R. (2002). Diversity of scenarios in information extraction. In *Proceedings of LREC: 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain.
- Jeffries, L. (2015). Textual meaning and its place in a theory of language. *Topics in Linguistics*, 15.
- Ji, H. and Grishman, R. (2011). Knowledge base population: successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon. Association for Computational Linguistics.
- John, G. and Langley, P. (1995). Estimating continuous distributions in bayesian. In *Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann.

- Järvinen, T., Laari, M., Lahtinen, T., Paaajanen, S., Paljakka, P., Soininen, M., and Tapanainen, P. (2004). Robust language analysis components for practical applications. In *Proceedings of 20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland.
- Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as a corpus. *Computational Linguistics*, 29(3):333–347.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Lee, D. Y. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72.
- Li, J., Ji, H., and Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 73–82, Sofia, Bulgaria.
- Lijffijt, J. and Nevalainen, T. (2017). A simple model for recognizing core genres in the bnc. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, 19.
- Lucas, N. (2005). The enunciative structure of news dispatches, a contrastive rhetorical approach. In Ilie, C., editor, *Language, culture, rhetoric*.
- Macleod, C., Grishman, R., and Meyers, A. (1994). Creating a common syntactic dictionary of english. In *Proceedings of the International Workshop on Shared Natural Language Resources*, Nara, Japan.
- Maedche, A., Neumann, G., and Staab, S. (2005). Bootstrapping on ontology-based information extraction system. In *Studies in Fuzziness and Computing. Intelligent Exploration of the Web.*, pages 345–359. Physica-Varlag GmbH.
- Mann, W. C., Matthiessen, C. M. I. M., and Thompson, S. (1989). Rhetorical structure theory and text analysis. Technical report, University of Southern California.
- Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

- Martin, J. (1992). *English text: system and structure*. John Benjamins.
- M.Hall, Frank, E., Holmes, G., Pfahringer, B., P.Reutemann, and , I. W. (2009). The WEKA data mining software: an update. In *SIGKDD Explor.Newsl.11(1)*.
- MUC-6 (1995). Proceedings of the 6th message understanding conference (MUC-6). Morgan Kaufmann.
- MUC-7 (1998). *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Morgan Kaufmann.
- Patwardhan, S. and Riloff, E. (2006). Learning domain-specific information extraction patterns from the web. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, pages 66–73. ACL.
- Piskorski, J. and Yangarber, R. (2013). Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer-Verlag.
- Pivovarova, L., Huttunen, S., and Yangarber, R. (2013). Event representation across genre. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Atlanta.
- Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization.. In *Advances in kernel methods: support vector learning*. MIT Press, Cambridge.
- Redeker, G. (2000). Coherence and structure in text and discourse. In Black, W. and Bunt, H., editors, *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics*, pages 233–263, Amsterdam. Benjamins.
- Riloff, E. (1995). Little words can make a big difference in text. In *Proceedings of SIGIR-95*.
- Sager, N., Nhand, N. T., Nevin, B. E., and Johnson, S. (2002). The computability of strings, transformations, and sublanguage. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 79–120.
- Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. (2007). Ontology-based information extraction for business intelligence. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*. Springer-Verlag.

- Sahran, A. and Imran, H. (2009). Machine learning approach for automatic document summarization. In *Proceedings of World Academy of Science: Engineering & Technology*, pages 103–109.
- Santini, M. (2006). Web pages, text types, and linguistic features: Some issues. In *ICAME Journal*, pages 67–86.
- Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, 1(3).
- Saukkonen, P. (2001). *Maailman hahmottaminen teksteinä. Tekstirakenteen ja tekstilajien teoriaa ja analyysia*. Yliopistopaino.
- Schiffirin, D. (1981). Tense variation in narrative. *Language*, 1(57):45–62.
- Scollon, R. (1998). *Discourse As Social Interaction : A Study of News Discourse*. Longman Pub Group.
- Sekine, S. (1997). The domain dependence of parsing. In *Proc. Conf Fifth Conference on Applied Natural Language Processing*, pages 96–102. ACL.
- Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers*, pages 83–91.
- Shinyama, Y. and Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Human Language Technology Conference of the NAACL. HLT-NAACL*.
- Sinclair, J. (1991). *Corpus Concordance and Collocation (Describing English Language)*. Oxford Univ Pr.
- Soricut, R. and Marcu, D. (2006). Discourse generation using utility-trained coherence models. *Proceedings of Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL-06)*, page 803–810.
- Sperber, D. and Wilson, D. (1986). *Relevance*. Blackwell, Oxford.
- Strzalkowski, T., Wang, J., and Wise, B. (1998). A robust practical text summarization. In *AAAI Technical Report SS-98-06*.

- Svartvik, J. (1990). *The London–Lund corpus of spoken English: Description and research*, volume 82. Lund University Press.
- Taboada, M. and Mann, W. C. (2006a). Applications of rhetorical structure theory. *Discourse Studies*, 8(4).
- Taboada, M. and Mann, W. C. (2006b). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3).
- Tanskanen, S.-K. (2006). *Collaborating towards Coherence - Lexical Cohesion in English discourse*. Number 146 in Pragmatics & Beyond New Series. John Benjamins.
- van Dijk, T. A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Longman, London.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Vihavainen, A. (2011). Event relevance in information extraction. Master’s thesis, Helsinki University.
- Williams, G. (2002). *In search of representativity in specialised corpora: Categorisation through collocation*, pages 43–64. 22. John Benjamins Publishing Company.
- Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D., and Steinberger, R. (2007). Combining information about epidemic threats from multiple sources. *Proc. RANLP-2007 MMIES Workshop, International Conference on Recent Advances in Natural Language Processing*.
- Yangarber, R. and Grishman, R. (1997). Customization of information extraction systems. In Velardi, P., editor, *International Workshop on Lexically Driven Information Extraction*, Frascati, Italy. Università di Roma.
- Yangarber, R. and Grishman, R. (1998a). NYU: Description of the Proteus/PET system as used for MUC-7 ST. In *MUC-7: Seventh Message Understanding Conference*.
- Yangarber, R. and Grishman, R. (1998b). Transforming examples into patterns for information extraction. In *Proceedings of the TIPSTER Text Program Phase III*. Morgan Kaufmann.

- Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000a). Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany.
- Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000b). Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of the Conference on Applied Natural Language Processing ANLP-NAACL*, Seattle, WA, USA.
- Yangarber, R., Jokipii, L., Rauramo, A., and Huttunen, S. (2005). Extracting information about outbreaks of infectious epidemics. In *Proceedings of the HLT-EMNLP 2005, Demonstration*, Vancouver, Canada.
- Yates, A. (2009). Extracting world knowledge from the web. *IEEE Computer*, 42(6).
- Yin, W., Schütze, H., Xiang, B., and Zhou, B. (2016). Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Appendices

Appendix A

Examples of extraction

A.1 Intra-sentential

- MED20001210_11.26.23_7202-51
“Ebola fever has killed 156 people, **including** 14 health workers, in Uganda since September. -incl. by descriptor (people-health workers: subclass)
- MED20001211_22.19.10_13510-2
“**Of these cases**, 15 have died (12 nurses, a doctor, a medical assistant and an ambulance driver), 4 have been discharged, while 7 are still undergoing treatment, 3 ** in Masindi and 4** in Gulu. [how to describe conjunction?]
- MED20001211_22.19.10_13510-7
“The overall death toll has risen to 160, **after** 2 patients in Gulu and 2 in Masindi died since Tue 5 Dec 2000.
- MED20001211_22.19.10_13510-9
“In total 7 new cases in Gulu and none in Masindi have been confirmed since Tuesday, **raising** the overall number of people who have suffered from Ebola fever to 406.

A.2 Intra-paragraph inter-sentential

- 23.11.27-27286
“The public health officer responsible for Kombo south, Buba Manjang, has confirmed diagnosing at least **6 cases** of dysentery in Sifoe, Kartong and Gunjur recently. In July last year, a dysentery outbreak killed **6 people** and devastated many more in the same area, Kombo South, through a contaminated public well.

- 23.11.27-27286
*“Dysentery has this week claimed **2 more inmates** at King’ong’o GK Prison. The deaths bring to **6** the number of prisoners killed by the disease in the last 3 weeks. **The 2** died at the Nyeri Provincial General Hospital where they had been undergoing treatment after complaining of diarrhea and severe stomach pains.*
- 23.11.27-27286
*“AHMEDABAD: On Thursday **6** more patients with gastroenteritis were brought to the Chhipa Welfare Hospital from Jamalpur ward while municipal authorities continued to make contradictory statements about the figures involved. Of these patients, 2 were admitted to the Chhipa hospital. With **30 people** having contracted the waterborne disease on Wednesday [15 Nov 2000] from the Jamalpur ward, **the new cases** increase the toll to **36** in 2 days.*

A.3 Inter-paragraph

- *“KAMPALA: The Ministry of Health has instituted an investigation into the spread of Ebola hemorrhagic fever among **health workers** in Gulu and Masindi as the death toll among such workers rose to **15**. The Assistant Commissioner for National Disease Control, Dr. Alex Opio, said the investigation would also cover the persistence of the disease in the 2 districts. Briefing the press at the Ministry headquarters in Wandegaya yesterday [8 Dec 2000], Opio said some of the fallen health workers were not working in the Ebola wards. A total of **26 health workers** have contracted the virus in Gulu and Masindi, according to the Health Ministry records. Of these **cases**, **15** have died (12 nurses, a doctor, a medical assistant and an ambulance driver), 4 have been discharged, while **7** are still undergoing treatment, **3** in Masindi and **4** in Gulu.”*
- *“The overall death toll has risen to **160**, after **2 patients** in Gulu and **2** in Masindi died since Tue 5 Dec 2000. In total **7 new cases** in Gulu and none in Masindi have been confirmed since Tuesday, raising the overall number of **people** who have suffered from Ebola fever to **406**.”*

Appendix B

Examples of launch

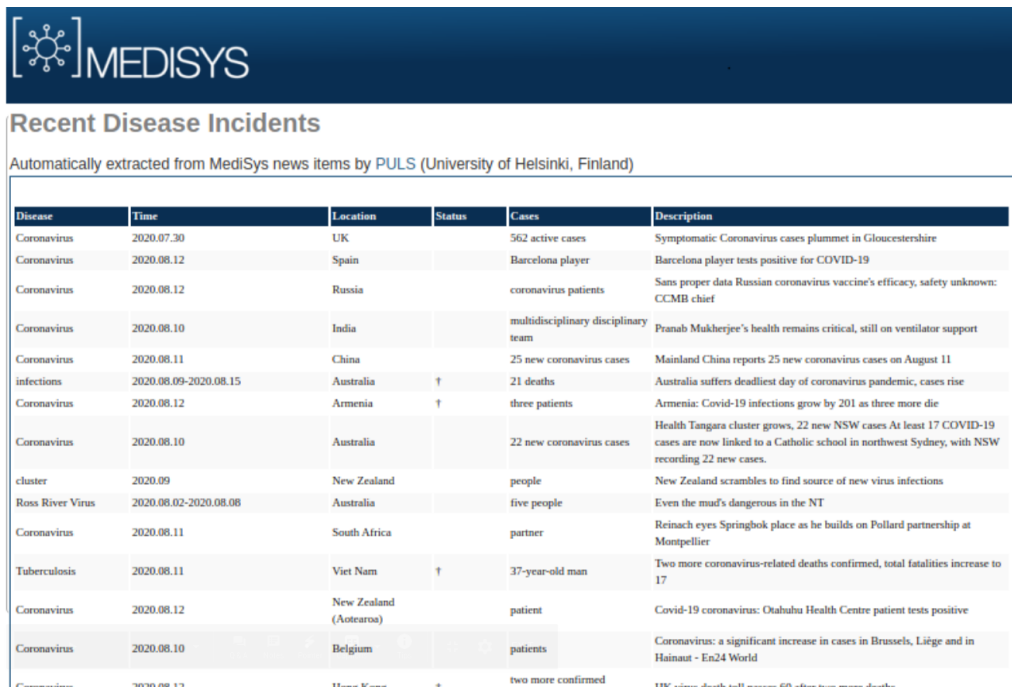
Table B.1: Examples of *launch* in Product Launches scenario, in three corpora

<i>Business Articles</i>	#	<i>General News</i>	#	<i>Abstracts</i>	#
campaign	216	campaign	1333	service	3499
<u>investigation</u>	183	<u>investigation</u>	1185	range	3263
service	114	<u>attack</u>	592	products	2359
website	86	<u>attacks</u>	517	product	2035
product	85	<u>appeal</u>	329	<u>campaign</u>	1874
program	82	bid	285	line	1749
<u>initiative</u>	82	<u>inquiry</u>	273	project	1423
fund	81	<u>operation</u>	255	version	1195
<u>attacks</u>	79	<u>project</u>	235	card	933
it	69	program	233	production	931
project	68	<u>probe</u>	217	services	921
products	65	programme	210	flights	888
<u>attack</u>	64	<u>offensive</u>	208	fund	786
series	60	service	200	system	783
<u>review</u>	60	series	193	application	767
range	58	<u>crackdown</u>	188	website	685
version	56	<u>initiative</u>	187	model	643
programme	54	<u>war</u>	182	<u>offer</u>	612
<u>action</u>	49	it	182	<u>tender</u>	576
<u>plan</u>	48	website	171	brand	556
<u>satellites</u>	45	<u>strike</u>	160	it	526
<u>offensive</u>	44	<u>drive</u>	159	programme	490
platform	42	<u>manhunt</u>	140	<u>projects</u>	479
services	41	<u>plan</u>	135	drink	472
consultation	41	<u>missiles</u>	131	site	406
scheme	37	<u>review</u>	127	models	403
round	37	<u>scheme</u>	118	fragrance	395
network	37	<u>missile</u>	110	series	394
<u>probe</u>	35	<u>rockets</u>	104	facility	389
site	34	<u>run</u>	102	cream	389
<u>inquiry</u>	30	<u>report</u>	101	collection	380
	2311		5699		19162

Appendix C

Real-world application of IE

Figure C.1: Infectious Disease Outbreaks scenario running live on the MediSys



MEDISYS

Recent Disease Incidents

Automatically extracted from MediSys news items by PULS (University of Helsinki, Finland)

Disease	Time	Location	Status	Cases	Description
Coronavirus	2020.07.30	UK		562 active cases	Symptomatic Coronavirus cases plummet in Gloucestershire
Coronavirus	2020.08.12	Spain		Barcelona player	Barcelona player tests positive for COVID-19
Coronavirus	2020.08.12	Russia		coronavirus patients	Sans proper data Russian coronavirus vaccine's efficacy, safety unknown: CCMB chief
Coronavirus	2020.08.10	India		multidisciplinary disciplinary team	Pranab Mukherjee's health remains critical, still on ventilator support
Coronavirus	2020.08.11	China		25 new coronavirus cases	Mainland China reports 25 new coronavirus cases on August 11
infections	2020.08.09-2020.08.15	Australia	†	21 deaths	Australia suffers deadliest day of coronavirus pandemic, cases rise
Coronavirus	2020.08.12	Armenia	†	three patients	Armenia: Covid-19 infections grow by 201 as three more die
Coronavirus	2020.08.10	Australia		22 new coronavirus cases	Health Tangara cluster grows, 22 new NSW cases At least 17 COVID-19 cases are now linked to a Catholic school in northwest Sydney, with NSW recording 22 new cases.
cluster	2020.09	New Zealand		people	New Zealand scrambles to find source of new virus infections
Ross River Virus	2020.08.02-2020.08.08	Australia		five people	Even the mud's dangerous in the NT
Coronavirus	2020.08.11	South Africa		partner	Reinach eyes Springbok place as he builds on Pollard partnership at Montpellier
Tuberculosis	2020.08.11	Viet Nam	†	37-year-old man	Two more coronavirus-related deaths confirmed, total fatalities increase to 17
Coronavirus	2020.08.12	New Zealand (Aotearoa)		patient	Covid-19 coronavirus: Otahuhu Health Centre patient tests positive
Coronavirus	2020.08.10	Belgium		patients	Coronavirus: a significant increase in cases in Brussels, Liège and in Hainaut - En24 World
Coronavirus	2020.08.12	France	+	two more confirmed	Two more coronavirus cases confirmed in France