Complexity of Event Structure in IE Scenarios

Silja Huttunen, Roman Yangarber, Ralph Grishman
Courant Institute of Mathematical Sciences
New York University
{silja,roman,grishman}@cs.nyu.edu

Abstract

This paper presents new Information Extraction scenarios which are linguistically and structurally more challenging than the traditional MUC scenarios. Traditional views on event structure and template design are not adequate for the more complex scenarios.

The focus of this paper is to show the complexity of the scenarios, and propose a way to recover the structure of the event. First we identify two structural factors that contribute to the complexity of scenarios: the scattering of events in text, and inclusion relationships between events. These factors cause difficulty in representing the facts in an unambiguous way. Then we propose a modular, hierarchical representation where the information is split in atomic units represented by templates, and where the inclusion relationships between the units are indicated by links. Lastly, we discuss how we may recover this representation from text, with the help of linguistic cues linking the events.

1 Introduction

Information Extraction (IE) is a technology used for locating and extracting specific pieces of information from texts. The knowledge bases are customized for each new topic or *scenario*, as defined by fill rules that state which facts are needed for constitution of an extractable *event.* A scenario is a set of predefined facts to be extracted from a large text corpus, such as news articles, and organized in output *templates.*

Our experience with customizing our IE system called Proteus (Grishman, 1997; Grishman et al., 2002) to new scenarios suggests that the lexical and structural properties of the scenario affect the performance of the system. To make an IE system flexible for tasks of varying complexity, it is essential to conduct a linguistic analysis of the texts relating to different scenarios.

In this paper, we focus on the Infectious Disease Outbreak scenario (Grishman et al., 2002), and the Natural Disaster scenario (Hirschman et al., 1999) collectively called the "Nature" scenarios. During the customization of the IE system to the Nature scenarios, we encountered problems that did not arise in the traditional scenarios of the Message Understanding Conferences (MUCs). This included, in particular, delimiting the scope of a single event and organizing the events into templates.

We identify two structural factors that contribute to the complexity of a scenario: first, the scattering of events in text, and second, inclusion relationships between events. These factors cause difficulty in representing the facts in an unambiguous way. We proposed that such event relationships can be described with a modular, hierarchical model (Huttunen et al., 2002).

The phenomenon of inclusion is widespread in the Nature scenarios, and the types of inclusions are numerous. In this paper we present preliminary results obtained from our corpus analysis, with a classification and distribution of inclusion relationships. We discuss the potential for recovery of these inclusions from text with the help of the linguistic cues, of which we show some examples.

This paper will argue that a thorough linguistic analysis of the corpus is needed to help recovery of the complex event structure in the text.

In the next section we give a brief description of the scenarios we are investigating. In section 3 we review the problems of scattering, inclusion and event definition, and propose a method for representing template structure. In section 4 we present examples of the linguistic cues to

Disaster	*Date*	*Location*	*VictimDead*	*Damage*
tornado	Sunday night	Georgia	one person	motel

Disease	*Date*	*Location*	*VictimDead*	*VictimSick*
Ebola	since September	Uganda	156 people	-

Table 1: Disaster Event and Disease Event

recover the complex event structure, followed by discussion in section 5.

2 Background

2.1 Information Extraction

Our IE system has been previously customized for several news topics, as part of the MUC program, such as Terrorist Attacks (MUC, 1991; MUC, 1992) and Management Succession (MUC, 1995; Grishman, 1995). Subsequently to the MUCs, we customized Proteus to extract, among other scenarios, Corporate Mergers and Acquisitions, Natural Disasters and Infectious Disease Outbreaks.

We contrasted the Nature scenarios with the earlier MUC scenarios (Huttunen et al., 2002). The "traditional" template structure is such that all the information about the main event can be presented within a single template. The main events form separate instances, and there are no links between them. Management Succession scenario presents a slightly more complicated template structure, but it is still possible to present in one template. The traditional representation is not adequate to represent the complex structure of the Nature scenarios.

In the next section, we give a short description of the Nature scenarios.

2.2 Scenarios

For the **Natural Disaster** scenario, the task is to find occurrences of disasters (earthquakes, storms, etc.) around the world, as reported in newspaper articles. The information extracted for each disaster should include the type of disaster, date and location of the occurrence, and the amount of human or material damage.

An example of a Natural Disaster template is in table 1, extracted from the following news fragment:

> "[...] tornadoes that destroyed a Georgia motel and killed one person in a mobile home Sunday night."

For the **Infectious Disease Outbreak** scenario, the task is to track the spread of epidemics of infectious diseases around the world. The system has to find the name of the disease, the time and location of the outbreak, the number of victims (infected and dead), and type of victims (e.g., human or animal). The next example is a fragment of a disease outbreak report, and the extracted facts are shown in table 1.

> "Ebola fever has killed 156 people, [...], in Uganda since September."

3 Structure of Events

The complex event structure in Nature scenarios is partly due to the fact that the events are reported in a scattered manner in the text.

By scattering of events we mean that their components are not close to each other in the text, and a typical text contains several related events. This is partly because the articles are often in a form of an update, where the latest reported damages contribute to the total damages reported earlier, over several locations and over different time spans.

The example in table 2 illustrates scattering in the Disease scenario. It is a fragment of an update about a cholera epidemic in Sudan, from the World Health Organization's (WHO) web report. The locations are highlighted in italics and the victim counts are in boldface, to show the scattering. In this example there are six separate *mentions*—partial descriptions of the event in text—giving the number of infected and dead victims, in Sudan, and in two locations within Sudan. Paragraph (1) reports the number of victims in Sudan, 2549 infected, and 186 dead. In paragraph (2), the focus is shifted to another location in Sudan, and new numbers are reported. Paragraph (3) gives the respective

(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**), [...]

Table 2: Example of a Disease Outbreak Report

Disease	*Location*	*Infected*	*Dead*
Meningococcal	Sudan	2549	186
	Bahar aj Jabal State	1437	99
	White Nile	197	15

Table 3: Facts from Disease Outbreak Report

numbers for yet another location in Sudan. The mentions are summarized in table 3.

3.1 Inclusion Relationships

As we frequently observe in the Nature scenarios, the information in the various mentions in table 2 is overlapping, and the mentions partially include each other.

For example, the numbers for infected victims in paragraph (2) and (3), contribute to the total number of infected cases in paragraph (1). The extraction system should be able to extract all the numbers for this text. The problem is how to group these mentions into a template in an unambiguous and coherent way. It is impossible to represent an event with overlapping information in a single template, since it consists of multiple numbers of victims in several areas and several time intervals.

For the purpose of handling this phenomenon, we first introduce a distinction between *outbreaks* and *incidents*. An incident is a short description, or a mention, of one occurrence that relates to an outbreak. It covers a single specific span of time in a single specific area. An outbreak takes place over a longer period of time, and possibly over wider geographical area: it consists of multiple incidents.

In general, one incident may *include* others, which give further detailed information.

Therefore, we analyze the news fragment in table 2 as containing six incidents, with two types of inclusions: first, inclusion by status, where the dead count contributes to the infected count of the same area, and second, inclusion by location, where the numbers of infected cases in Bahar aj Jabal State, in paragraph (2), and in White Nile, in (3), contribute to the infected count in Sudan, in paragraph (1).

The Natural Disaster scenario poses further complications for this schema. The scattering is complicated by the relationship of *causation*: the main disaster triggers derivative disasters (sub-disasters), which in turn may cause damages that contribute to the overall damage. This is illustrated by the news fragment in table 4, from the New York Times. Names of disasters are in bold, and the damages are italicized.

In table 4, paragraph (1), a disaster includes rain and winds, which cause flooding. In paragraph (3), the human damages caused by snow are included in the total human damages caused by the storm in (2). The derivative disasters and their damages often take place in several locations, appearing relatively far in the text from the first mention of the main disaster. The final logical representation of the event should be such that the effects of the sub-disasters could be traced back to the main event.

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

(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. [...]

Table 4: Example of Disaster Reporting

the infected count, as in paragraph (2) of table 2.

- 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.

3.2 Type and Distribution of Inclusions

To investigate the extent of inclusions and their distribution by type, we analyzed 40 documents related to Nature scenarios.[1]

To confirm the feasibility and applicability of this approach, we manually tagged the inclusion relationships present in these documents. Table 5 shows the number of incidents found in the documents, as well as the number and the types of inclusion. There are also multiple inclusions: e.g., infected *health workers* in a town in Uganda are included in the total number of infected *people* in the whole country: this is inclusion by both *case-descriptor* and *location.*

Multiple inheritance also occurs: in table 2, 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 5, we show only the inclusion in the immediately preceding parent.

Scenario	*Disease*	*Disaster*
Documents	26	14
Words	9 500	6500
Incidents	125	112
Inclusions	57	81
time	6	6
location	19	20
status	19	1
case-descriptor	6	1
case-desc/location	3	–
disease	1	–
causation	–	19
causation/location	–	11
causation/time	–	3
time/location	–	7
disaster	–	5
disaster/location	–	2
damage	–	4
others	3	2

Table 5: Type and Number of Inclusion

3.3 Hierarchical Template Structure

Our proposed solution is to have a separate template for each incident. Once we have broken down the information into smaller incident templates, the inclusion relationship between them is indicated by *event pointers.* This approach makes it possible to represent the information in a natural and intuitive way.

The final template for the Infectious Disease scenario is shown in table 6. Note that there is a separate slot indicating the parent incident.

Disease Name
Date
Location
Victim Number
Victim Descriptor
Victim Status
Victim Type
Parent Event

Table 6: Infectious Disease Template

[1] The 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 Disease Outbreaks, a total of 26 documents from NYT, Promed, WHO, and ABC.

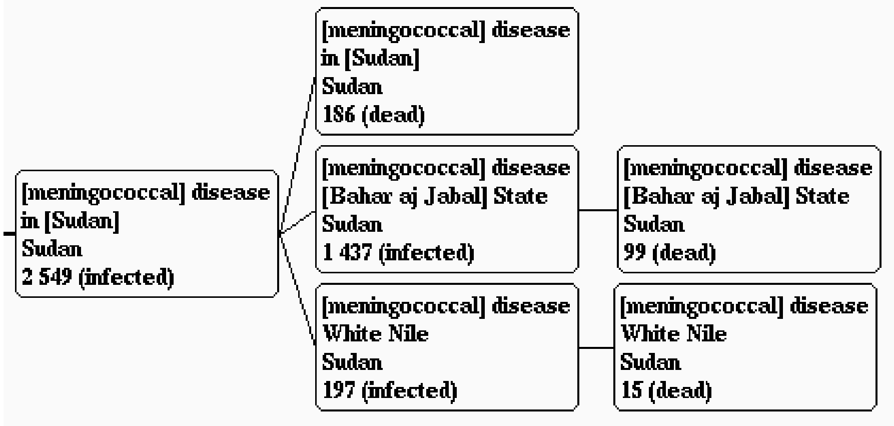

Figure 1: Infectious Disease Outbreak

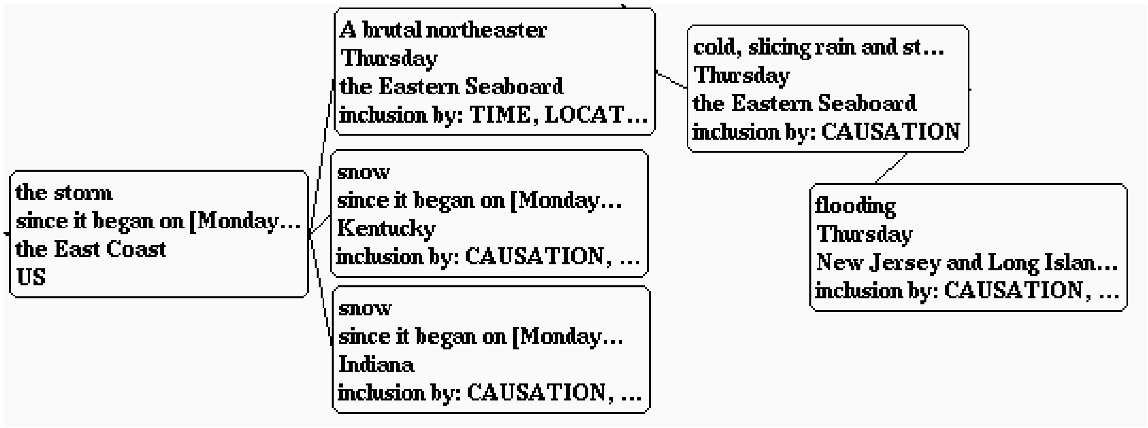
the storm
since it began on [Monday...
the East Coast
US
A brutal northeaster
Thursday
the Eastern Seaboard
inclusion by: TIME, LOCAT...
snow
since it began on [Monday...
Kentucky
inclusion by: CAUSATION, ...
snow
since it began on [Monday...
Indiana
inclusion by: CAUSATION, ...
cold, slicing rain and st...
Thursday
the Eastern Seaboard
inclusion by: CAUSATION
flooding
Thursday
New Jersey and Long Islan...
inclusion by: CAUSATION, ...

Figure 2: Natural Disaster

Figure 1 is a graphical representation of the inclusion relationships among the incidents extracted from the Disease report in table 2. The figure shows the main incident with several sub-incidents. Two of the sub-incidents have, in turn, sub-incidents. The types of inclusions are shown in the last row.

Figure 2 shows a graphical representation of inclusion by causation in Natural Disaster scenario. The incidents are extracted from table 4.[2] There is a causation relationship between the incidents. It is important to recover the long causation chains from the text.

As a result, the templates are simple, but there are typically many templates per document. The separation of incidents affects the process of extraction, since we can now focus on looking for smaller atomic pieces first. Then we must address the problem of linking together related incidents as a separate problem in the overall process of IE.

4 Linguistic cues

The process of tracking the inclusion relationships between the incidents is not trivial. A human reader uses the cohesive devices in the text to construct the connections between parts of text (see e.g., (Halliday and Hasan, 1976; Halliday, 1985)). Finding the relationship between incidents may be a less complex task than tracking cohesion through an entire text or discourse. Our task is limited to finding the cohesive de-

[2] 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.

vices connecting a small set of pre-defined facts, that may occur nearby within one sentence, or are separated by one or more sentence boundaries. Our goal is to locate the cues in the text, and use them to automatically recover these relationships.

An example of a linguistic cue is in the following fragment of an update from table 4:

> **Elsewhere** along the East Coast, 19 deaths have been attributed to the storm [...]

Elsewhere indicates a shift in the focus from one location to another and there is probably no inclusion between the following and immediately preceding mention of the damages.

We have identified several linguistic cues that signal the presence or absence of an inclusion relationship between two incidents. These cues can be one of following types:

- Specific *lexical items*, which can be e.g., adverbs, verbs, prepositions, connectives. *Elsewhere* in the previous example implies that damages caused by the following disaster do not contribute to the damages of the immediately preceding disaster.
- Two expressions in separate incidents which are related in the *scenario-specific concept hierarchy*, may indicate the presence and also the direction of an inclusion, e.g., *health worker* is included in *people*; names of plants, animals and terms referring to human beings, are hyponyms of *victim*.
- Locative or temporal expressions that are in a hierarchical relationship in a location hierarchy or in the implicit time hierarchy, often indicate presence or direction of inclusion.
- Elliptical elements create cohesion. Ellipsis indicates the presence of a parent incident earlier in the text. In paragraph (3) of table 4, in the parent incident we observe a case descriptor, *deaths*, which is elided in the two sub-incidents.
- Anaphora: anaphoric reference usually indicates the absence of an inclusion between two incidents, merging into one. For example, in table 4, paragraph (3), *the 19 deaths* is coreferential with *19 deaths* caused by the storm in paragraph (2).
- Coordination tends to indicate the absence of inclusion relationship. For example, when two incidents are conjoined by *and* and do not share information about location or time, there is typically no inclusion. However, there are cases where other cues override this general tendency.

These cues often do not appear in isolation, and they may interact.

We give an example of three lexical items and their role as an indicator of inclusion in the Infectious Disease Outbreak Scenario. Consider the preposition *with*[3], the participle *including* and the finite verb *include*.

> “More than 500 cases of dengue hemorrhagic fever were reported in Mexico last year, with 30 deaths, Ruiz said.”

The 30 deaths are included in the 500 cases. The direction of the inclusion is reversed in the following example:

> ”Disease has killed 10 persons, with 242 cases having already been reported.”

The latter incident includes the former. Here additional cues are provided by the concept hierarchy, and the numbers: a smaller number cannot include a larger one.

The following illustrates the participle *including* as cue:

> Ebola fever has killed 156 people, *including* 14 health workers, in Uganda since September.

The incidents are connected by *including*, which also indicates the direction explicitly. Additional information is obtained from the case-descriptors, related in the concept hierarchy.

The context for such “trigger” words as they indicate inclusion, is that the trigger appears between two incidents, preceding and preceded

[3] In the case of *with* we look only at free prepositions, that is, those not bound to a preceding verb (Biber et al., 1999).

by a quantified NP[4] and optional phrases or items from the concept hierarchy.

> Q {case-descriptor | status} [reported | get sick | time | location | disease] [,] *trigger* Q {case-descriptor | status}

These triggers can indicate inclusion also inside a parenthetical phrase, preceding a quantified NP, as in table 2 in paragraph (2).

The trigger *include* (as a finite verb) functions similarly, but can also occur between sentences:

> [...] the Ugandan Ministry of Health has reported [...] 370 cases and 140 deaths. This figure *includes* 16 new confirmed cases in Gulu [...]

In our training corpus, when these cue words occurred in this context, they consistently indicated an event inclusion relation.

5 Discussion

Complexity of a scenario seems to depend of multiple factors. The notion of complexity, however, has not been investigated in great depth. Some research on this was done by (Bagga and Biermann, 1997; Bagga, 1997), classifying scenarios according to difficulty by counting distances between “components” of an event in the text. In this way it attempts to account for variation in performance across the MUC scenarios.

Our analysis suggests that the type and amount of inclusion relationships depend on the nature of the topic. In such scenarios as Management Succession and Corporate Acquisitions, an event usually occurs *at one specific point in time.* By contrast, the Nature events typically take place *across a span of time and space.* As the event “travels” and evolves, its manifestations are reported in a piecewise fashion, sometimes on an hour-by-hour basis.

An extensive linguistic analysis of the corpus is necessary to resolve these complex issues. For evaluation and training, we are building test and training corpora, totaling 70 documents and annotated with inclusion relationships.

[4] Here the case descriptor or status can be elided: however, one of quantifiers should have a case descriptor or a status.

Acknowledgments

This research is supported by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center San Diego, and by the National Science Foundation under Grant IIS-0081962.

This paper does not necessarily reflect the position or the policy of the U.S. Government.

References

A. Bagga and A. W. Biermann. 1997. Analyzing the complexity of a domain with respect to an information extraction task. In *Proc. 10th Intl. Conf. on Research on Computational Linguistics (ROCLING X).*

A. Bagga. 1997. Analyzing the performance of message understanding systems. In *Proc. Natural Language Processing Pacific Rim Symposium (NLPRS'97).*

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English.* Longman.

R. Grishman, S. Huttunen, and R. Yangarber. 2002. Real-time event extraction for infectious disease outbreaks. In *Proc. HLT 2002: Human Language Technology Conf.*, San Diego, CA.

R. Grishman. 1995. The NYU system for MUC-6, or where's the syntax? In *Proc. 6th Message Understanding Conf. (MUC-6)*, Columbia, MD. Morgan Kaufmann.

R. Grishman. 1997. Information extraction: Techniques and challenges. In M. T. Pazienza, editor, *Information Extraction.* Springer-Verlag, Lecture Notes in Artificial Intelligence, Rome.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English.* Longman, London.

M.A.K. Halliday. 1985. *Introduction to Functional Grammar.* Edward Arnold, London.

L. Hirschman, E. Brown, N. Chinchor, A. Douthat, L. Ferro, R. Grishman, P. Robinson, and B. Sundheim. 1999. Event99: A proposed event indexing task for broadcast news. In *Proc. DARPA Broadcast News Workshop*, Herndon, VA.

S. Huttunen, R. Yangarber, and R. Grishman. 2002. Diversity of scenarios in information extraction. In *Proc. 3rd Intl. Conf. of Language Resources and Evaluation, LREC-2002*, Las Palmas de Gran Canaria, Spain.

1991. *Proc. 3th Understanding Conf. (MUC-3).* Morgan Kaufmann.

1992. *Proc. 4th Message Understanding Conf. (MUC-4).* Morgan Kaufmann.

1995. *Proc. 6th Message Understanding Conf. (MUC-6).* Morgan Kaufmann.