Diversity of Scenarios in Information Extraction

Silja Huttunen, Roman Yangarber, Ralph Grishman

Courant Institute of Mathematical Sciences
New York University
{silja,roman,grishman}@cs.nyu.edu

Abstract

This paper discusses/presents problems of template structure for Information Extraction. We investigate these problems in the context of two new Information Extraction scenarios which are linguistically and structurally more challenging than the traditional MUC scenarios. By a scenario we mean a predefined set of facts to be extracted from text. Traditional views on event structure and template design are not adequate for the more complex scenarios.

We identify two structural factors that contribute to the complexity of a scenario: first, the scattering of events in text, and second, inclusion relationship between events. These factors cause difficulty in representing the facts in an unambiguous way. Traditional views on event structure and template design are not adequate for the more complex scenarios. We propose that these kinds of event relationships can be better described with a modular, hierarchical model.

1. Introduction

Our experience with customizing our Information Extraction system (Grishman, 1997) to new tasks suggests that evaluation is strongly affected by the lexical and structural properties of the scenario.

Evaluation of performance is a central step in the process of customization. We employ the standard MUC-style evaluation strategy and tools. Weaker performance of an Information Extraction system on one scenario, as compared to another, may hide more fundamental problems than only a weaker knowledge base, i.e., a less complete lexicon or semantic patterns.

In this paper, we focus in particular on the Infectious Disease Outbreak scenario (Grishman et al., 2002), and the Natural Disaster scenario—we shall refer to these collectively as the "Nature" scenarios. During the customization of the Information Extraction system to the Nature scenarios, we encountered problems with definition of facts that did not arise in the traditional scenarios of the Message Understanding Conferences (MUCs). In particular, we encountered difficulties with delimiting the scope of a single event, and with organizing the events into templates.

This paper will argue that the traditional views on event structure and template design are not adequate for some of the new, richer scenarios.

We can identify two structural factors that impact on the performance of our Information Extraction system. First, the events, or components of events, seem to be much more widely spread in text, or *scattered*, in the Nature scenarios than in the MUC tasks. Second, they interlock and relate to each other in complex ways, forming *inclusion* relationships.

These relationships cause difficulty in representing the facts in an unambiguous way, raising important questions about how best to organize the extracted facts into templates, and how to define the structure and the extent of a fact.

In the next section we give the background, a brief description of the scenarios we are investigating, and an example of a traditional MUC-style template. In section 3. we present the problems of scattering and event definition; in section 4. inclusion relationships and our proposed method for representing template structure. In section 5. we discuss issues related to the corpus analysis.

2. Background

2.1. Information Extraction

Information Extraction (IE) is a technology used for locating and extracting specific pieces of information, or *facts*, from texts. IE systems are commonly based on pattern matching, using sets of patterns of increasing linguistic complexity (Appelt et al., 1995; Yangarber and Grishman, 1998). Each pattern is a regular expression, matching a syntactically and semantically typical construction that states a sought fact. The patterns are customized for each new topic or *scenario*, as defined by fill rules. The fill rules specify which pieces of information constitute an extractable *event*, or fact. The facts are extracted from a large text corpus, such as news articles, and organized in output *templates*.

Our IE system, called Proteus, has been previously customized for several news topics, as part of the MUC program, such as Management Successions (MUC, 1995; Grishman, 1995). Subsequently to the MUCs, we customized Proteus to extract Corporate Mergers and Acquisitions, Natural Disasters and Infectious Disease Outbreaks (Grishman et al., 2002), among other scenarios.

In this study, we contrast earlier MUC scenarios, Terrorist Attacks and Management Succession, with the Nature scenarios, Natural Disasters and Infectious Disease Outbreaks. In the next sections, we give a short description of each of the scenarios.

2.2. Description of Scenarios

For the topic of **Terrorist Attacks**, investigated under MUC-3 and MUC-4, (MUC, 1991; MUC, 1992), the system has to find terrorist incidents, and for each incident extract the place and time of the attack, a description of the victims, the responsible party, and the weapon used in the attack. The following sample text contains a description of a terror event. The corresponding filled template is in table 1.

Time	*Location*	*Victim*	*Terrorist*	*Weapon*
yesterday	San Salvador	6 Jesuit priests	unknown	car bomb

Table 1: Example of "Terrorist Attack" Template

Person	*Company*	*Post*	*Date*	*Status*
P.I. Wilber	Drug Emporium	chairman	April	out

Table 2: Example of "Management Succession" Template

> "Six Jesuit priests were killed when a car bomb exploded near San Salvador yesterday [...]."

Note that all the templates shown here are simplified versions of the actual templates. We show only the central slots, omitting the rest for the sake of clarity. The MUC-4 template, for example, had 25 slots.

For the topic of **Management Succession**, the system has to find reports about top-level corporate positions changing hands. For each succession report, the system must determine the name of the company, the post, and the name of the person who is leaving or assuming the post [1]. A short example of a Management Succession template is in table 2, with the fact extracted from the following fragment of text:

> "In April, Drug Emporium retired its founding chairman, Philip I. Wilber [...]."

The original training and test corpus, for customizing and evaluating the performance of the system, consisted of 200 articles from the Wall Street Journal (WSJ).

We had explored the **Natural Disaster** scenario as part of the Event-99 project (Hirschman et al., 1999). The IE task was to find occurrences of disasters (e.g., earthquake, storm, flood, etc.) around the world, as reported in newspaper articles. The information extracted for each disaster should include: the type of disaster, when and where it occurred, how much material damage it caused, and how many people were killed or injured. An example of a Natural Disaster template is in table 3, extracted from the following news fragment:

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

Our corpus for Natural Disasters consisted of transcribed broadcast news reports from ABC, CNN, PRI, Voice of America (VOA) and written news from the Associated Press (AP) and the New York Times (NYT), a total of 54 documents.

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 affected victims (infected and dead), and type of victims (e.g., human or animal). The following example is a fragment of a disease outbreak report, and the corresponding template in table 4.

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

For customization, we used a corpus consisting of articles from ProMed, a mailing list where medical professionals in the field contribute updates on epidemics and outbreaks of diseases around the world, the World Health Organization (WHO) web reports of disease epidemics, and articles from the New York Times (totaling about 140 documents).

2.3. Template Structure: Management Succession

We now focus on the problem of template structure. Consider the Terrorist Attacks scenario. The template structure is "flat", in the sense that all the information about a given event can be presented within a single template, in one row of the table. The events form separate instances, they are not linked in any way, so that the flat representation is adequate for this scenario.

Management Succession scenario presents a slightly more complicated template structure. The next example is an fragment of a news report from the Wall Street Journal:

> "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."

The post of chairman of Drug Emporium was vacated by Mr. Wilber last April, and the same post was assumed by Mr. Kriegle last week, and this is a direct succession. The crucial observation is that if each of these facts were represented in a separate template (or row of the table), the connection between them, i.e., that one officer succeeds the other and there was no one in between, may be lost.

For the Management Succession scenario (MUC-6), the extracted facts are represented in an "object-oriented" view. The template is an object consisting of slots, whose value can be a string, number, or an object (another list of slots with values). A Succession event template is shown in the following schema, with three slots, which are the **post**, the **company**, and the **changes** in the post. The **changes** slot may contain one or more *transition* objects, each of which

[1] There is other information relating to the change of post that the system must extract; for finer details, refer to (MUC, 1995)

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

Table 3: Example of "Natural Disaster" Template

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

Table 4: Infectious Disease Outbreak Template

contains two pieces of information: who is the **officer** involved in the transition, and whether the person is leaving or assuming the post (**status** is either "in" or "out"). Each person is described by an object containing the name, etc.

post:
company:
changes: *transition*→**status**: {in, out}
officer: *person:*→ **name:**
title:
transition-2 ...
...

Figure 1: Object-oriented template for MUC-6

Table 5 shows five facts extracted from the above news report, and transferred from the object-oriented view into a table. Note that when the information is transferred from the object-oriented view to the table view, we don't want to lose the explicit links between the transitions in a single *succession*. To preserve the link in the table view, we add a column for co-indexing records belonging to the same succession—the rightmost column of the table.

3. Scattering of Facts in Texts

The Nature scenarios pose still deeper problems for this kind of template representation. In this section we explore the complex event structure we encounter in these scenarios.

We first observe the overall structure of a news article and the *scattering* of events. By scattering we mean that the components of the sought events do not appear together near each other in the text, and a typical article may contain several related events. This is due in part to the fact that the articles are often in a form of update reports, where the latest damages are reported and added to the total amounts of damages that occurred earlier and had been reported in detail previously.

The text in figure 2 is a segment of an update about an outbreak of Ebola in Uganda, from ProMed. The locations are highlighted in italics and the numbers of victims in boldface. There are multiple facts in this report, and they are scattered widely in the text. In this example there are ten separate *mentions*—partial descriptions of the event in text—describing victims infected with Ebola. The mentions relate to two locations, and several time intervals.

The problem is how to group these mentions into a template in a coherent way. Paragraph (1) reports the number of deaths among health workers for this update, 15, which contributes to a total number, 26, in paragraph (2). The total is further divided into numbers of victims that are dead[2] or sick (*undergoing treatment*)[3]; the sick victims are further divided by their location. In other words, the total number of 26 victims is further analyzed by status (dead, sick or infected) and location. The information from the mentions from paragraph (2) is summarized in table 6.

The mapping of the locations and victims in this example is complex, yet it is quite typical for descriptions of disease outbreaks.

Disease	*Location*	*Victims*
Ebola	Gulu and Masindi	26 infected
	Gulu and Masindi	15 dead
	Gulu and Masindi	7 sick
	Masindi	3 sick
	Gulu	4 sick

Table 6: Facts from Paragraph (2)

It is clear that this list of facts does not reflect the information in the text, since the structure and relationship between these numbers is lost.

4. Toward a Solution

4.1. Inclusion Relationships

In the example in figure 2 we observe that the information in the various mentions is overlapping, and the mentions partially include each other. This phenomenon is common in the Nature scenarios.

For example, in paragraph (3) of figure 2, the 2 lives lost in Masindi plus the 2 in Gulu since Tue 5 Dec 2000 contribute to the total of 160 for the entire outbreak. It is important that the extraction system be able to extract *all* numbers from this paragraph—the total for the year, as well as the new numbers for this update—because these mentions contain complementary information. It becomes difficult to represent these types of events within a single template or

[2]Note that this is the same as the number for the update, repeated from paragraph (1).

[3]In this task, we do not attempt to extract information about patients who have recovered.

Person	*Company*	*Post*	*Date*	*Status*	*Link*
P.I. Wilber	Drug Emporium	chairman	April	out	♣
P.I. Wilber	Drug Emporium	CE	April	out	♢
G. Wilber	Drug Emporium	vice chairman	April	in	
D.L. Kriegle	Drug Emporium	chairman	last week	in	♣
D.L. Kriegle	Drug Emporium	CE	last week	in	♢

Table 5: Example of "Management Succession"

(1) 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**.
(2) 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 , 4 have been discharged, while **7** are still undergoing treatment, **3** in *Masindi* and **4** in *Gulu*.
(3) The overall death toll has risen to **160**, after **2 patients** in *Gulu* and **2** in *Masindi* died since Tue 5 Dec 2000.
(4) 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**.

Figure 2: Example of a Disease Outbreak Report

a single row in a table, since they consists of multiple numbers of victims in several areas, over different time spans.

We make a distinction between *outbreaks* and *incidents* for the purpose of handling this phenomenon. An outbreak consists of several reported atomic units, or incidents. An incident is a short description, or a mention, of a specific occurrence that relates to an outbreak. It covers a single specific span of time in a single specific area. An outbreak, rather, takes place over a longer period of time, and possibly over wider geographical area. An outbreak is made up of incidents.

Note, that in general one incident may *include* others, which give further detailed information.

Example: We analyze paragraph (3) to contain three incidents: one covering the overall total of 160 victims, one covering the 2 patients in Gulu, and for the 2 patients in Masindi. The two new incidents, containing two victims each, are included in the total number of cases, 160. The relationship between the parent incident (the overall death toll 160) to the two sub-incidents (2 in Gulu, 2 in Masindi) is that of locative *inclusion*.

In this scenario we observe the following inclusion relationships:

- location
- time: e.g., total for this update included in total for the year
- status: dead or sick cases included in the number of infected cases, as in paragraph (2) of figure 2.
- victim type or descriptor: e.g., "people" includes "health workers", "women", "men"

Figure 3 shows a graphical representation of the inclusion relationships among the incidents in paragraph (2). At the top of the tree there is the total number of 26 cases, including the 15 dead victims, as well as the seven sick who are in treatment. There is an inclusion by status between the parent incident and the two sub-incidents. Further, there are three sick persons in Masindi and four in Gulu that contribute to the total of seven sick; these are inclusions by location.

Due to this complexity it becomes unclear how all of this information can be consistently represented within a single event template. We propose to split up the description of the outbreak into incidents, which makes it possible to represent the information in a natural and intuitive way.

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, but that is outside the scope of this paper.

4.2. Hierarchical Template Structure

Our proposed solution is to make a separate template for each incident. Figure 3 shows the hierarchical relationships among the incidents in paragraph (2) of figure 2. It shows one incident with several sub-incidents, and there are two types of inclusions. Each extracted template represents an atomic event or sub-event, which is not further analyzable.

Once we have broken down the information into the smaller incident templates, the inclusion relationship between them is indicated by *event pointers*, or *hierarchical event links*. These links capture the relationship between the incidents and their sub-incidents.

The final template for the Infectious Disease scenario then is shown in table 8.

The template slot *Case Number* contains the number of victims for the given incident. *Case Descriptor* describes the victims, and *Case Status* indicates whether the victims are infected, sick or dead. Note that not all incidents must contain a number of victims: sometimes an outbreak is mentioned in a location, but no descriptions or numbers of victims are given. Such a mention still creates an incident template.

As a result, we have a simpler template, but typically many templates per document.

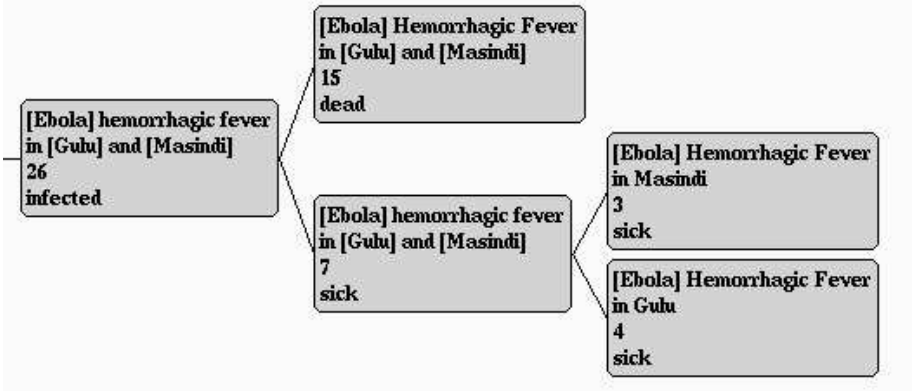

Figure 3: Disease Template

Disaster	*Date*	*Location*	*NumberDead*	*ID*
northeaster/ rain and winds	Thursday	the Eastern Seaboard	–	1
flooding	–	N.J. and L.I.	–	2
storm	since Monday	East Coast	19	3
tornado	–	Southern Florida	–	4
rain and winds	–	Georgia to New Jersey	–	5
snow	–	Kentucky	5	6
snow	–	Indiana	2	7
snow	–	Princeton, W.Va	2	8

Table 7: Natural Disasters

Disease Name
Date
Location
Case Descriptor: (people, cows, ...)
Case Number
Case Status: (infected, sick, dead)
Victim Type: (human, animal, plant)
Hierarchical Event Link

Table 8: Template of Disease Incident

4.3. Causation Relationships

The Natural Disaster scenario poses further complications. One disaster may have several manifestations and consequences, which can themselves be considered disasters, e.g, a storm may cause mud slides and flooding. The phenomenon of scattering is complicated by the additional relationship of *causation*: the main disaster triggers derivative disasters (sub-disasters), which in turn may cause damages, e.g., harm to humans and destruction of property. This is illustrated in figure 4, a fragment of a news report from the New York Times. Names of disasters are in bold, and the damages they caused are italicized.

In figure 4, paragraph (1), a disaster includes rain and winds, which cause flooding. Though the flooding could be considered as damage, we instead treat it as a derivative sub-disaster since it could cause damages of its own. The damages of the sub-disasters contribute to the overall damage of the main disaster.

Hence, we add one more inclusion relationship to our list: inclusion by causation.

Table 7 lists the separate mentions extractable from the news report. The relationships between the incidents and the weather conditions are unclear from this representation. To make this representation useful, it is necessary to reflect the relationships as they are expressed in the text.

The purpose of the analysis is to achieve a final logical representation of the event, where the effects of the sub-disasters are traceable back to the main event. This is easily done if the sub-events are linked to their parent events.

The derivative disasters and their damages often take place in several locations, and so may appear relatively far in the text from the mention of the main disaster. Consider the snow that caused accidents in paragraph (4). The chain of causation is quite long, and the victims are given at the end of the chain (e.g. traffic accidents or roof collapses in paragraph (4)). Under our representation model the relationship is indicated by the chain of links (though it may be difficult in practice to link all of the lowest-level damages to the top-level disaster).

Figure 5 shows the graphical presentation of the incidents in this text, linked by several inclusion relationships: by causation, by time, and by location.

Note, that cumulative events since Monday appear higher in the tree, but are mentioned later in the text. It is clearly a non-trivial practical task for the Information Extraction system to establish the correct linkage.

(1) 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.
(2) 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**, *power failures* and temporary *road closings* in coastal areas. No injuries or deaths were reported.
(3) Elsewhere along the East Coast, *19 deaths* have been attributed to the **storm** since it began on Monday.
(4) *The 19 deaths* include *five* in accidents on **snow**y 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.

Figure 4: Example of Disaster Report

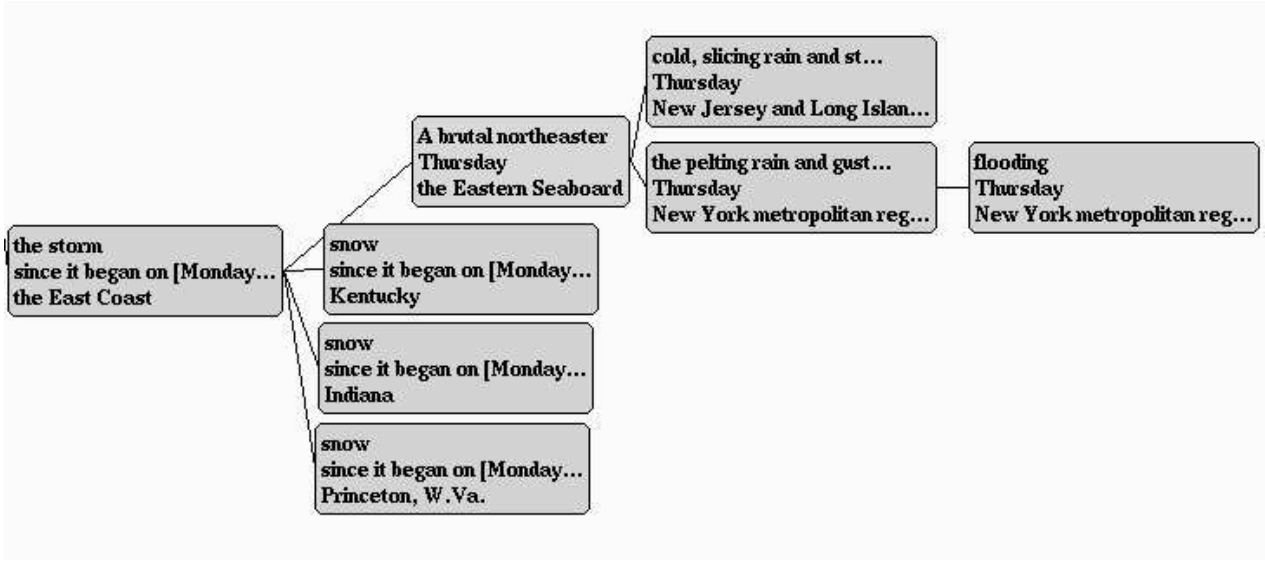

Figure 5: Disaster Template

To conclude the discussion of inclusion relationships between incidents, we note that different scenarios exhibit different types of inclusions, and in new scenarios we may encounter still other types of relationships.

5. Discussion

Complexity of a scenario seems to depend on multiple factors. The notion of the complexity of scenarios, however, has not been investigated in great depth in the context of Information Extraction. Some research on this was done by (Bagga and Biermann, 1997; Bagga, 1997), who tried to classify scenarios according to difficulty by counting distances between certain elements of an event in the text. In this way it attempted to account for variation in performance across several MUC scenarios.

Our approach to assessing complexity is somewhat different. In analyzing the structure of an event, we try to understand how its component sub-events are linked together to form the complete picture of the event.

The Management Succession scenario does not exhibit complex links between events. Nor do the hiring/firing events tend to be reported in a fragmentary fashion, where the complete picture is built up from the pieces[4].

[4]This applies only to top-level managers. In cases of mass layoffs, there may indeed begin to emerge a structure similar to those of the Nature scenarios.

In MUC-6, the template representation is centered on a post within a company. The only type of link, then, among the extracted records is the connection between multiple transitions relating to the same post—i.e., when a succession is reported, there is an explicit link between the events of vacating and occupying the post.

Beyond this, in the MUC scenarios, the need to extract links between events did not arise.

The main difference between the earlier hierarchical models, such as the one for MUC-6, in figure 1, and what we propose is that we have a hierarchy of events, not just participants in events.

We have analyzed a training corpus consisting of documents for each of the Nature scenarios, to confirm the feasibility and applicability of this approach, as shown in Table 9. We used the same training corpus for analysis of relationships among incidents, as well as for customization of the IE system and evaluation of performance.

Scenario	*Diseases*	*Disasters*
Documents	18	14
Words	5000	6500
Events	82	115
Inclusions	49	76

Table 9: Number of Inclusions in Nature Scenarios

For the Natural Disaster Scenario we analyzed 14 articles from the New York Times, ABC, AP, CNN, and Voice Of America. For Disease Outbreaks, the 18 documents were mostly from the NYT, ProMed, WHO reports, and ABC.

We manually tagged the inclusion relationships present in the 14 documents about Natural Disasters and the 18 documents about Infectious Disease Outbreaks. The table shows the number of event templates (incidents) extracted from these documents and the number of inclusion relationships that were found.

This analysis is part to a wider study of structural and linguistic phenomena of events in texts.

Our finding that some scenarios exhibit more complex structure than others raises two questions: what features of the text account for the variation in complexity, and how are the inclusion relationships indicated in the text. We do not attempt to address these questions in this paper, but include a few notes to indicate directions for future research.

Our initial experience suggests that the structuring of these reports does not necessarily depend on the source of the documents. The purpose or function of the text may have an impact on its structural organization. E.g., if the function of the natural disaster report were to inform airplane pilots of severe weather conditions in a certain area, different linguistic expressions might be used: the focus of the text would not be on the damages and on the description of individual episodes.

In general, it may be that while the foreground facts are scattered, the background information is presented in a more compact form, since it is there for explanatory purposes.

The scattering seems to be partly due to editorial structure. 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 exist between them. In the Nature scenarios, cause and effect are often in separate clauses, sentences or even paragraphs. In the Business scenarios, the components of the fact tend to group much more compactly.

Our analysis suggests that the type and amount of inclusion relationships depend on the topic and the style of reporting. The facts seem to be presented in a different manner in different scenarios: different grammatical, lexical and discoursive means used as cohesive devices to tie together the pieces of information into facts. First, in the Business scenarios like Management Succession and Corporate Acquisitions, an event usually occurs *at one specific point in time*. By contrast, in the Nature scenarios, events are not punctual, but rather 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. Further, the definition of a fact is complicated in Natural Disasters 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.

The hierarchical model of template structure proposed here assigns a separate template instance to each textual mention of a (possibly partial) description of an event, and establishes hierarchical links from the smaller sub-parts up to their containing events (or sub-events). We believe this is an appropriate way to begin analyzing and addressing the problem of complexity of IE scenarios. Since we are operating with strictly defined pieces of information reduced to templates with relatively simple semantics, our task is very concrete; how the facts are represented in a text, and how they are connected to each other.

The process of reconstructing the complete logical picture from the logical fragments is not trivial. The human reader picks up linguistic cues in the text which indicate the form of the report. E.g., in a text like "More than 500 cases of dengue hemorrhagic fever were reported in Mexico last year, with 30 deaths, Ruiz said.", the reader is informed that the 30 deaths are included in the 500 cases. These cues, which tie the components of the large fact together, may not be easy to pick up automatically. This poses a complex but separable technical problem, which will be addressed in further research.

In conclusion, we should note that a proper evaluation of performance on the Nature scenarios should take the complex structure into account. In particular, the standard MUC scoring scheme will be able to tell us only how well the system recovered the set of incidents from the document.

In the MUC scenarios, the system has to extract a *set* of events or incidents; that set completely describes the answer. However, in the presence of hierarchical relationships, the correct answer is a *graph* (in the case of Nature scenarios, a tree), i.e., the set of incidents *together* with hierarchical links.

Hence, we should also measure how well the hierarchical structure among the extracted incidents corresponds to the correct structure in the document.

Acknowledgements

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.

6. References

Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Meyers, and Mabry Tyson. 1995. SRI International FASTUS system: MUC-6 test results and analysis. In *Proc. Sixth Message Understanding Conf. (MUC-6)*, Columbia, MD, November. Morgan Kaufmann.

Amit Bagga and Alan W. Biermann. 1997. Analyzing the complexity of a domain with respect to an information extraction task. In *Proc. 10th Int'l Conf. on Research on Computational Linguistics (ROCLING X)*, pages 175–94, August.

Amit Bagga. 1997. Analyzing the performance of message understanding systems. In *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, pages 637–640, December.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002. Real-time event extraction for infectious disease outbreaks. In *Proceedings of HLT 2002: Human Language Technology Conference*, San Diego, California, March.

Ralph Grishman. 1995. The NYU system for MUC-6, or where's the syntax? In *Proc. Sixth Message Understanding Conf. (MUC-6)*, pages 167–176, Columbia, MD, November. Morgan Kaufmann.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In Maria Teresa Pazienza, editor, *Information Extraction*. Springer-Verlag, Lecture Notes in Artifi cial Intelligence, Rome.

Lynette Hirschman, Erica Brown, Nancy Chinchor, Aaron Douthat, Lisa Ferro, Ralph Grishman, Patricia Robinson, and Beth Sundheim. 1999. Event99: A proposed event indexing task for broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, February 28–March 3.

1991. *Proceedings of the Third Message Understanding Conference (MUC-3)*. Morgan Kaufmann, May.

1992. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann, June.

1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, November. Morgan Kaufmann.

Roman Yangarber and Ralph Grishman. 1998. Transforming examples into patterns for information extraction. In *Proc. of TIPSTER Text Program Phase III*. Morgan Kaufmann.