PLUM: Contextualizing News For Communities Through Augmentation

Sara Kristiina Elo

Diplôme d'Informatique
Université de Genève, Switzerland
October 1992

Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

February 1996

Author

Program in Media Arts and Sciences
December 13, 1995

Certified by

Kenneth B. Haase
Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by

Stephen A. Benton
Chair
Departmental Committee on Graduate Students
Program in Media Arts and Sciences

PLUM: Contextualizing News For Communities Through Augmentation

Sara Kristiina Elo

Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning
on December 13, 1995
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences at the
Massachusetts Institute of Technology

Abstract

The transition of print media into a digital form allows the tailoring of news for different audiences. This thesis presents a new approach to tailoring called *augmenting*. Augmenting makes articles more informative and relevant to the reader. The PLUM system augments news on world-wide natural disasters that readers often find remote and irrelevant. Using community profiles, PLUM automatically compares facts reported in an article to the reader's home community. The reader, browsing through annotations which PLUM generates, discovers, for example, the number of people affected by the disaster as a percentage of the home town population. The reader can also view the article augmented for other communities. By contextualizing disaster articles and making them more informative, PLUM creates a sense of connectedness.

Thesis Advisor:
Kenneth B. Haase
Assistant Professor of Media Arts and Sciences

This research was sponsored by the News in the Future Consortium.

PLUM: Contextualizing News For Communities Through Augmentation

Sara Kristiina Elo

The following people served as readers for this thesis:

Reader

Jack Driscoll
Visiting Scholar
MIT Media Laboratory

Reader

Mitchel Resnick
Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Table of Contents

Preface

My grandmother did not believe we lived in a house in Mogadishu. From her home in Finland, night after night, she watched the evening news, which showed images of refugee camps on the border of Somalia and Eritrea. We could not convince her that most of Somalia was not a war-zone or that people lived a peaceful life. She could not explain to her worried friends why we had moved, when life was good back in Finland. When all of Somalia did become a war-zone 10 years later, TV had few new images to convey the reality.

An exaggerated idyllic image of a remote country can be as misleading as a miserable one. The head of a French donor organization visited a leprosy centre in Southern India. To encourage donations, he wanted to bring photographs back to Paris. He needed a picture of a little boy smiling by the well. With the sunset. And perhaps a hibiscus bush. My 83-year-old friend Sister Regina shook her head. She did not approve of the idyllic image the visitor was trying to capture. After all, she was running a leprosy centre, not a holiday resort.

Motivated by these experiences, I wanted to build a computer system that could help portray a realistic image of foreign countries. I knew that my system ought to know about different cultures, but how to encode something so complex into a computer algorithm? PLUM does not know much at all about 'cultures' - an ambiguous concept even for people. But it does know locations of countries and basic facts about them, ethnic origins of people, and how languages divide into groups. It also knows how to compare large numbers with familiar entities in an attempt to make them more tangible.

Why did I build PLUM around natural disasters? The United Nations declared the 1990's the International Decade for Natural Disaster Reduction (IDNDR). I attended the Media Round Table of the IDNDR World Conference where attendees pondered how to convince media to cover 'Good News on Disasters', or stories where a prepared community avoided a major disaster. I was left with a second question: Could I build a computer program to change the perspective of breaking disaster news?

And so I built PLUM. It is my first attempt to design a system which adds contextual information to disaster news, news that often seems remote or routine.

Thanks to...

I want to thank my advisor Kenneth Haase for guidance. Thanks, Ken, for allowing me to follow my interests and combine text processing with disasters.

My readers, Jack Driscoll and Mitchel Resnick, receive my big thanks for advice and good suggestions over the past year. Jack, thank you for your editorial comments as well.

Thank you to the Aardvarks, the self-made research group with Joshua Bers, Jeff Herman, Earl Rennison, Deb Roy, Nicolas Saint-Arnaud. 'Varks, you provided the most constructive critique and brainstorming sessions.

To the Machine Understanding group, with Mike Best, Janet Cahn, Anil Chakravarthy, Warren Sack, Push Singh and Zhong-hui Xu, as well as Suguru Ishizaki and Tomoko Koda, thanks for being more than colleagues. You are good friends who gave valuable suggestions.

A big thanks to Walter Bender, your enthusiasm for PLUM gave me a lot of confidence. Thanks to Pascal Chesnais for being the other side of the 'Fishwrap with Plum Sauce' dish and to Jon Orwant for writing the PLUM-Fishwrap server and 'baby prunes'.

My appreciation goes to the CRED Research Centre for the Epidemiology of Disasters for allowing me to use their disaster database in this project.

David Dean, who read through endless drafts of this thesis, begging me to write English, not techno-jargon, you are my best and most critical reader. And credit goes to you for the name "Peace Love and Understanding Machine".

And ultimately, thanks to my parents who, in 1979, picked up us kids and 6 suitcases, starting the Elo family's nomadic life from Somalia. Without them this thesis would not have a preface, nor I a personal motivation.

Chapter 1. Scenario

August, 1995. 7pm in Bellefontaine, a rural town in central Ohio. Dora Newlon, 57, turns on her compute to read the augmented news of the day:

Augmented News for Bellefontaine, Logan County, Ohio, United States.

NIAMEY, Niger (Reuter)

Flooding caused by record seasonal rains in the desert state of Niger killed 42 people and left almost 127,000 homeless, the state news agency ANP said Friday.

Heavy rain since July has destroyed nearly 74,100 acres of crops and killed 6,800 animals, the agency said. The worst affected areas are the central Maradi region and the western Dosso and Niamey regions. The losses are estimated at $52 million.

The average annual rainfall of 14 inches has fallen in a single day in some areas. On August 12, national television was forced to halt broadcasts because its studios were flooded.

Niger is located in Western Africa, between Algeria and Nigeria. Niger is about six times the size of Ohio.

There is no record of Nigeriens living in Bellefontaine, but 30 people of first ancestry Sub-Saharan African live here.

The languages spoken in Niger: French, Hausa, Djerma. Census data categorizes French as 'French-or-French-Creole language'. 43 people in Bellefontaine speak one of 'French-or-French-Creole-languages' at home.

127,000 people is roughly the same as 10 times the people living in Bellefontaine.
127,000 people is roughly the same as 3 times the people living in Logan county.

The total population of Niger is 9,000,000.
127,000 people is the same as 1 person out of 71.

To cover $52 million, every household in Bellefontaine would pay about $1100.

$52 million is about 1% of Niger's GDP, $5.4 billion.

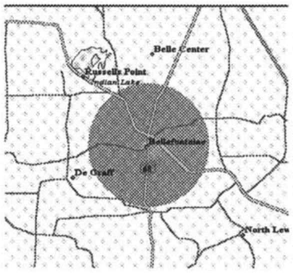
National product per capita
Niger: $650
United States: $24,700

The last serious flood in Niger occurred in August 1988, when 80,000 people were affected, 20 killed and when the total losses were $10,200.

The most serious flood in USA occurred in 1913, when 732 people died and total losses were $200,000

Agriculture in Niger: accounts for roughly 40% of GDP and 90% of labor force; cash crops - cowpeas, cotton, peanuts; food crops - millet, sorghum, cassava, rice; livestock - cattle, sheep, goats; self-sufficient in food except in drought years.

74,100 acres is equivalent to a circle with a radius of 6 miles, shown on the local map:



The total area of Niger is 1.267 million sq km. 74,100 acres, or 285 sq km, is less than 1% of the total land area.

The original article on the Niger flood does not immediately relate to Dora's life in Bellefontaine. Like most of us, she probably knows no one there, where it is, or the difference between Nigeriens and Nigerians. The augmented article provides explanations localized to Dora's home town as well as contextual information on Niger.

Chapter 2. Motivations for Augmenting Text

Local newspapers often rely on wire services such as Associated Press and Reuters for news outside their community. Small newspapers like the Bellefontaine Daily cannot afford to send reporters to cover far-away events. From the incoming news wires, editors choose the articles to include in their paper. Apart from labeling articles that refer to the state, senator or congressman of the client newspaper, wire services rarely indicate the relevancy of their articles to the readership's community. Outside of the obvious references, the local journalists must research the implications of reported events for their home community. When a highway bill passes the Senate, a journalist uses insight, the local library, or other resources to "localize" an article before press time. This is harder with foreign news. When news of the Niger flood arrives, the local journalist must get acquainted with this distant place and, under deadline, scramble to find good resources. Given these pressures, smaller newspapers often reprint international news wires without further refinement for the local readership.

Computer technology has yet to significantly improve the content of news. Most news organizations employ computers to make quantitative improvements, to cut costs, produce faster, and generate better graphics. While 79% of newspapers surveyed by Cable & Broadcasting had computer graphics capability, only 29% had a computerized library and even fewer used information-gathering tools such as CD-ROM databases [Cable & Broadcasting, 1994]. Technology can do more in the newsroom. While it is unlikely that computers take over the creation of news stories, computers play a major role in the on-line versions of many print papers. In an effort to attract readers to their on-line services, newspapers are seeking ways to add value to the digital paper. Unlimited by column space, an on-line newspaper integrating archives of historical articles and other background material can be a meaningful resource. A digital article becomes a gateway to exploring related resources.

Digital news is a young media both for the newspaper industry and the readers. Since a computer allows tailoring of information, digital news can be made meaningful to an individual reader. According to the cognitive science and psychology communities, people understand something new in terms of something they have understood previously (e.g. [Schank 1990]). This supports tailoring news by relating it to familiar concepts in the home community. It can bring news 'closer to home'.

The ideal computer program would present us news according to our personal experiences: when Uncle Heikki is traveling in India, I read news about an earthquake in Southern India carefully. A computer cannot know this unless it has detailed and up-to-date information about each reader. Such information is hard to acquire and maintain. It is easier with publicly available information on a geographic community. Information about demographics, weather history, and geography of a city evolve more slowly than information about an individual. Furthermore, the privacy of this information need not be secured. Contextualizing news to a person's community, not to the person, is more feasible.

Tailoring news by augmenting may also help counter misconceptions. Foreign disaster news often fosters a tragic image of the developing world. The public has "an impression that the developing world is exclusively a theater of tragedy... This misconception is as profound as it is widespread," said Peter Adamson,[1] author of UNICEF's annual State of the World's Children report [Cate 1993]. Misconceptions arise from ignorance and lack of familiarity. The current style of reportage of tragic disasters may only exacerbate these misconceptions. News that clearly explains the scope of the disaster or gives a scale to interpret large figures, provides a more realistic image.

[1] Referring to the 1993 World Vision UK public opinion survey.

Chapter 3. Biased Disasters

TV news broadcasts tend to cover breaking news on disasters with destruction, deaths or injuries. The portrayal of these events influences our image of the disaster-stricken country.

Before the 1980's disaster news depicted helpless, passive victims and heroic saviors [Bethall 1993]. Even relief agencies were criticized for the imagery they used pleading for donations. For example, the extreme Save the Children poster in 1981 pictured a healthy plump white hand holding the hand of a dying African child and read 'Sentenced to Death: Save the Innocent Children.' In 1989, the Code of Conduct on Images and Messages relating to the Third World was adopted to promote a more realistic and informative view of the developing world.

More recently, the Media Round Table at the IDNDR World Conference on Natural Disaster Reduction in May 1994 proposed solutions for a more accurate reporting of disasters. One suggestion is that a reporter return to the site of a disaster. The follow-up article should describe what was learned from the event to prepare and reduce the impact of future disasters. News agencies should cooperate with local disaster managers in charge of preparedness. A disaster manager could suggest story topics to news agencies who don't have the expertise or the staff to investigate the progress at disaster sites. In short, media should cover examples of successful mitigation of disasters, or 'good news' on natural disasters.

Despite the efforts to change the depiction of natural disasters, television images still portray a patronizing view of victims of a disaster. While print offers more space for analysis of the situation, misconceptions about misery in developing countries exist. Viewers and readers tend to generalize from the drama. When a new disaster strikes, the predisposed audience may exaggerate the scale of the disaster.

People may also misunderstand the scale of a disaster simply because large numbers are hard to understand. Powers of Ten [Morrison 1982] explains how all familiar objects can be measured in six orders of magnitude. We can see a few millimeter insect on our palm, while even the tallest trees and buildings are never over a hundred meters. Numbers smaller or larger are hard to imagine. It is difficult to picture the evacuation of 200,000 people or the size of the area required to provide them temporary shelter. A poor knowledge of world geography and distances also causes misunderstandings. For example, when a hurricane hit the Southern Caribbean some years back, tourists cancelled their trips to Jamaica [IDNDR 1994].

This is the same as travelers to New York City staying home when a hurricane hits central Florida.

Disaster news is appropriate for augmentation, because natural catastrophes occur on every continent. Two communities across the world with different cultural backgrounds and life-styles may have little in common. Like people who have lived through a catastrophe feel a bond with others with similar experiences [Wraith 1994], communities that have survived a disaster may feel sympathy and willingness to help one another. By pointing to similar events in two communities, they may feel connected.

From a computational point of view, disasters news is also ideal for augmentation. Automatic extraction of content works best on formulaic text with a restricted vocabulary. The reporting of disaster news tends to follow patterns and lends itself well to automatic analysis. Furthermore, digital background databases pertinent for augmenting disasters are available.

In summary, disaster news is appropriate data for several reasons:

Disaster news is a partial description. Disaster news leaves positive facts unsaid. It evokes misconceptions in readers who generalize from the drama.

Foreign disaster news depicts faraway places. The geographic spread makes for interesting and educational comparisons.

Disaster news reports numbers. Large numbers are difficult to understand. This can lead to misunderstanding the scope of a disaster.

Disaster news tends to follow patterns. The automatic processing of text is more accurate.

Chapter 4. Tailoring Information

The Australian Committee for Disaster Preparedness held a campaign on the hazards of cyclones [Rynn. 1994]. Its poster directed to mainland Australia depicts a Caucasian person flying away with his household affairs and his cow. When the campaign expanded to the Solomon Islands and Vanuatu, it was tailored to appeal to local people. The new poster depicts a native islander being blown away with a pig, the typical domestic animal on the islands. Tailoring to different cultural contexts is not a new concept for graphic designers or advertisers.

Although translation is considered a literal rendering from one language into another, translated text is sometimes altered to fit an audience. A simple but illustrative example is the Finnish translation of Richard Scarry's Storybook Dictionary [Scarry 1966]. The original apple pie on the dinner table is labeled as a potato casserole. Since a Finnish apple pie has no crust on top, Finnish children would not recognize an American-style pie. A faithful translation was sacrificed to avoid misunderstandings.

Traditional print media sees its audience as a mass and sends the same printed message to all readers [McQuail 1987]. Traditionally, when a page was printed with manually assembled metal fonts, cost prohibited tailoring for different readers. Once page layout was computerized, printing different versions became technically simple. However, a newspaper can afford to hire an editor for a special issue only for a sizeable readership. The New Jersey Journal recently launched an electronic journal for the Indian community. [http://nif.www.media.mit.edu/abs.html#india] The India Journal combines international news articles related to India with locally written ones to produce an interesting journal. While an editor puts time and effort into tailoring the content, a computer can automatically and instantaneously adapt information in more than one way.

Information can be dynamically tailored to a desired 'look', or style. Weitzman and Wittenburg [Weitzman 1995] present a computer system that generates different spatial layouts and graphical styles for the same multi-media document. Using a visual grammar based on one document's style, another document can be laid out in the same style. For example, their system transforms the table of contents of the conservative Scientific American to look like one out of WIRED, a publication with an avant-garde lay-out. While the content remains the same, the style of the presentation is tailored for a specific purpose or reader.

Expert systems, computer programs that answer questions about a narrow subject, adapt to the user's level of knowledge. A tailored presentation by an expert system should not present any information obvious to the user or include facts the user cannot understand. The expert system TAILOR [Paris 1993] describes an engine in terms appropriate for a hobbyist or an engineer.

Filtering, a common tailoring technique for digitally distributed news, matches articles with a reader's interest model [Yan 1995]. The simplest reader profile consists of keywords that describe topics of interest. The reader selects the keywords and maintains the personal profile up-to-date. Because keywords fail to describe complex events and relationships, filtering with keywords is not satisfactory. Using a list of keywords to describe a topic yields better results. More sophisticated approaches propose autonomous agents to update a profile by analyzing a person's e-mail and calendar. Because current text processing tools are not reliable enough to do this autonomously, the user needs to verify the profile. Webhunter [Lashkari 1995] uses yet another approach. It generalizes from profiles of other readers with similar interests and proposes a tailored selection of web documents to a person. While filtered information can save a reader's time, it may sacrifice the diversity of information. It is generally agreed, that a day's tailored news needs to be accompanied by a selection of news compiled with another method or by an editor.

The on-line newspaper Fishwrap personalizes news for members of the MIT community [Chesnais 1995]. In addition to specifying topics of interest or choosing to receive news from their home region, readers can follow the interests of the Fishwrap readership. While browsing articles, readers can add meaningful and interesting ones to Page One. Fishwrap presents the latest recommended articles on Page One in an order reflecting the interests of the whole community.

Peace Love and Understanding Machine, PLUM, differs from previous work on digital tailoring, because it adds explanations to existing articles. PLUM concentrates on one subject presenting all articles on natural disasters with explanations that relate to the home community of the reader. PLUM operates on the reasonable assumption that residents are familiar with their home town. Since PLUM does not maintain personal profiles, readers' privacy is not at risk. Also, a single community profile permits tailoring news to all residents of the community.

Chapter 5. What Kinds of Augmentations Does PLUM generate?

5.1. Four Classes of Augmentations

PLUM augments facts reported from the disaster site and generates four classes of augmentations. The examples below show how the augmentations vary for the three prototype home communities Boston, Massachusetts, Bellefontaine, Ohio, and Helsinki, Finland. (The examples are better viewed at http://www.media.mit.edu/people/elo/plum-explanation.html)

5.1.1. PLUM draws comparisons between disaster site and home community

PLUM refers to linguistic and ethnic similarities. It shows how many people with origins in the disaster-struck country live in the reader's home community. It also shows how many people speak its language(s) at home. If there is no information on a specific origin or language, PLUM uses more general information about continents and language groups, taking into account the most recent changes in country borders.

PLUM also generates comparative statistics from the World Fact Book. These statistics vary from one augmentation to another. The reader can follow an html-link to the World Fact Book and explore it in more detail.

When PLUM augments a feature of a disaster, it generates two types of augmentations. Local augmentations refer to the home community and global augmentations to the disaster-struck country. If PLUM augments the same article for several home communities, it generates a global augmentation only once presenting it to all home communities.

Dora in Bellefontaine only sees the local augmentation for Bellefontaine and the global augmentation 'A bit of background on Soviet Union.'

```
Subject: Final Russian quake death toll put at 1,989
[...] MOSCOW ( Reuter ) - Almost 2,000 people died
in last month's massive earthquake on the far east-
ern Russian island of Sakhalin, the deputy governor
of the region said Wednesday. [...]
```

AUGMENTATION FOR HELSINKI

```
There is no record of people from Russia living in
Helsinki, but 4,521 people of 'ex-ussr-origin' live
in Helsinki. [Finnish Census Data]

Russia is roughly 2 times the size of Europe.
Also, Russia is 56 times the size of Finland.
Population density:
```

Russia 9 people/sqkm, Finland 17 people/sqkm

Today's comparative statistics
Female life expectancy:
Russia 74 years, Finland 80 years
...
[World Fact Book - Russia]
[World Fact Book - Finland]

AUGMENTATION FOR BOSTON

10,565 people of 'first-ancestry-russian' live in Boston. [US Census Data]

The languages spoken in Russia: Russian, other. 3,211 people in Boston speak Russian at home. [US Census Data]

Russia is roughly 2 times the size of United States.
Also, Russia is 840 times the size of Massachusetts.
Population density:
Russia 9 people/sqkm, Massachusetts 296 people/sqkm

Today's comparative statistics
Electricity consumption:
Russia 6,782 kWh, United States 12,690 kWh
...
[World Fact Book - Russia]
[World Fact Book - United States]

AUGMENTATION FOR BELLEFONTAINE

Russia is roughly 2 times the size of United States.
Also, Russia is 160 times the size of Ohio.
Population density:
Russia 9 people/sqkm, Ohio 103 people/sqkm

Today's comparative statistics
Male life expectancy:
Russia 64 years, United States 73 years
...
[World Fact Book - Russia]
[World Fact Book - United States]

A BIT OF BACKGROUND ON RUSSIA

Russia is located in Northern Asia (that part west of the Urals is sometimes included with Europe), between Europe and the North Pacific Ocean.

The day of independence of Russia: 24 August 1991 (from Soviet Union)
[World Fact Book - Russia]

The United States has a large immigrant population with significant ethnic communities in many cities. These immigrant populations are almost certainly interested in news from their country of origin. In addition, their colleagues and neighbors may be interested, because they know someone to whom it matters. For the rest of the community, the ethnic and linguistic similarities with the disaster site may be interesting facts. It may also evoke a feeling of sympathy and a sense of a smaller world.

5.1.2. PLUM expands on the history of disasters

PLUM refers to the most serious disaster in the history of the stricken country and the home community. It also compares the frequency of disasters in the two countries. If no disasters of the same type have occurred in the home country, PLUM refers to ones in the neighboring countries. If PLUM has augmented previous articles on the same disaster, the current article is linked to the archived articles.

```
Subject: Hundreds dead as China fears worst floods
this century
SHANGHAI, China ( Reuter ) - China fears its worst
flooding disaster this century with rising waters
already killing hundreds of people and devastating
farms and fisheries in its eastern region. [...]
```

AUGMENTATION FOR BOSTON / BELLEFONTAINE

```
The most serious flood in United States occurred in
March, 1913. People killed: 732; total losses:
$200,000 [CRED Disaster Data Base]

Record of serious floods between years 1970 and 1990
in United States:

1990 *
1989
1988 * * * *
1987
1986 * * *
1985
1984 * *
1983 * * *
1982 * *
1981 *
1980 * * *
1979 *
1978 * *
1977 * * *
1976 *
1975
1974
1973 * *
1972 *
1971
```

1970

AUGMENTATION FOR HELSINKI

There is no record of past floods in Finland.

The most serious flood in neighboring Russia occurred in June, 1993: people affected: 6,500; people killed: 125 [CRED Disaster Data Base]

A BIT OF BACKGROUND ON CHINA

The most serious flood recorded in China occurred in July, 1931, when 3,700,000 people were killed. [CRED Disaster Data Base]

Record of serious floods between years 1970 and 1990 in China:

```
1990 * * * * * * * * *
1989 * * *
1988 * * * * * * * * *
1987 * * * * * * * * * *
1986 * * * *
1985 * * * * * * * * *
1984 * *
1983 * *
1982 * * * * * *
1981 * * * * *
1980 * * * *
1979
1978
1977
1976
1975
1974
1973
1972
1971
1970 * *
```

ARCHIVES OF MOST RECENT FLOODS IN CHINA

Thu, 29 Jun 95: Flood, Drought Plagues China
Wed, 28 Jun 95: Heavy rains kill 100 in central China province

Journalistic techniques inspired this class of augmentations. When a disaster occurs in a distant country, journalists often place the event in the context of the history of similar disasters. This helps the reader understand how rare or common the disaster is for the stricken country. For example, floods occur almost yearly in Bangladesh, but rarely in Nepal. A journalist may take this technique one step further and remind the readers of similar disasters in the home country. Personal experiences affect us deeply, as well as those of our family or close friends. Experiences of third parties that get told by a fourth have less significance. PLUM attempts to make a distant disaster more tangible, by relating it to one that has occurred closer to the reader's community.

A journalist's work also inspired linking an article with archived articles covering the same event. When journalists write a story, they consult past articles on the topic. Pathfinder, the on-line news service of Times Inc. at http://www.pathfinder.com, also allows readers to access its archives. Linking current articles to historic ones broadens the perspective, as shown by their coverage of the Hiroshima bombing from 1945 to 1995. In PLUM, previous articles show the evolution of the disaster since its onset. Skimming the subject lines provides a summary of this evolution.

The PLUM archives currently contain some 150 disaster articles collected since May 1995. Until October 1995 the archives grew as PLUM processed incoming news. The articles complement the CRED disaster database containing statistics up to 1993. The PLUM archives are not part of the Fishwrap version of PLUM.

5.1.3. PLUM provides a yardstick for quantities and areas

To make numbers more meaningful to a reader, PLUM uses techniques inspired by Richard Wurman's book, Information Anxiety [Wurman 1989]. Wurman points to the need for a 'personal media-measuring stick.' He wants to turn 'statistics into meaningful numbers, data into information, facts into stories with value.' Wurman seeks common sense explanations. As an example of a meaningless entity, Wurman takes the size of an acre. If he explains that an acre equals 43,560 square feet, we still cannot imagine its size. But, if he says it is roughly the size of a football field, we have a good idea. PLUM explains a hurricane's 100-mile trajectory using a well-known path in the home town. But PLUM does not have common sense and does not know about common objects. To explain that flood waters in Vietnam rose high enough to cover Boston's Longfellow Bridge, PLUM would have to know the location of Longfellow Bridge, its height, and, to be accurate, the amount of water in Charles River as well as the topography of the river banks. Instead of generating comparisons of this detail, PLUM uses census data, overlays on digital maps, and local distances to provide a scale for understanding numbers.

Augmentation for Finland: A football field is about the same size as a soccer field.

Augmentation for Boston: 100 miles is the same as driving from Boston to Provincetown.

- **PLUM compares numbers of people affected** by a disaster to the population of the home town and to a local group of people approximately the same size. It also compares the number to the total population of the disaster-stricken country.

```
Subject: Russia rescue teams stop search for quake
```

> victims
>
> MOSCOW (Reuter) - Rescue teams in the Russian far eastern oil town of Neftegorsk Friday officially stopped the search for survivors of a giant earthquake which officials fear killed at least **2,000** people. [...]

AUGMENTATION FOR HELSINKI

> 2,000 is 1 out of every 250 people living in Helsinki.
> 2,000 is also 2 times all the farmers living in Greater Helsinki.
> [Finnish Census Data]

AUGMENTATION FOR BOSTON

> 2,000 is roughly the same as 1 out of every 290 people living in Boston.
> 2,000 is also 1 out of 4 people 85 years or older living in Boston.
> [US Census Data]

AUGMENTATION FOR BELLEFONTAINE

> 2,000 is 1 out of every 6 people living in Bellefontaine.
> 2,000 is also all the children under 8 living in Bellefontaine.
> [US Census Data]

A BIT OF BACKGROUND ON RUSSIA

> The total population of Russia is 149,608,953 (July 1994 est.). 2000 people is the same as 1 person out of 75000. [World Fact Book - Russia]

•**PLUM compares numbers of families** mentioned in the article to the number of families in the home community. It also compares the average size of a family in the two countries.

> Subject: 172 die in Vietnam floods
> [...] An official from the government's natural relief committee told Reuters 218,415 houses had been submerged or swept away in six delta provinces and **146,550** families needed emergency food assistance. [...]

AUGMENTATION FOR HELSINKI

> 146,550 is roughly all the families living in Hels-

inki. [Finnish Census Data]

The average size of a family in Helsinki is 3. [Finnish Census Data] In Finland, the average fertility rate is 1.79 children born/woman, while in Vietnam it is 3.33 children born/woman. [World Fact Book - Vietnam]

AUGMENTATION FOR BOSTON

146,550 is roughly all the families living in Boston.
[US Census Data]

The average size of a household in Boston is 2.51.
[US Census Data]
In United States, the average fertility rate is 2.06 children born/woman, while in Vietnam it is 3.33 children born/woman
[World Fact Book - Vietnam]

AUGMENTATION FOR BELLEFONTAINE

146,550 is roughly 45 times the families living in Bellefontaine.
[US Census Data]

The average size of a household in Bellefontaine is 2.56. [US Census Data] In United States, the average fertility rate is 2.06 children born/woman, while in Vietnam it is 3.33 children born/woman. [World Fact Book - Vietnam]

•**PLUM compares numbers of houses** mentioned in the article to the number of households in the home community.

Subject: Hundreds dead as China fears worst floods this century
[...] Shanghai's Liberation newspaper Friday put the Jiangxi toll much lower at 64. It said **220,000** houses in the area had been swamped and put the cost of damages at $500 million. [...]

AUGMENTATION FOR HELSINKI

220,000 is roughly all the households in Helsinki.
[Finnish Census Data]

AUGMENTATION FOR BOSTON

220,000 is roughly all the households in Boston.
[US Census Data]

AUGMENTATION FOR BELLEFONTAINE

220,000 is roughly 46 times the households in Bellefontaine.
[US Census Data]

•**PLUM explains large dollar-amounts** by calculating how much every home town household would have to pay to cover the amount. It also compares an amount to economic statistics of the home and foreign country. Because economic indices are not well understood concepts, PLUM provides a hyper-link to a page explaining how these statistics are calculated.

Subject: 172 die in Vietnam floods
[...] The communist party newspaper Nhan Dan quoted Agricultural Minister Nguyen Cong Tan as saying the floods had caused the loss of an estimated 200,000 tons of rice and damage worth **$54 million**. [...]

AUGMENTATION FOR HELSINKI

To cover $54,000,000, every household in Helsinki would pay $423 or 1,691 FIM. [Finnish Census Data]

National product per capita
  Finland: $24,700
  Vietnam: $1,000
[World Fact Book 1993 - Vietnam]
[World Fact Book - Finland]

What do these statistics mean? Click here.

AUGMENTATION FOR BOSTON

To cover $54,000,000, every household in Boston would pay $459. (The average yearly household income in Boston is $29,180) [US Census Data]

National product per capita
  United States: $24,700
  Vietnam: $1,000
[World Fact Book - Vietnam]
[World Fact Book - United States]

What do these statistics mean? Click here.

AUGMENTATION FOR BELLEFONTAINE

To cover $54,000,000, every household in Bellefontaine would pay $16,580. (The average yearly household income in Bellefontaine is $25,221) [US Census Data]

National product per capita
United States: $24,700
Vietnam: $1,000
[World Fact Book - Vietnam]
[World Fact Book - United States]

What do these statistics mean? Click here.

A BIT OF BACKGROUND ON VIETNAM

$54,000,000 is less than 1 o/oo of Vietnam's GNP - purchasing power equivalent - $72 billion
[World Fact Book - Vietnam]
What do these statistics mean? Click here.

•**PLUM overlays a shadow of an area of land** on the home town map

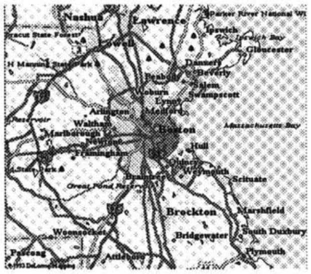
Subject: Record rains flood Niger
[...] Heavy rain since July has destroyed nearly **74,100 acres** of crops and killed 6,800 animals, the agency said. [...]

AUGMENTATION FOR HELSINKI

74,100 acres is equivalent to a circle with a radius of 10 miles. Sorry, no map available for overlay.

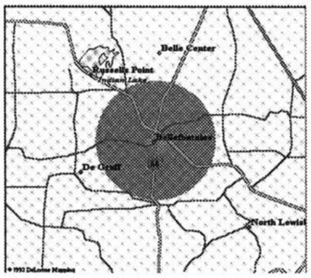
AUGMENTATION FOR BOSTON

74,100 acres is equivalent to a circle with a radius of 6 miles, shown here on the local map.
[DeLorme Mapping]



AUGMENTATION FOR BELLEFONTAINE

74,100 acres is equivalent to a circle with a radius of 6 miles, shown here on the local map.
[DeLorme Mapping]



A BIT OF BACKGROUND ON NIGER

The total area of Niger is 1.267 million sq km. 74,100 acres, or 258 sq km, is less than 1% of the total area. [World Fact Book - Niger]

•**PLUM compares distances** mentioned in the article to frequently traveled or well known routes in the home community. Chapter 10. “FactBase: Public or Collected Data?” discusses in detail this type of augmentation.

Subject: Swiss Photographer May Have Ebola
[...] All the cases are concentrated in and around Kikwit, **250** miles east of the Zairian capital, Kinshasa. [...]

AUGMENTATION FOR BOSTON

250 mi is roughly the distance from Boston to New York City.

AUGMENTATION FOR BELLEFONTAINE

250 mi is roughly the distance from Bellefontaine southeast to Charleston, WV.

•**PLUM adds background information** on the agriculture in the disaster-struck country.

Subject: Flood, Drought Plagues China
[...] About 9 million people and 8.3 million heads of livestock are short of drinking water and 22 million acres of **farmland** have been parched, the China Daily said. [...]

A BIT OF BACKGROUND ON CHINA

Agriculture in China: accounts for 26% of GNP; among the world’s largest producers of rice, potatoes, sorghum, peanuts, tea, millet, barley, and pork; commercial crops include cotton, other fibers, and oilseeds; produces variety of livestock products;

```
basically self-sufficient in food; fish catch of
13.35 million metric tons (including fresh water and
pond raised) (1991).
[World Fact Book - China]
```

5.1.4. PLUM links disaster articles to related WWW sites

Readers can add links to related World Wide Web sites using an html form. These added WWW sites may provide background information, address issues of preparedness, or point to discussion groups. With the contributions of its readers, the PLUM database continues to grow. For example, the Federal Emergency Management Agency's web-site explains how to prepare for the different natural disasters and why and how such disasters occur.

PLUM links an augmentation to its source database allowing further exploration. Currently, only the World Fact Book is available on WWW. PLUM generates an html document of comparative statistics from the World Fact Book entries of the home country and the disaster stricken country. The comparison includes the world's smallest and largest values for each statistic.

```
Japan - United States comparison

location
Japan: Eastern Asia, off the southeast coast of Rus-
sia and east of the Korean peninsula
United States: North America, between Canada and
Mexico

area total area
Japan: 377,835 sq km
United States: 9,372,610 sq km
Most in the world: Russia 17,075,200 sq km
Least in the world: Vatican City 0.44 sq km

coastline
Japan: 29,751 km
United States: 19,924 km
Most in the world: Canada 243,791 km
Least in the world: (landlocked) Afghanistan,
Andorra, ...

etc...
```

5.2. Critique of Augmentations

Arguably, no right way exists to compare facts from different cultures and societies. Facts are sensitive to context and subject to interpretation. For example, if an article reports the evacuation of 500 families in Vietnam, a comparison with Boston cannot be readily drawn. In Vietnam, families may include several generations of relatives, while many American families are nuclear or separated. PLUM cannot resolve this difference in definition. Nor can it represent the value of money in different cultures.

Simply converting between currencies does not suffice. A farmer in a subsistence economy who loses a buffalo and a plough has lost his livelihood. In dollars, this may amount to $200. The equivalent in Bellefontaine would mean the loss of Dora's hardware store. Since PLUM cannot reason this way, it compensates by proposing several different interpretations for the facts in the article. Each entity is augmented several ways to give the most accurate account.

An interesting way to make a disaster more tangible would be to simulate it in the home community: "The flood in Vietnam is as severe as if the waters of Charles River rose to reach Central Square in Cambridge, MA." This kind of analogy is extremely hard to draw. Natural disasters are complicated processes. They depend on a number of factors and have a range of consequences. For instance, the magnitude of an earthquake depends on ground density, fault lines, water table levels, stress levels in the ground, and other factors yet to be discovered. The destruction it causes depends on building materials, architecture, the direction of the earth's movement, the presence of precursors, and the preparedness of the community. The Kobe Earthquake in January 1995 illustrated that even the time of day is an important factor: the earthquake struck early in the morning when people were cooking with open flames. This resulted in a high number of fires in the city. Since no formula can accurately predict the consequences of a disaster, simulating a disaster in the home community would be a simplification. At worst, constructing such an analogy could give the impression that all factors of the natural disaster are known and well understood.

Chapter 6. Description of the PLUM System

PLUM consists of four components. The **Parser** analyzes in-coming disaster news. The **FactBase** contains all descriptive and statistical resources.The **RuleBase** contains the rules for augmenting. The **Augmenter** executes these rules to produce the augmented text. The reader receives the augmented news on the World Wide Web.

6.1. Parser

Text processing by computer is difficult. A computer program cannot infer the meaning of natural language accurately. Language in non-restricted text is ambiguous and sensitive to context. However, stylized text, with a limited vocabulary or sentences that follow patterns, allows a more accurate extraction of content. PLUM takes advantage of the structure and wording of stereotypical disaster articles.

6.1.1. The disaster template

The PLUM Parser extracts the following features from a disaster article:

COUNTRY-AFFECTED:	principal country affected
DISASTER-TYPE:	earthquake, flood, avalanche, famine, fire, heat wave, volcano, tsunami, epidemic, storm, landslide, cold wave.
PEOPLE-AFFECTED:	numbers of people reported (usually describes people evacuated, stranded, homeless, etc....)
PEOPLE-KILLED:	numbers of people killed
LOSSES:	$ amounts (usually describes losses or aid)
FAMILIES-AFFECTED:	numbers of families (usually describes families evacuated, homeless, etc....)
LAND-AFFECTED:	land area in acres, sq mi, sq km, ha (usually describes land flooded, burnt, etc....)
CROPS-ANIMALS-AFFECTED:	crops or animals reported
HOUSES-AFFECTED:	numbers of houses, dwellings, homes, huts (usually describes houses destroyed, submerged, etc....)
DISTANCES:	distances in mi or km reported

The goal of the Parser is not to understand every detail in an article, but only the characteristic items reported. Because disaster articles tend to be formulaic, the ten features of the disaster template, or a subset of them, appear in most articles. Initially, the PLUM disaster template consisted of the first five features, selected after the format of the CRED Natural Disaster Database. Since the CRED database provides global statistics on natural disasters, the data on one disaster is limited. After analyzing fifty disaster articles, the five other features were added to the PLUM template. The new features occur frequently in articles and are relatively easy to extract automatically.

6.1.2. Pattern matching

The PLUM Parser extracts the disaster features using pattern-matching techniques. A pattern describes how a feature is reported in a disaster article.

For example, all words matching one of the patterns below are candidates for the feature COUNTRY-AFFECTED in the disaster template:

- the name of a country, such as "Russia"
- an adjective denoting a nationality, such as "Russian"
- a possessive of a country name, such as "Russia's"

Values for the feature PEOPLE-KILLED are all numbers matching one of the following patterns:

a number followed by "people" or its synonym, followed within 5 words by the passive verb "killed" or its synonym

(e.g. `3 people in the town were killed`)

- active verb "killed" or synonym, followed within 5 words by a number and "people" or synonym

(e.g. `the flood killed at least 3 residents`)

- a number and "dead"

(e.g. `3 dead`)

- a number and "people" or synonym, followed within 5 words by active verb "died" or synonym

(e.g. `3 children may have drowned`)

- a number and "died" or synonym

(e.g. `3 perished in the waters`)

- "death toll" followed within 5 words by a number

(e.g. `the death toll has risen to 3`)

- "claimed" or synonym, followed within 5 words by a number and "lives"

(e.g. `the flood claimed more than a hundred lives`)

The patterns were constructed using 50 articles as examples. They were tested on another 50 articles and corrected accordingly (See 8.1. Quantitative evaluation) Annex 1 lists the complete patterns for the ten features.

To analyze articles, PLUM uses Professor Ken Haase's multi-scale parser [Haase 1993]. The multi-scale parser tags every word in text with the class and the stem of the word. For example, in `"30 people were evacuated"` `"evacuated"` is tagged as #("evacuated" :PASSIVE-VERB "evacuate"). The class and stem information in the tag is useful. It allows constructing patterns that take into account the tense of a verb or ignore whether a noun is singular or plural.

The words accepted as synonyms in a pattern, e.g. `"drowned"` and `"died"`, or `"people"` and `"residents"`, were compiled manually after studying the wording in example articles. Incorporating a digital thesaurus, such as Wordnet [Miller 1990], could help to determine synonyms or related words automatically. However, since wire services use a relatively standard vocabulary to describe disasters, the hand-constructed patterns give satisfactory accuracy as shown in Section 8.1.

6.1.3. Example of Parser output

> Subject: **Quake** kills 70 in **Russia's** Far East-agency
> Date: Sun, 28 May 95
> MOSCOW, May 28 (Reuter) - **Seventy** people were killed when a strong **earthquake** hit the island of Sakhalin in **Russia's** Far East, Interfax news agency said on Sunday. It quoted the deputy governor of the region, Vitaly Gomilevsky, as saying that about 2,000 of the **3,500** residents of the northern town of Neftogorsk had been buried. Only **500** people had so far been found alive, it said. The **quake** struck in the early hours of Sunday morning local time when **most** people in Neftogorsk were asleep. Strong fog was hampering rescue efforts, Interfax said, quoting Gomilevsky. Alexander Yudin, a spokesman for the local civil defence headquarters in the island's capital Yuzhno-Sakhalinsk, said the epicentre of the **quake** was offshore **80 km** beyond the northern tip of the island, Cape Elizaveta. It measured 7.5 on the Richter scale, he said by telephone. The southern-most point of the island, which is rich in oil and gas, is just **30 miles** from **Japan's** northern coast.

```
COUNTRY-AFFECTED:       "Russia"
                        plum extracted:
                        ("Russia" sentence 0)
                        ("Russia" sentence 1 word 22)
                        ("Japan" sentence 8 word 21)

DISASTER-TYPE:          "earthquake"
                        plum extracted:
                        ("earthquake" sentence 0)
                        ("earthquake" sentence 1 word 15)
                        ("earthquake" sentence 4 word 2)
                        ("earthquake" sentence 6 word 23)
```

```
PEOPLE-AFFECTED:          (70 3500 500)
                          plum extracted:
                          ("Seventy" sentence 1 word 8)
                          (3500 sentence 2 word 20)
                          (500 sentence 3 word 2)
                          ("most" sentence 4 word 14)

PEOPLE-KILLED:            plum extracted:
                          ("Seventy" sentence 1 word 8)

LOSSES:                   -

FAMILIES-AFFECTED:        -

LAND-AFFECTED:            -

CROPS-ANIMALS-AFFECTED:-

HOUSES-AFFECTED:          -

DISTANCES:                (30 80) (mi km)
                          plum extracted:
                          (30 sentence 8 word 18)
                          (80 sentence 6 word 26)
```

The Parser detects the type of the disaster -- `earthquake, quake` -- as keywords and synonyms of "earthquake". If different types of disasters are mentioned in the article, the type occurring most often or in the subject line is selected. If several types occur the same number of times, it selects the one occurring earliest in the article.

The Parser extracts all country names, country adjectives, and country possessives from the text (`Russia's, Russia's, Japan's`) It selects as the country affected -- `Russia` -- the name of a country occurring most often in the article or in the subject line. If several names occur the same number of times, it selects the one occurring earliest in the article.

It extracts the number of people killed -- `Seventy` -- because it matches the pattern *"a number and 'people' or synonym, followed within 5 words by passive verb 'killed' or synonym"*.

The Parser finds the numbers of people affected by the disaster -- `Seventy, 500, 3500, most` -- because they match the pattern *"a number and 'people' or synonym"*.The word `Seventy` is ignored, because it also fills the more constrained pattern for number of people killed. The word `most` matches this pattern because the multi-scale parser tags it as a number. Because PLUM cannot translate it into an integer, it is ignored. In the phrase `2,000 out of 3,500 residents` PLUM fails to resolve what `2000` modifies. Thus, it is not extracted.

Any number followed by “mi” or “km” is extracted as a distance.

6.1.4. Discussion

Naturally, disaster articles report more than the PLUM Parser patterns detect. An article may report local and foreign assistance, reasons the disaster was extraordinary, the history of disasters in the area, or it may quote a survivor or a local official. Because it is difficult to build patterns for complex concepts, the disaster template does not include them. A pattern is difficult to construct for an idea expressed over several sentences or paragraphs. It may also be difficult if the idea can be worded in many different ways. PLUM augments numeric values and specific characteristics of a disaster. The rest of the article is left untouched.

6.2. FactBase

The FactBase contains information on the readers’ home communities, the disaster-struck countries and the history of natural disasters. Two large cities, Boston, Massachusetts, and Helsinki, Finland, and a small rural town, Bellefontaine, Ohio, were the initial prototype home communities. Buenos Aires, Argentina, was added later to provide the perspective of a third continent.

6.2.1. Data structures in PLUM FactBase

The FactBase is stored in FramerD [Haase 1995], a frame-based representation language. The basic object in FramerD is a frame. Minsky [Minsky 1986] introduced frames as a data structure for Artificial Intelligence programs to store knowledge. A frame has slots of values. For example, a thesis document can be described as a frame:

```
(thesis
    :has-part ‘(chapter-1 chapter-2)
    :is-a ‘document
    :title “PLUM”)
```

Because a slot can store a pointer to another frame (chapter-1 and chapter-2 can be names of other frames), a set of frames can be linked to create a network of information. Frames are an appropriate way to represent the FactBase because it is a collection of objects with cross-references.

FramerD uses the Dtype library to construct data objects, ship them over networks, and store them in data base files. The data can be stored in distributed data servers, by defining distinct pools of objects. The PLUM objects all reside in a single pool for simplicity.

FramerD is practical because it allows creating objects and only later specifying relationships between them. In FramerD PLUM maintains the original labels of the three databases, so that new data from these databases is easy to incorporate.

6.2.2. Home community data

The description of each home community was compiled from the country’s census data. The data for a home community describes its population by origin, age, language spoken at home, occupation, income level. For a complete description of a community, see Annex 2. The FactBase contains this data for the city, and the county, the state and the country in which it is located. Each community also includes maps and local distances.

The home community data is organized in layers using frames.

```
(Boston
    :part-of “Suffolk”)
(Suffolk
    :has-part ‘(“Boston”)
    :part-of “Massachusetts”)
(Massachusetts
    :has-part ‘(“Suffolk” ... )
    :part-of “Unites States”
    :is-a ‘state)
(United States
    :has-part ‘(“Massachusetts” ...)
    :part-of “America”
    :is-a ‘country)
```

If data does not exist at the city level, the augmenting algorithm uses the :part-of relation between frames to search at the county level. The layered representation allows PLUM to zoom out from city data all the way to country data.

The linguistic and ethnic information is also stored in the home community frame. For example, the information pertaining to Ukraine in the Boston frame is describes as follows:

```
(Boston:
    :russian-language 3211
    :first-ancestry-ukrainian 1277
    ...)
```

6.2.3. Country data

The CIA World Fact Book provides background information on every country of the world, its people, government, geography, economy, communications and defence forces. For the complete data for a country, see http://www.media.mit.edu/people/elo/cia/index.html. PLUM employs this data as a ‘yard-stick’ to help explain the impact of a disaster on a stricken country. It also uses the data to compare the disaster-stricken country with the home country of the reader.

The World Fact Book is available on the Internet as ASCII files, not as a database. In order to use the data, it was necessary to parse the information and convert it to a usable format. Every attribute of a country was converted from text into a frame.

```
before    Belgium
          LAND-BOUNDARIES:
          "total 1,385 km, France 620 km, Germany 167 km, Luxembourg 148 km,
          Netherlands 450 km"

after     Belgium-LAND-BOUNDARIES:
             :list                  ("total" "France" "Germany" "Luxembourg" "Netherlands")
             :total-number          1385
             :total-unit            "km"
             :France-number         620
             :France-unit           "km"
             :Germany-number        167
             :Germany-unit          "km"
             :Luxembourg-number     148
             :Luxembourg-unit       "km"
             :Netherlands-number    450
             :Netherlands-unit      "km"
```

The conversion was difficult to do automatically, because the World Fact Book does not use a consistent notation across the collection. For example, commas, colons and semi-colons are used inconsistently to mark boundaries between different types of data. Eventually, using complicated functions, all useful fields were automatically converted. As a result of the conversion, PLUM knows, for example, countries adjacent to each other. If no floods have occurred in Belgium, PLUM searches for historical floods in the neighboring countries.

6.2.4. Cross-indexing home community and country data

Because the databases compiled into the Factbase use different names and conventions for the same entities, it was necessary to cross-index the data. Every country frame has a slot ‘origin-tags which contains the name of the origin of the country in a home community frame. Similarly, each language frame has a slot ‘language-tags which contains the name of the language in a home community frame. For example, should a disaster strike Ukraine, PLUM would search for the slot ‘first-ancestry-ukrainian in a home community.

```
(Ukraine
    :origin-tags ‘(first-ancestry-ukrainian)
    ...)
(French
    :language-tags ‘(french-or-french-creole-language)
    ...)
```

6.2.5. Disaster history data

PLUM also incorporates the CRED Disaster Database [CRED 1993]. The database contains the history of world-wide natural disasters since the beginning of the century. For example, the record of the great Chinese earthquake of 1976 is:

```
(cred-76050
    :date-yy            76
    :date-mm            7
    :date-dd            27
    :country-affected   “China”
    :type               ‘earthquake
    :people-killed      242000
    :people-affected    0
    :people-homeless    0
    :people-injured     164000
    :losses             7000000)
```

Because the CRED Data Base was designed to record global statistics on natural disasters, it does not contain information about the specific location. For example, the data on the Chinese earthquake of 1976 would be more meaningful if its epicentre, the city of Tangshan, was recorded in the database. People remember events such as “the North Ridge earthquake”, not “the earthquake that struck United States in January 1994”.

6.2.6. Cross-indexing disaster and country data

The disaster data is cross-indexed with the country data. Each World Fact Book country frame has a pointer the CRED country frame which contains slots for all past disasters in the country. This permits easy retrieval, for example, of all earthquakes that have occurred in China.

6.2.7. Need for a standard format

The lack of a standard database format complicates the use of public data sources. While computers access and let users read remote documents in ascii, html, or other formats, no standard exists that allows computers to access and *use* remote databases. Most data collections on the Internet contain little information, apart from html tags, about the structure of their content. A data source should be accompanied with a standard machine-readable description of its content. Presently, every project tapping into an on-line database probably wastes time reformatting the data for its own purpose.

For example, The CIA World Fact Book could be tagged to differentiate the data types. The data file on Belgium could have the following format:

```
<RECORD> <TITLE> Belgium </TITLE>
  ...
  <FIELD> <TITLE>land boundaries</TITLE>
      <LIST>
      <ELT> <NAME>total</NAME> <NUMBER>1,385</NUMBER> <UNIT>km</UNIT>
                          </ELT>
      <ELT><NAME>France</NAME> <NUMBER>620</NUMBER> <UNIT>km</UNIT>
                          </ELT>
      ...
      </LIST>
  </FIELD>
  ...
</RECORD>
```

6.3. RuleBase

The RuleBase defines how PLUM augments text. Each feature in the disaster template is associated with one or more augmentation rules. Some rules define how to compare distances, areas, quantities, and currencies in the disaster site against those in the home community. Other rules describe how to add background facts about the disaster-struck country. Because the rules are not specific to a home community, PLUM can accommodate new communities. Annex 4 lists the complete set of augmentation rules.

6.3.1. An example rule

When PLUM fires, or executes, a rule, it searches for data in the FactBase. For example, the following two rules define how to augment the feature PEOPLE-AFFECTED. Let's assume the extracted value is 1,841.

rule "Find the factor between home community population and number of people affected"

produces `1,841 is roughly 1 out of 270 people in Helsinki.`

```
rule "Start at the city level of home community
        Loop: look for a group of people which is a multiple of the number of people affected,
            Find the factor between the two
        Redo loop for next level up until level is country
        Select group of people with smallest factor"
```

produces `1,841 is roughly 1 out of 27 people aged 75 or older living in Greater Helsinki.`

6.4. Augmenter

For every feature of the disaster, the Augmenter fires the associated augmenting rules, finds the appropriate data in the FactBase, and adds the generated augmentations as hyper-links to the original article.

6.4.1. Sentence generation

PLUM generates the text for an augmentation by filling in a sentence template with the appropriate information from the FactBase. For example, the sentence template below generates the comparison in size of two countries, where *factor* is the relationship between the two sizes,

```
if *factor* is 1
      write: *site* is roughly the same size as *home*
if *factor* is greater than 1
      write: *site* is roughly *factor* times the size of *home*
if *factor* is less than 1
      write: *home* is roughly *factor* times the size of *site*
```

It produces sentences, such as,

> China is roughly the same size as United States.
>
> Bangladesh is roughly 7 times the size of Massachusetts.
>
> Finland is roughly 3 times the size of Guatemala.

6.4.2. Map generation

The Augmenter overlays on the map of the reader's home town a shadow of any area of land mentioned in the article. Because people are generally familiar with the map of their home town, the overlay is effective for explaining the size of an area. Another way to explain an area of land would be to name a neighborhood of the same size in the home community, e.g. "500 acres is the same as the size of Back Bay in Boston." However, unless a map of the neighborhood accompanies such a comparison, it only works for neighborhoods with well-defined borders. As shows Chapter 10, most neighborhoods have ill-defined borders in residents' minds. Furthermore, if an area affected by a disaster is compared to a specific neighborhood, people might draw further analogies between the two areas. In the case of Back Bay, readers may think that the affected area is an expensive part of town or one near a shoreline.

Generic city and state maps at different scales overlain with varying sizes of circles explain a land area of any size. Because a circle is an abstract shape, it is clearly not the actual shape of the land area mentioned in the article.

The procedure for an overlay is simple: Given an area of land and the scale of the map, PLUM calculates the equation of a circle with the same surface area. It then turns dark all white background pixels that fall within the circle preserving roads and other landmarks, as shown in Fig 6.1.

6.4.3. Augmented text as hyper-text

Augmentations alter the reading of the disaster story. The original linear story becomes a hyper-text document. Hypertext is suitable for PLUM, because it allows branching from the original article to the augmentations. Augmentations are a click away from the body of the article. Readers can explore at their own pace the augmentations they choose. Since PLUM augments an article for several home communities, with another click, the readers can view the Niger flood from Heikki's perspective in Helsinki.

On the other hand, reading a hyper-text document is choppy, because links appear in the middle of the text. It is also easy to get lost when following links, because there is no global reference point. To keep the reader from losing the context, the augmentation page repeats the augmented word in the surround-

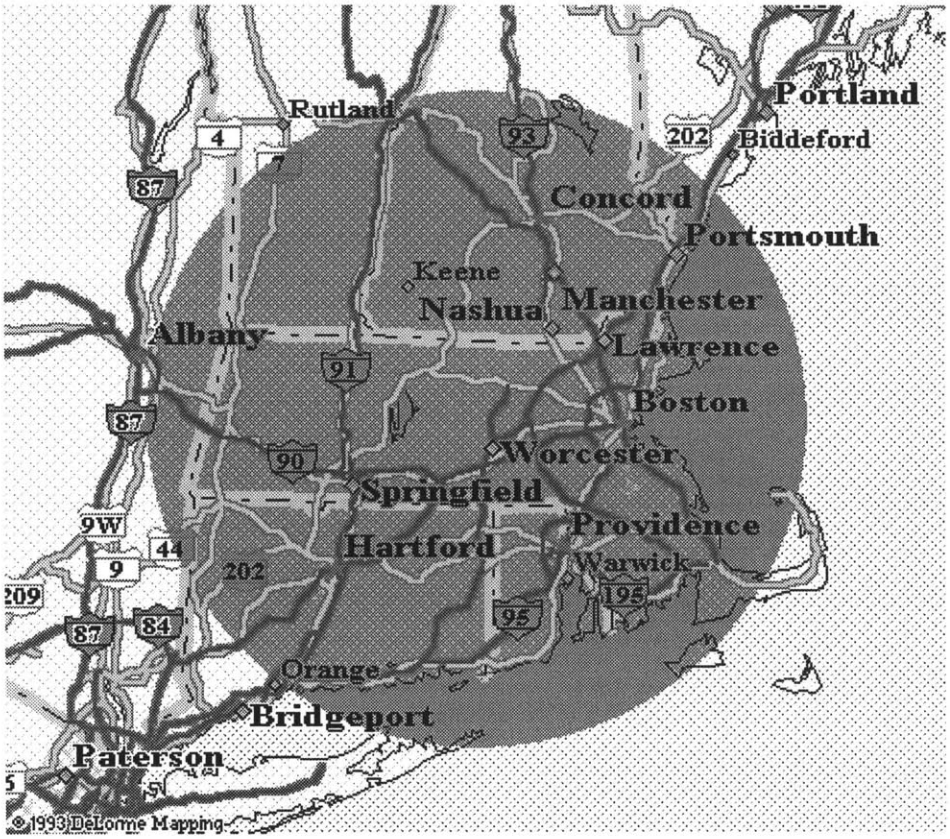


Fig 6.1

ing sentence. The recent version 2.0 of Netscape allows displaying more than one html document at once. The Reader would be able to view the original article in one window and the augmentation in another.

Chapter 7. Integrating PLUM with Fishwrap

7.1. Fishwrap with PLUM sauce

Fishwrap, the on-line news service for the MIT community, is an ideal testing ground for PLUM. Fishwrap personalizes news to readers by allowing them to select filters: readers can subscribe to news on topics of their interest or reports from their home town, state or country. PLUM is an appropriate extension to Fishwrap, because augmenting complements filtering: readers receive news with localized contextual information as well as news of their interest.

In the first stage of PLUM in Fishwrap, all disaster news is augmented for the readers' current home town, Boston. They can also view augmentations for Buenos Aires, Bellefontaine and Helsinki to see how explanations vary from one city to another. Later, if a representation for every state capital in the United States is added, American students will be able to read disaster news augmented for their home state. The design of PLUM supports adding new home communities from census data, as described in Annex 3 - How to Add a Home Community.

7.2. Communication between PLUM and Fishwrap

Fishwrap forwards to PLUM all articles containing a disaster keyword: storm(s), volcano(s), eruption, earthquake(s), quake(s), flood(s), flooding, landslide(s), epidemic(s), virus, drought(s), typhoon(s), tsunami(s), fire(s), cyclone(s), tornado(s), avalanche(s), famine, heat wave, cold wave. This simple initial test does not reject articles where disaster keywords appear in contexts other than disasters, such as in "`People were flooding to the market place.`" PLUM applies a second test to the articles. If an incoming article reports at least four features from the disaster template and names a disaster a sufficient number of times in relation to its length, PLUM returns the augmented article to Fishwrap. Otherwise, PLUM rejects the article. This second test attempts to assure that PLUM only augments true disaster articles, not ones that refer to a disaster in passing.

Fishwrap does not communicate directly with PLUM. Instead it sends requests to a Perl server named Orwant after its author. Orwant handles the messages between PLUM and Fishwrap. It can accept requests from several clients guaranteeing that an augmented article is returned to the proper client. Fig 7.1 illustrates the protocol of communication between Fishwrap and PLUM.

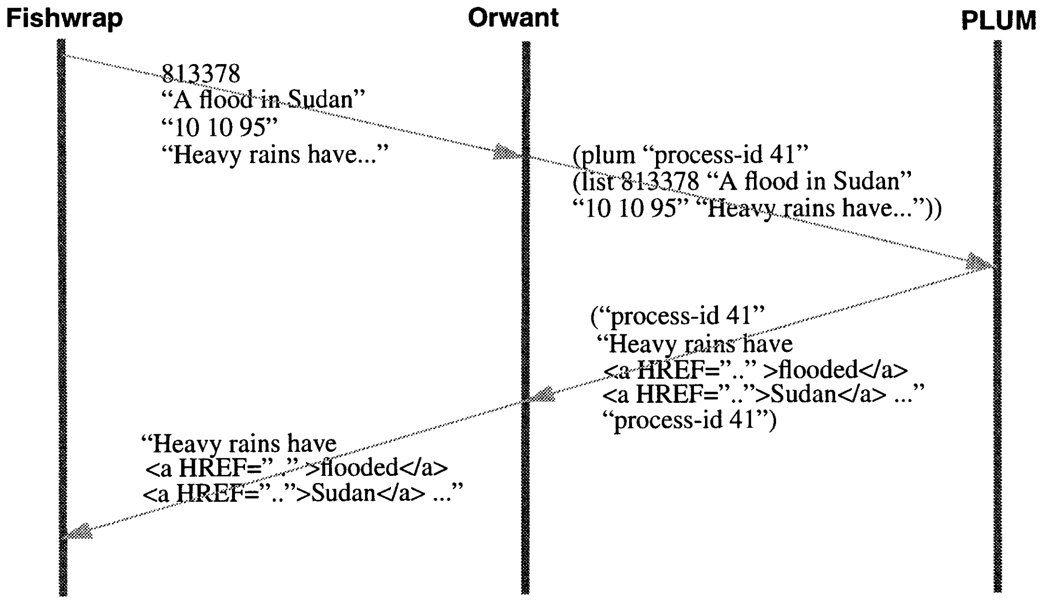

Fishwrap
Orwant
PLUM
813378
"A flood in Sudan"
"10 10 95"
"Heavy rains have..."
(plum "process-id 41"
(list 813378 "A flood in Sudan"
"10 10 95" "Heavy rains have..."))
("process-id 41"
"Heavy rains have
<a HREF=".." >flooded</a>
<a HREF="..">Sudan</a> ..."
"process-id 41")
"Heavy rains have
<a HREF=".." >flooded</a>
<a HREF="..">Sudan</a> ..."

Fig 7.1

The PLUM server has been running with few interruptions since October 7, 1995. It augments roughly 40 Associated Press wire service articles a day. Associated Press sends on the average 2 updates for every article. Thus, PLUM generates on the average 13 distinct augmented disaster articles every day.

7.3. Augmented articles in Fishwrap

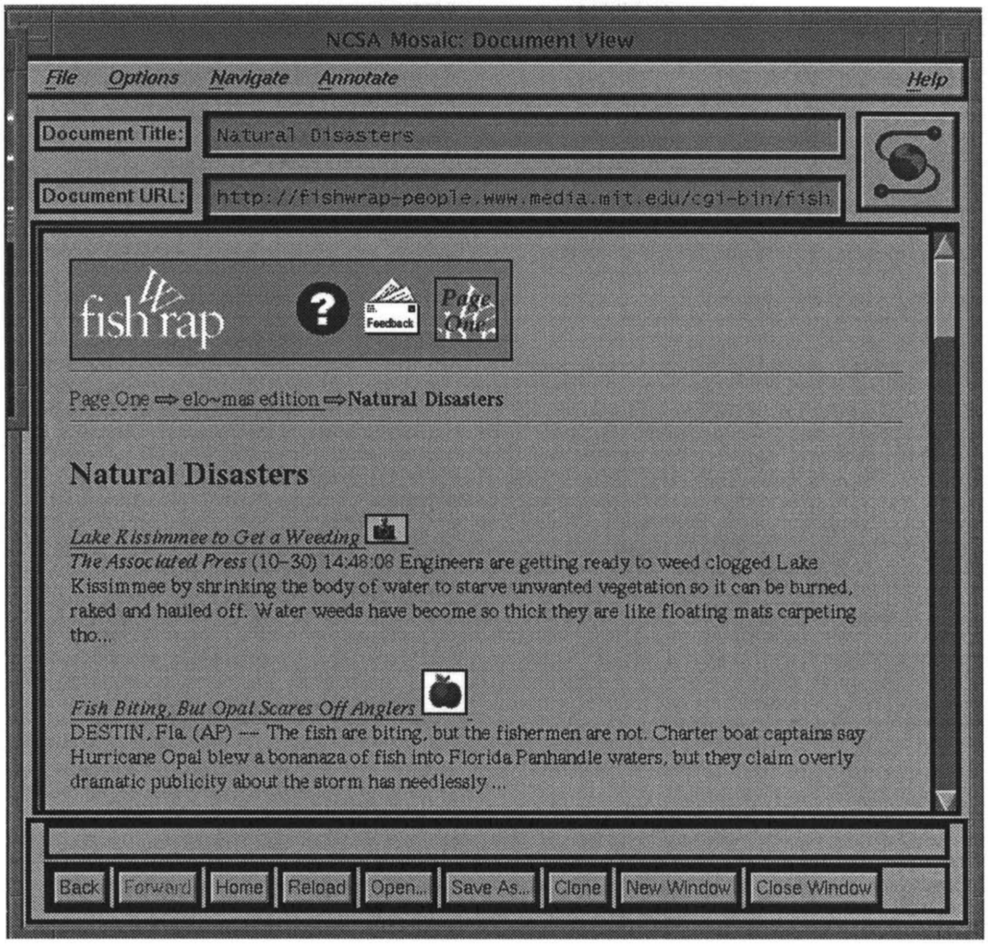
An augmented article appears in Fishwrap with a small PLUM icon next to the title, as shown in Fig.7.2.

When the reader clicks on the title, the browser reveals the full body of the article with augmented words as underlined hyper-links, as shown in Fig. 7.3

When the reader clicks on an augmented word, the augmentation is displayed in a separate html document, as illustrates Fig. 7.4.

The augmentation page consists of three parts. At the top, to remind the reader of the context within the original article, is the augmented word in bold in the surrounding sentence. Below are links to augmentations for the other three home communities. The augmentation itself is divided into a local augmentation relating to the reader's home town and a global augmentation providing background information about the disaster-stricken country.

From the augmentation page, the reader can access the World Fact Book entry PLUM used to generate



Document Title: Natural Disasters

Document URL: http://fishwrap-people.www.media.mit.edu/cgi-bin/fish

Page One ⇨ elo~mas edition ⇨ **Natural Disasters**

Natural Disasters

Lake Kissimmee to Get a Weeding
The Associated Press (10-30) 14:48:08 Engineers are getting ready to weed clogged Lake Kissimmee by shrinking the body of water to starve unwanted vegetation so it can be burned, raked and hauled off. Water weeds have become so thick they are like floating mats carpeting tho...

Fish Biting, But Opal Scares Off Anglers
DESTIN, Fla. (AP) — The fish are biting, but the fishermen are not. Charter boat captains say Hurricane Opal blew a bonanaza of fish into Florida Panhandle waters, but they claim overly dramatic publicity about the storm has needlessly ...

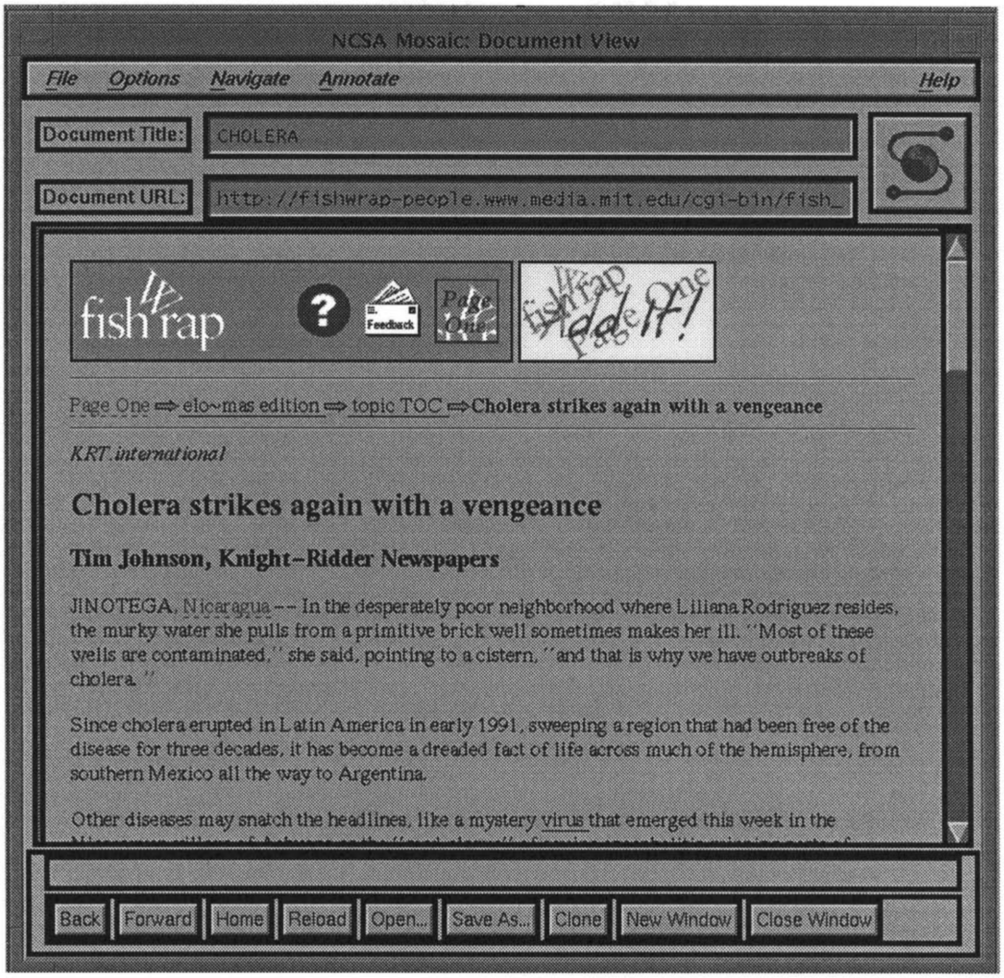
Fig. 7.2

the augmentation. The reader can also access an html form to send feedback.

7.4. How can readers get involved in PLUM?

7.4.1. Feedback

Like letters to the editor in a print newspaper, Fishwrap allows readers to comment on articles. Comments are displayed at the bottom of an article. Readers can also send feedback directly to PLUM.-Readers are invited to send their opinion and to indicate which augmentations they prefer and why. They can also suggest new types of augmentations and expansions. Since no right way exists to explain facts in a news article, readers may not agree with PLUM's augmentations. When the augmentations need improvement, rules in the RuleBase must be modified, added or deleted. Because the rules are implemented in the programming language Lisp, only a programmer is able to change them. PLUM cannot modify its rules without human intervention.



CHOLERA

http://fishwrap-people.www.media.mit.edu/cgi-bin/fish_

Page One ⇨ elo~mas edition ⇨ topic TOC ⇨ **Cholera strikes again with a vengeance**

KRT.international

Cholera strikes again with a vengeance

Tim Johnson, Knight-Ridder Newspapers

JINOTEGA, Nicaragua -- In the desperately poor neighborhood where Liliana Rodriguez resides, the murky water she pulls from a primitive brick well sometimes makes her ill. ''Most of these wells are contaminated,'' she said, pointing to a cistern, ''and that is why we have outbreaks of cholera.''

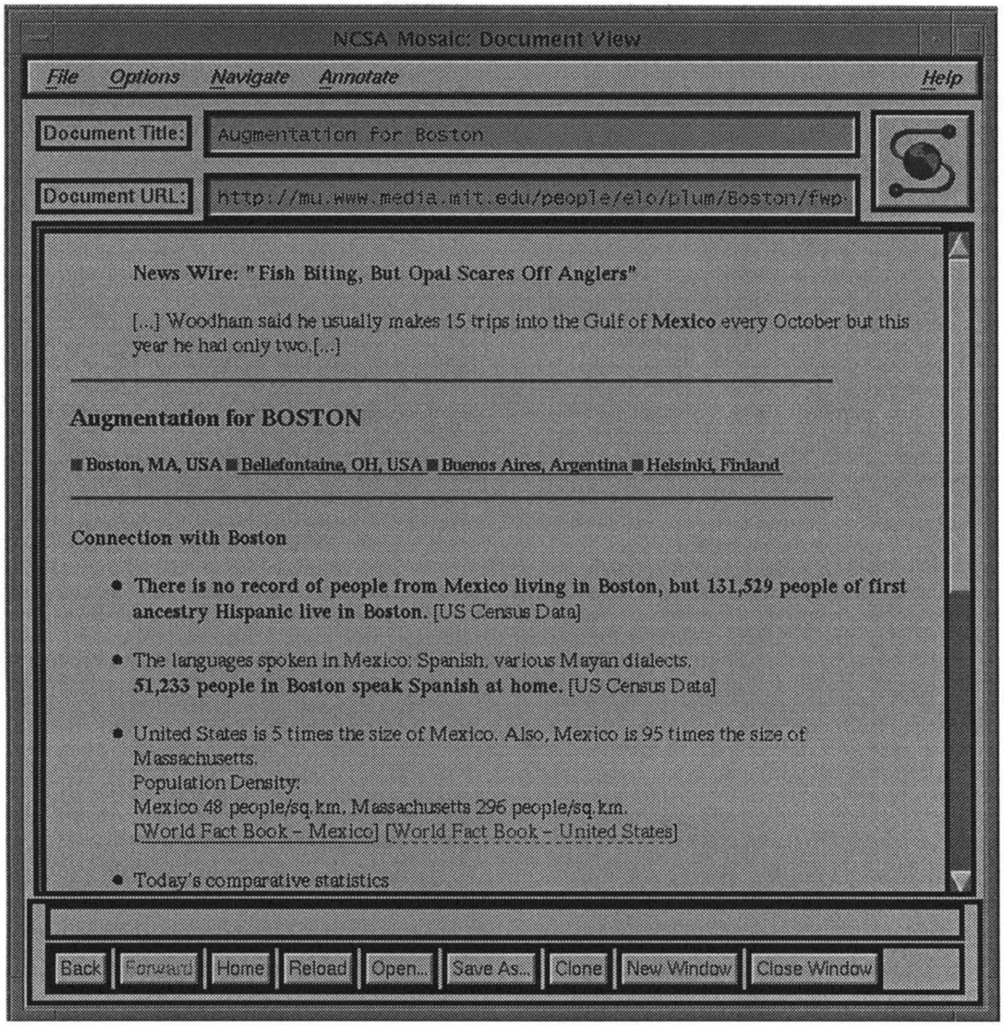
Since cholera erupted in Latin America in early 1991, sweeping a region that had been free of the disease for three decades, it has become a dreaded fact of life across much of the hemisphere, from southern Mexico all the way to Argentina.

Other diseases may snatch the headlines, like a mystery virus that emerged this week in the

Fig 7.3

The reader feedback is summarized in the Chapter 8, Evaluation of PLUM. Unfortunately, the Fishwrap readership has sent little feedback. A reason for the passive response may be the nature of the news PLUM augments. In general, people may not read news on disasters. Furthermore, if a reader does not go beyond the three-line summary to view the full body of an article, he or she does not see the augmentations. In addition, because Fishwrap has not advertised PLUM as a new feature, the augmented articles may go unnoticed.

One person suggested that PLUM automatically process feedback from readers. An html augmentation page could include two buttons, "I like this augmentation" and "I don't like this augmentation", to send PLUM a message. However, such a binary choice does not allow a reader to describe the reasons for liking or disliking an augmentation. Does the reader not like the augmentation in general or for this particular article? Furthermore, PLUM could only react to the message "I don't like this augmentation" by deleting the rule that produced it when rewriting the rule may be sufficient to produce



Document Title: Augmentation for Boston

Document URL: http://mu.www.media.mit.edu/people/elo/plum/Boston/fwp-

News Wire: "Fish Biting, But Opal Scares Off Anglers"

[...] Woodham said he usually makes 15 trips into the Gulf of **Mexico** every October but this year he had only two.[...]

Augmentation for BOSTON

■ **Boston, MA, USA** ■ **Bellefontaine, OH, USA** ■ **Buenos Aires, Argentina** ■ **Helsinki, Finland**

Connection with Boston

- **There is no record of people from Mexico living in Boston, but 131,529 people of first ancestry Hispanic live in Boston.** [US Census Data]
- The languages spoken in Mexico: Spanish, various Mayan dialects.
  51,233 people in Boston speak Spanish at home. [US Census Data]
- United States is 5 times the size of Mexico. Also, Mexico is 95 times the size of Massachusetts.
  Population Density:
  Mexico 48 people/sq.km. Massachusetts 296 people/sq.km.
  [World Fact Book - Mexico] [World Fact Book - United States]
- Today's comparative statistics

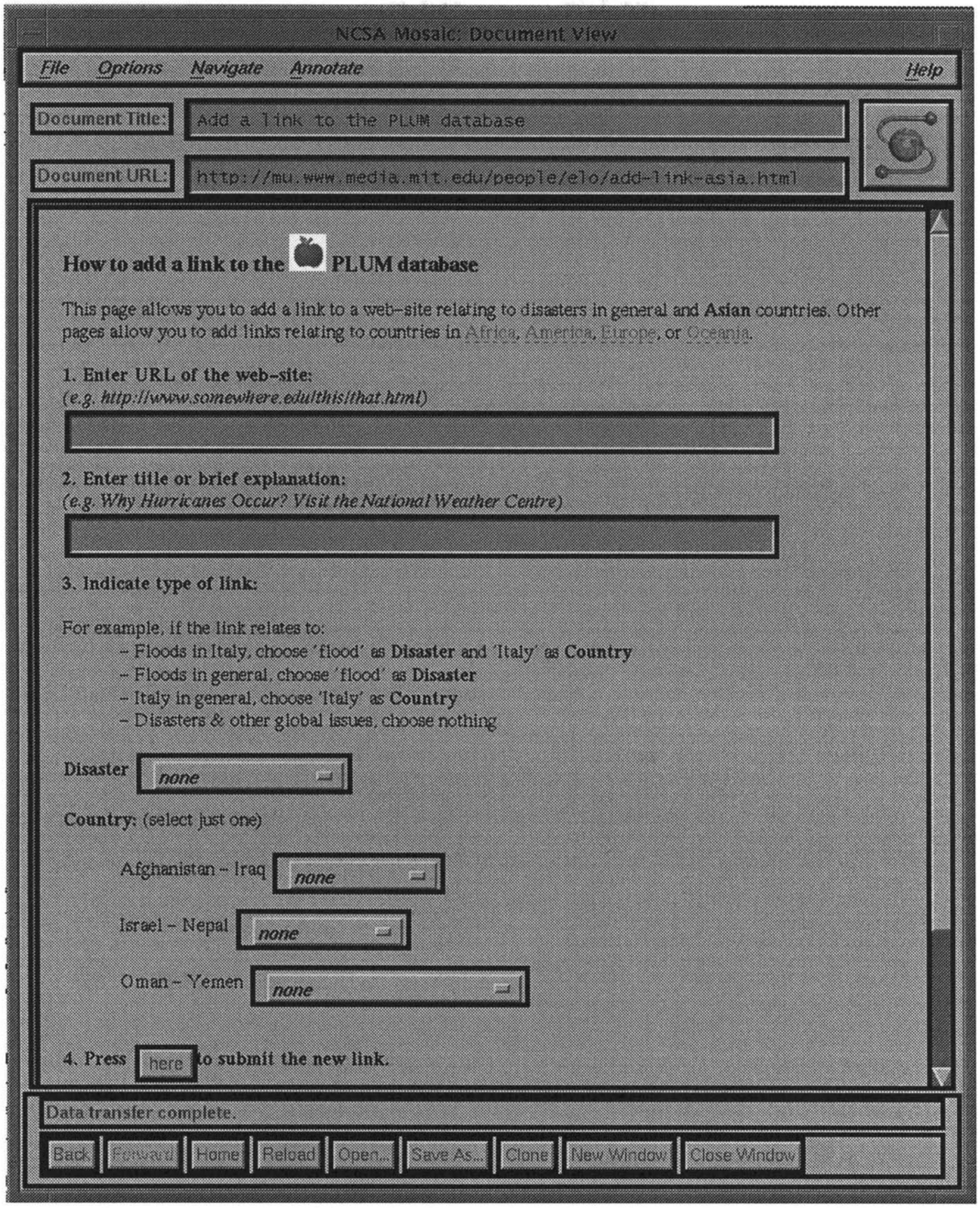
Fig. 7.4

an acceptable augmentation.

7.4.2. Adding related web-sites

Only a few hours after an earthquake struck Kobe in January 1995, discussion groups and web-sites sprang up on the Internet. Since PLUM does not search the Internet for information related to disasters, readers who know of related web-sites and news groups are a valuable source of information. By allowing readers to add web pointers to the PLUM database, the otherwise static PLUM FactBase acquires new information. An augmented article becomes a gateway to related information on the Internet.

A reader can submit addresses of web-sites using the html-form shown in Fig 7.5. The reader chooses one of four options to indicate how the web-site relates to the article. For example, after reading an



Document Title: Add a link to the PLUM database

Document URL: http://mu.www.media.mit.edu/people/elo/add-link-asia.html

How to add a link to the PLUM database

This page allows you to add a link to a web-site relating to disasters in general and **Asian** countries. Other pages allow you to add links relating to countries in Africa, America, Europe, or Oceania.

1. Enter URL of the web-site:
(e.g. *http://www.somewhere.edu/this/that.html*)

2. Enter title or brief explanation:
(e.g. *Why Hurricanes Occur? Visit the National Weather Centre*)

3. Indicate type of link:

For example, if the link relates to:
- Floods in Italy, choose 'flood' as **Disaster** and 'Italy' as **Country**
- Floods in general, choose 'flood' as **Disaster**
- Italy in general, choose 'Italy' as **Country**
- Disasters & other global issues, choose nothing

Disaster *none*

Country: (select just one)

Afghanistan - Iraq *none*

Israel - Nepal *none*

Oman - Yemen *none*

4. Press here **to submit the new link.**

Data transfer complete.

Fig. 7.5

article on floods in Italy, the reader may add a web-site on

- the given natural disasters in the given country, such as "Italian floods this century"
- the given natural disaster in general, such as "How to prepare for a flood"
- the country in general, such as "Historical Sites In Italy"
- issues related to natural disasters in general, such as "Disaster Relief Organizations"

Chapter 8. Evaluation of PLUM

PLUM can be evaluated at two levels

- a quantitative evaluation measures how robust PLUM is as a computer system
- a qualitative evaluation measures how useful it is for readers

8.1. Quantitative evaluation

The success of PLUM depends partly on the performance of the Parser. If the Parser extracts facts accurately from the articles, the Augmenter will augment appropriate concepts. In general, an information extraction program is robust if it accurately processes texts not used as a model in the design. The extraction patterns for PLUM were designed using 50 sample disaster articles. They were tested on 50 other disaster articles. Because some features of a disaster are more difficult to detect than others, the patterns vary in accuracy. Table 1 shows the number of times PLUM correctly extracted a feature, the number of times a feature appeared in the article but PLUM missed it, and the number of times PLUM extracted an incorrect feature.

Table 1:

	% correct of detected	appeared,was not detected	detected, did not appear
COUNTRY-AFFECTED	88%	12%	0%
DISASTER-TYPE	96%	0%	4%
PEOPLE-AFFECTED	96%	0%	4%
PEOPLE-KILLED	64%	32%	4%
DOLLAR-AMOUNTS	96%	4%	0%
FAMILIES-AFFECTED	100%	0%	0%
LAND-AFFECTED	96%	4%	0%
CROPS-ANIMALS-AFFECTED	90%	10%	0%
HOUSES-AFFECTED	94%	6%	0%
DISTANCES	100%	0%	0%

PLUM may incorrectly guess the country affected when an article mentions several country names an equal number of times. When articles report on an overseas department, such as Dutch St.Maarten or Portugal's Azores, PLUM often located the disaster to the mother country, because the overseas departments are not listed in the World Fact Book. In addition to country names, PLUM extracts US states. In some cases when an article reports an event in a large city, such as Chicago, it mentions no country or state. These errors could be avoided if PLUM included a list of all geographic entities.

When an article mentions more than one disaster, PLUM may incorrectly extract the type of disaster. For example, if an article reports a flood but refers several times to the previous year's drought, PLUM may augment the article as a drought instead of a flood.

As Table 1 shows, the patterns for detecting numbers of people killed are the least accurate. This is normal. The patterns for extracting numbers of houses, areas of land, or numbers of people, are simple constructs with two elements, a number modifying a noun. The patterns for extracting numbers of people killed are constructs with three or more elements, as shown in Section 6.1.2. A death can be reported using many different words and expressions. The patterns synthesize the most frequently used wordings in English language disaster news. If an article uses an unusual wording, such as in the sentences below, the Parser fails to detect the numbers in bold, because it cannot resolve what the number quantifies.

> Of the dead, **33** were from Negros Occidental.
>
> Cholera has killed 45 in the Hemisphere, including
> **30** in Nicaragua.
>
> The death toll in a flash flood in a western Turkish
> town rose to **70** on Wednesday.

Land areas are not detected when expressed in a non-standard way, such as

> 100-ft-by-100-ft area

Some of the errors where a feature was detected without being reported could be eliminated by preprocessing of the text. Some multi-word proper names such as `The Philippine Air Force` should be made into one entity, so that `Philippine` is not taken to describe a country. Also, the part-of-speech tagger occasionally erroneously glues two consecutive numbers into one, such as 17,300 extracted from

> On August **17, 300** people were evacuated.

8.2. Erroneous augmentations

PLUM's Motif interface could be developed further to allow the edition of generated augmentations. A person should be able to reject or modify the generated augmentations. However accurate PLUM's parsing or augmenting rules, unpredictable turns of phrases can result in erroneous parsing and nonsensical augmentations. For example, the following sentence could result in an erroneous augmentation:

> `Patients are flooding the hospitals in Dhaka, as the epidemic continues to spread.`

Because '`flooding`' is a synonym of the keyword '`flood`', PLUM extracts '`flooding`' and '`epidemic`' as candidates for the feature TYPE-OF-DISASTER. Let's suppose the rest of the article does not mention any more disaster keywords. Since both '`flood`' and '`epidemic`' occur once, TYPE-OF-DISASTER is set to '`flood`' because it occurs first. Hence, PLUM erroneously augments the article with references to the history of floods in Bangladesh.

Even more nonsensical augmentations could result when an article uses a natural disaster as a metaphor to describe an event unrelated to disasters:

> `Apple Computer Inc., taking advantage of the lull before Hurricane Windows strikes the computer industry with full force later this month, plans to introduce three new models in its desktop Power Macintosh product line today...[New York Times, Aug 7, 1995. p.D4]`

As mentioned earlier, PLUM tests whether an incoming article is truly about a natural disaster. If an article mentions the disaster too few times considering its length, PLUM rejects it. PLUM also rejects an article if it reports less than four features in the disaster template. An article that employes a disaster metaphor rarely reports other disaster features, such as people, land, or houses. Thus, PLUM rejects the Apple Computer article because it only fills two features in the disaster template, COUNTRY-AFFECTED, United States, and TYPE, hurricane.

Currently, PLUM augments all numbers that quantify people. For example, in

> **`Fifteen`** `children were evacuated from the roof of the school building.`

PLUM augments 15, even though it is not a large number difficult to understand. Depending on the nature of the complete article, augmenting 15 may be sensible. Currently, PLUM does not discard any numbers because an absolute threshold may not apply to all articles.

8.3. Qualitative evaluation

Unlike the Parser, the rest of the PLUM system cannot be evaluated by counting errors. Since PLUM explains disaster news to readers, their reactions are important. People can easily judge if an augmentation is sensible or not. Below are selected comments from Fishwrap readers, Media Lab students and faculty:

Q: I prefer to see `'30000 is 1 out of 15 people in Boston'` over `'30000 is all the people living in Waltham'`. I have some sense of the size of Boston and I can always imagine a group of people and think of 1 out of 15 people. I have no idea about the sizes of other cities around Boston so you would be imposing on me an image which is not meaningful.
A: PLUM now compares numbers of people to the home town population as well as to another local group of people approximately the same size.

Q: How can I make PLUM relate news to New York City? It would make more sense to me than Boston.
A: PLUM does not contain a description for all cities in the US. However, a mechanism exists to add new cities compiled from Census Data, if there is sufficient interest.

Q: Why not maintain a history of the augmentations and vary them from one time to another?
A: PLUM now keeps track of the comparative statistics it generates on the home site and the disaster-struck country. The statistics permute, such that two articles reporting from the same country contain different comparisons.

Q: I'd like to see the source PLUM uses to generate the explanations.
A: The World Fact Book is now in HTML form document in the PLUM web-site, accessible directly from the augmentations. The CRED database is not public and cannot by put on-line.

Q: Readers should be able to get involved somehow.
A: It's true, the web allows readers to be active without much effort. In addition to the Fishwrap commenting facility, readers can add pointers to web-sites that relate to PLUM articles. The submitted pointers become part of PLUM's FactBase.

Q: You should add basic definitions of all the natural disasters.
A: The Federal Emergency Management Agency web site with the definitions is linked to all augmented articles.

Q: There is no absolute truth. Many different augmentations should be chosen.
A: PLUM generates most of the time more than one explanation for an augmented item.

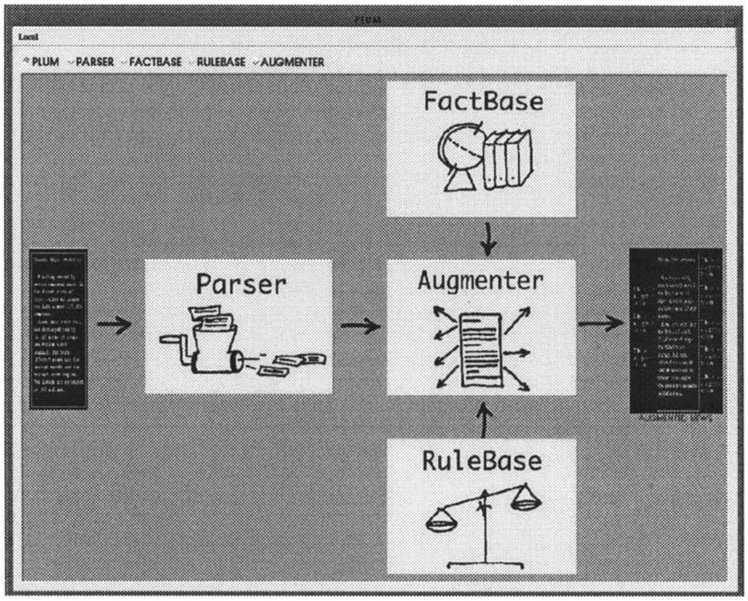
Q: How to enlarge to other domains, or 'unnatural disasters' such as a civil strives or technological disasters?
A: If we start to enlarge the domain, we may want to include an editor in the process. The system could propose appropriate resources and possible augmentations, and the editor would select the final output.

Q: I think an editor would have to check the links and the contents of augmentations. Does PLUM include a program to do that in a simple and fast way?

A: PLUM has an interface for modifying the augmentations. However, since the emphasis has been mostly on readers, the interface would have to be developed further before taking it to a newsroom.

Chapter 9. PLUM Icons

The Motif Interface is useful for demonstrating the project for visitors. It illustrates the relationship of the different components in PLUM. The interface opens to a stylized flow chart of the system:

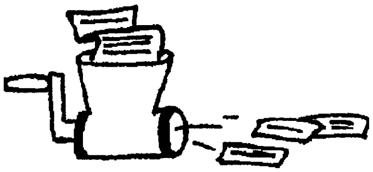


The flow chart illustrates how PLUM processes an article. In contrast with the digital PLUM project, I drew the icons and the arrows by hand. I wanted the flow chart of the system to look somewhat 'sketchy' and hand-made. The design of the icons was my final project for the Visual Design Workshop at the MIT Media Lab in Spring 1995.

A logo conveys the image of a company or a product and should clearly illustrate its function. A logo should not convey a different or contradictory message to different audiences. Matt Mullican witnessed how the meaning of his artwork changed, as he hung a flag in yellow and black on a museum in Brussels. He was unaware that the colors symbolize the Flemish Nationalists. The meaning of a sign depends on the cultural context. [Blauvelt]

Working toward universal logos, I used familiar objects to illustrate the functionality of each PLUM component. 'Filtering' of news streams brings to mind a colander. Because the Parser analyzes incoming news wires and extracts features from them, a colander is too static an object. I chose the meat grinder with extracted pieces of news flying out of it.

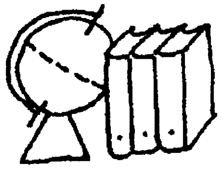
Parser



The working meat grinder characterizes the Parser as a dynamic process. However, the critique session of the Visual Design Workshop revealed to me that a meat grinder is not a universally known object. A Japanese student, accustomed to kitchen utensils for vegetarian diets, had never seen a meat grinder.

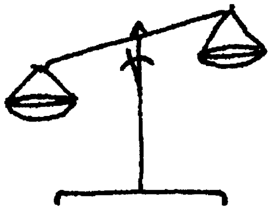
The FactBase icon consists of books and a globe. It is a fairly obvious representation for the collection of background resources, data bases and maps. This logo looks static and illustrates that the databases do not evolve over time. The workshop participants agreed that it was an appropriate representation for the FactBase.

FactBase

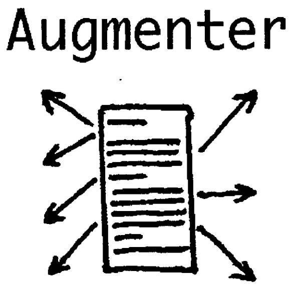


Several objects came to mind for depicting rules and standards contained in the RuleBase. A ruler sounds the same as 'rulebase'. Symbols used in law books, §§§, are appropriate but not familiar to everyone. I settled for the scale, because it suggests measures and standards.

RuleBase



Since augmenting an article is not a common concept, it is difficult to represent with a familiar object. A magnifying glass suggests adding detail, but does not illustrate the diverse ways in which augmenting occurs. To create a dynamic logo, I drew arrows flying out from the article.



Chapter 10. FactBase: Public or Collected Data?

If someone asked me where Fenway Park is located, I would not be able to say exactly. But I would remember a Sunday afternoon on the Green B line subway, when a wagon-full of sports fans got off...

The success of PLUM's augmentations depends on the nature of the data in the FactBase. After all, augmentation attempts to relate unfamiliar facts to well-known concepts in a geographic community. But what type of knowledge is shared by all residents of a community? What do people, regardless of occupation, income or level of education, gender, origin or age, know about their city? People learn facts about their home town explicitly when reading news or in a geography class at school. But mostly they acquire knowledge implicitly while interacting with others and going through daily activities.

10.1. The Image of a City

A part of people's knowledge about their city is their mental representation of the geographic area, or a *cognitive map*. In The Image of the City, Kevin Lynch studied what makes up a cognitive map. [Lynch 1960] Through extensive interviews, he collected cognitive maps from residents of three American cities, Los Angeles, Boston, and Jersey City, NJ. Comparing the mental maps suggests that the image of the same city varies significantly from person to person. Holahan [Holahan 1978] studied how cognitive maps are distorted in relation to a person's daily trips to work and home. If a person frequents an area, the size of the area is exaggerated in the person's mental map. In addition to individual residents' mental maps, a city also possesses a public image, a common mental picture carried by a large number of its inhabitants. The public image of a home community can suggest what to include in the PLUM FactBase.

Lynch offers interesting explanations for Boston's exceptionally vivid public image. Unlike Los Angeles with its many centres, Boston is the distinct core of the greater Boston area. While streets in Jersey City are all alike and are recognized only by the street sign, streets in Boston's neighborhoods have distinctive flavor. They are organized in a broad grid, a narrow grid or no grid at all. They also have contrast in age. Because Boston sits on a peninsula, the city has distinct borders with the sea and Charles River. It also has well-defined core, Boston Commons. Because people can see the Boston skyline from MIT campus, across the Charles River, it is easy to position buildings within the whole.

The public image of Boston in a computer program should encode discrete items that people remember about the city. Lynch's research demonstrates that people remember their city in terms of paths and districts, among other things. Paths are the predominant city elements people remember. However, a path is only memorable if it seems continuous. For example, Washington Street in Boston is well known around Filene's shopping area. Many people don't make the connection to the street in South End. Similarly, Causeway, Commercial Street and Atlantic Avenue are not perceived as one path because of the changes in name. However, Massachusetts Avenue is a long path traversing Boston and

Cambridge, while Storrow Drive follows the river front. People also remember paths from one landmark to another, such as Cambridge Street from Charles Street round-about to Scolley Square.

Lynch's study provides guidelines for encoding into PLUM city paths that minimize the differences in people's mental maps. When PLUM augments a distance reported in an article, it refers to a frequently traveled continuous path of the same length. It provides landmarks on the way between the origin and destination of the path. In automobilists' paths with one-way streets, it takes into account the direction. Despite these precautions, people may still perceive the distance of a path differently. Distance is subjective and depends on factors such as mood, weather, mode of transport, traffic. A map of the city with the path highlighted would probably be effective. However, it would require manually creating a map for each selected path.

According to Lynch, people also remember districts in their city. Districts are relatively large areas with a common characteristic: building type (Beacon Hill), use (Downtown shopping district), degree of maintenance, inhabitants (China Town). However, the boundaries of a district are often imprecise. People know Back Bay is bordered on two sides by the parks and Charles River, but the other two sides are fuzzy. Highlighting a district on a map would be clear, but creating individual maps would be tedious. Because people understand sizes of named districts differently, PLUM does not use districts to explain the size of an affected area. Instead, it overlays on the local map a shadowed circle the size of the affected area. A map is a conventional graphic representation most people know how to interpret. Seeing the overlay will help to understand the scale of the land affected. To make the overlay meaningful to a maximum of people, the shadow is centered on downtown Boston. It is a familiar reference point to most residents of Greater Boston.

10.2. PLUM's Knowledge of Boston Paths

The paths through Boston were encoded in PLUM using landmarks in the city. Curious to see whether landmarks compiled by Kevin Lynch were valid after 35 years, I asked five MIT Media Lab students to list their landmarks in Boston. Foreseeing the integration of PLUM into Fishwrap, this allowed me to see if an MIT student's mental landmarks overlap with that of Kevin Lynch's interviewees.

10.2.1. A questionnaire to satisfy my curiosity

Students Nicolas, Josh, Jeff, Deb and Earl listed different types of landmarks as defined in Webster's Dictionary:

1: a mark for designating the boundary of land

(Rivers, parks, hills, mountains, shore-lines are such geographical landmarks) The students listed

Boston Commons, Charles River, Atlantic Ocean, Beacon Hill, Boston Harbor, Fresh Pond, the Esplanade

2a: a conspicuous object on land that marks a locality

2b: an anatomical structure used as a point of orientation in locating other structures

The students listed the clock tower, Faneuil Hall, Longfellow bridge, Harvard bridge, Prudential Centre, the CITGO advertisement sign, the golden dome in Beacon Hill, Fenway Park, Bunker Hill, Logan Airport Tower as conspicuous objects in Greater Boston and the Green Building and the MIT Dome on MIT campus. Structures they use as a points of reference include Park Station T, Kendall Sq T, Central Sq T, Out-of-town News at Harvard Sq, the John Hancock building, Star Markets, South Station, Boston Public Library and Haymarket in Greater Boston, and the Student Centre, the Muddy Charles Pub and the Media Lab Lobby on MIT Campus.

3: an event or development that marks a turning point or a stage

Because four of the five students were from out of town, they listed significant events not necessarily specific to Boston. One could refer to such an event with "Remember when...?" and others who were present would recall it. The students listed extreme weather conditions (storm, heat, cold, "it snowed egg-sized hail", blizzard of '78), elections, referendums, a change in political leaders, a major local crime (murder), an earthquake, a flood, a sports team victory, US Hockey Team Olympic Championship, Super Bowl 1981, extreme traffic jam, July 4th fireworks.

In addition, the students colored in paths they frequently travel in Boston. The MIT students' maps overlapped to a large degree over downtown Boston, even though they live in different parts of Greater Boston.

Exploring the possibility to include local celebrities into the Boston knowledge base, the questionnaire asked to name well-known people. They listed Marvin Minsky, Seymour Papert, Noam Chomsky, Nicholas Negroponte, Mitch Kapor from the MIT Community (with an obvious Media Lab bias), and Tip O'Neal, Mike Dukakis, Neil Rudenstein, Mayor Menino, Governor Weld, Ted Kennedy from Boston.

10.2.2. Boston Paths

This rather simple questionnaire helped to come up with an initial set of frequently traveled paths in Boston:

"Going down Newbury Street from the Parks to Mass Ave" (1 mi)

"Following Storrow Drive on the bank of Charles River, from the BU bridge past the CITGO sign and the Esplanade until the Longfellow Bridge" (2.5 mi)

"Driving from Logan Airport through Sumner tunnel, past Haymarket, across the Longfellow bridge to Kendall Square" (5 mi)

"Riding the blue line T from Government Centre to Revere Beach" (5 mi)

"Going down Massachusetts Avenue from Symphony Hall, across Harvard Bridge, past MIT, Central Sq, Harvard Sq, Porter Sq to Davis Sq." (10 mi)

"Going from Boston to Concord and Walden Pond" (15 mi)

"Driving on 1A from Boston Centre to Salem" (17 mi)

"Going from Boston northwest to Nashua, NH." (40 mi)

"Going southwest from Boston to Hartford, CT." (102 mi)

"Driving from Boston to Provincetown at the tip of Cape Cod" (118 mi)

"Driving across the state of Massachusetts from Boston to Pittsfield at the western border of the state" (137 mi)

"Driving West from Boston to Albany, NY." (167 mi)

"Distance from Boston to New York City." (211 mi)

"Distance from Boston to Buffalo, NY." (463 mi)

"Distance form Boston to Chicago." (1015 mi)

"Distance from Boston to Orlando." (1285 mi)

"Distance from Boston to Seattle" (3088 mi)

10.3. Decision on PLUM's Knowledge

Studying Lynch's experiments revealed to me that residents' common knowledge about their city is not easy to compile. No recipe applies to all cities. The knowledge is constructed in people's minds as they interact with each other and go about their daily business in the city. The landmarks and paths people recall cannot be found in books. Although a tourist guide describes a city's famous monuments, residents may not use them as landmarks to orient themselves in the city. The monuments may go unnoticed during the daily trip to work. On the other hand, a vivid landmark to a resident, such as the obnoxious CITGO advertisement board, is not listed anywhere as a landmark of Boston. A physical object or a building may be distinct for reasons unpredictable even by its architect. One-way streets and public transportation routes also dictate how people think about their city. Only a local person would be able to pin-point meaningful information.

In addition to being difficult to collect, this kind of information is tedious to compile. To help in the process, PLUM contains a function to add descriptions of frequently traveled paths. PLUM saves the descriptions and uses them in subsequent augmentations to explain distances reported in disaster articles. In addition to geographic features, important current and past events in the city, local celebrities, and people's life-styles contribute to residents' common knowledge about their city. The small questionnaire illustrates the kinds of events and people MIT students consider to be landmarks. It confirms that natural disasters and extreme weather conditions are well remembered.

The investigation described in this chapter revealed the complexity of the data. Encoding representations of events and people requires encoding knowledge about the world. The computer system needs to be told how to use the information in augmentation. When an avalanche strikes Como, in Italy, how to tell Bostonians that a well-known public figure was born in the city? Or that the stone plaques on a famous landmark were imported from Como? Clearly, the computer would have to do sophisticated reasoning to come up with such complex explanations.

In conclusion, compiling local knowledge bases for the home communities was not feasible for the PLUM project. The question posed at the beginning of this chapter changes. "What type of knowledge is shared by all residents of a community?" becomes "How can PLUM take advantage of publicly available information on geographic communities?"

Chapter 11. Related Work

While many systems that analyze news seek understanding to better *retrieve* or *classify* articles, PLUM uses its understanding of news to *generate* and *explain*. The PLUM system is difficult to categorize, because it uses techniques from several fields of computation. While PLUM does not improve existing techniques, it integrates them to work together in one system. But the goal of PLUM is not only to see the different techniques cohabit one system. Its goal is to make readers understand news better. Because its success in improving understanding of news is difficult to measure, no obvious way to compare PLUM to other text-analysis systems exists. This chapter describes how the techniques PLUM uses situate within the fields of natural language understanding and generation.

11.1. Natural Language Understanding

The natural language processing community can be roughly divided into two camps. Researchers such as Gerald Salton use statistical techniques in processing text [Salton 1983]. For example, word frequencies and co-occurrences help to discover clusters of related documents. They also help determine if a document is a summary of another. While statistical techniques generalize to some extent across language boundaries, the other approach to natural language processing relies on the grammar and vocabulary of a specific language. The second approach grew out of the Artificial Intelligence community. The so-called AI natural language processing of the 70's and 80's aims to understand the detailed story told in text: who did what, when, where? [Schank 1977] Such story understanding is evaluated by asking questions to see if the system can draw inferences. AI natural language processing incorporates parsers or part-of-speech taggers to recognize the grammatical role of words.

Recent work in story-understanding seeks to accomplish specific tasks [Jacobs 1993]. A text interpreter should be able to accurately extract pre-defined facts from text on a given topic. Ideally, such a system could enter facts directly into a database from large collections of text on a constrained domain. PLUM belongs to this class of text processing systems.

The first text interpretation system, FRUMP, skimmed and summarized news on world events, including news on disasters [DeJong 1979]. FRUMP employed extensive knowledge about the topic of the text to improve its analysis. It used sketchy scripts to organize the knowledge on 60 different situations. A sketchy script described the most likely events reported in an article. While processing text, FRUMP's text analyzer tried to find these predicted events. In this way, FRUMP extracted the main thrust of a news story, but was incapable of processing details. While FRUMP looked for events such as 'the police arrested the demonstrators,' PLUM looks for data to fill the disaster template. Like FRUMP, PLUM does not understand complete articles. It only detects the ten features in the disaster template, the characteristic facts reported in a disaster article. Several systems like FRUMP competed against each other in the Message Understanding Conferences (MUC) in the late 80's and early 90's. The systems were to interpret texts on topics such as terrorism and military reports as fast and accurately as possible.

AI NLP also inspires some of the research on text-analysis in the Machine Understanding Group at the MIT Media Lab. Professor Ken Haase's on-going research [Haase 1995] processes text in order to match analogous sentences or phrases. Haase's system implements multi-scale parsing to process archives of news articles. It constructs a representation of the content based on each word's class, its role within the sentence and its entry in the digital thesaurus WordNet. It discovers, for example, that the actions in the phrases "Clinton named William Perry to be his secretary of defense" and "Jocelyn Elders was chosen by President Clinton to be the new surgeon general" are analogous.

The pattern matching techniques in PLUM resemble those used by another Machine Understanding Group project, SpinDoctor [Sack 1994]. Warren Sack's system analyzes news from Central America to detect the point of view of the author. It looks for patterns of words that reveal how an actor in the news is depicted. For example, if the government is described as 'criminal', the article is written from the point of view of the guerillas. PLUM and SpinDoctor both employ knowledge about the topic they analyze.

11.2. Natural Language Generation

Natural Language Generation is a field of joint research for linguists and computer scientists. Their goal is to design computer programs that generate grammatically and stylistically correct prose.

Natural language generation is often divided into two parts, "what to say" and "how to say it" (e.g. [Dale 1990]). First, a generator determines the content of text and its degree of detail. Then, it determines what words to use and in what order or style to present the content. An opposing view (e.g. [Appelt 1985]) criticizes this sequential processing, arguing that the two parts cannot be separated. For instance, the style of the presentation can also influence the choice of the content.

The simplest kind of natural language generation uses templates to produce text. SportsWriter, a basketball reporting program, generates sports articles by inserting player statistics into pre-composed sentence templates. It also uses direct quotes from observers of the game. A little variation in the choice of words makes for a relatively life-like effect. The SportsWriter benefits from the fact that sports reports in general have a repetitive style.

Automatic summarization of documents is another type of language generation. Instead of producing original text, most summarization techniques exploit the structure of a document to recognize key phrases in a document. The key phrases are then presented to the reader. [Salton 1983]

A sophisticated natural language generator varies its output depending on the context. According to the Register Theory within Systemic-Functional Linguistics [Halliday 1985], language varies along three dimensions:

field: the subject matter
mode: the purpose of the text, e.g. reference vs. tutorial
tenor: the relationship to the audience

A language generation system usually varies text along just one of the three dimensions. For example, an expert system varies its tenor by adapting to the user's level of knowledge of the subject matter [Paris 1993]. Augmentation varies text in more than one way. Augmenting a disaster article expands the field, because it includes background information on the history of disasters, the disaster-struck community and the reader's home community. Augmented text also alters the mode, stereotypical reporting of disasters, because references to the home community make the style more personal. An article written to a global audience acquires a more local tone.

For effective prose, language generation systems have to assure a rhetorical structure for the text [Dale 1990]. When generated text exceeds a few sentences, the system needs to present the points in a logical sequence, with a beginning, middle and end. PLUM avoids having to assure rhetorical structure because it does not rewrite the original story. It simply generates short sentences from templates and adds them as annotations to the original text. A simple augmentation becomes meaningful in the context of the article. The reader makes the connection between the two. In fact, linking the original text to an augmentation suggests a more sophisticated understanding and generation than is the case.

Chapter 12. Future Work and Development

12.1. Augmenting Beyond Disasters

Presently, PLUM only augments disaster articles. Such news lends itself to this kind of annotation. Its stylized reporting allows a fairly accurate automatic processing. Other stylized news topics PLUM could potentially augment include finance, sports and elections. All of these topics present stereotypical actors and situations. For each topic, a number of sample articles should be examined, to determine the characteristic features. Based on the recurring ways the features get reported, patterns could be constructed.

While a part of PLUM's RuleBase is specific to disasters, some rules apply to any domain. For example, a sports article might report the following:

```
10,000 fans were gathered at Fenway Park on Sunday
afternoon.
```

Since 10,000 quantifies a group of people, PLUM could augment it with an existing rule and compare it to the home town population. Similarly, the overlay of a shadow on the home town map is an effective way to illustrate the size of any area of land reported.

Using the existing parsing patterns, rules, and the extensive geographic data in the FactBase, PLUM can augment the location of any foreign news. Highlighting similarities between a distant country and the home town and country of the reader could be useful in contexts other than disasters.

If PLUM analyzed a new topic, new rules could be added into the rule base. For example, an article may report the result of an election.

```
Democrats won 72 of the 100 seats in the Senate.
```

A rule could define how to search a database on the history of elections for the last time the Democratic party won with this large a majority. Or it could pull up the names and e-mail address of the democratic senators in the home state.

In conclusion, PLUM can augment other stylized domains if patterns for extracting the characteristic items reported and domain-specific rules for augmenting are added. Some of the existing PLUM extraction patterns and rules may also apply to other domains, as shown in the examples above.

12.2. Global vs. Local Context

Augmentation mainly attempts to provide a *local* context for understanding news reports. Explaining the *global* context may also be helpful. When PLUM generates comparative statistics between two countries using the World Fact Book, it also shows the countries with the highest and lowest values for each statistic:

```
Japan - USA comparison
total area:
Japan: 377,835 sq km'
United States: 9,372,610 sq km
Most in the world: Russia 17,075,200 sq km
Least in the world: Vatican City 0.44 sq km
```

This is helpful in understanding the range of possible values. Providing a global context could be taken further in a digital encyclopedia where each entry could be examined within the context of the rest of the world as well as the context of the home community. Viewing the same information against different backgrounds illustrates that facts are relative and sensitive to context. Such juxtapositions may lead people to become more critical of numbers as truths.

12.3. Visualizing Augmentations

An augmented article is a suitable object for graphical visualization and dynamic layout. Because augmented articles consist of several levels of information, the original body, the augmentations, and the background resources, the PAD system could display them elegantly [Perlin 1993]. PAD allows defining different layers of text and zooming between the different layers to view more or less detail.

Positioning the augmentations in the margins of the original text, presents an interesting task of optimizing lay-out. [Ishizaki 1995] Suguru Ishizaki's agent-based approach to dynamic design could be applied to this problem. Ishizaki presents a theory whereby all objects in a layout possess local behaviours. Each object positions itself within the whole as best as it can. From the local interactions between neighboring objects emerges a global solution for the layout. If the body of an article and each augmentation were defined as objects, shifting the perspective from one home community to another would force the new set of augmentation objects to dynamically find a suitable layout.

Chapter 13. Conclusion

The computer system Peace Love and Understanding Machine, PLUM, *augments* news on natural disasters. By explaining reported facts in terms of a reader's home community, PLUM adds a context that helps the reader better understand disaster news. Augmenting is a new approach to tailoring digital news, since PLUM adds explanations to existing articles. Because personal profiles are difficult to maintain and necessary to protect, PLUM uses profiles of geographic communities. A profile compiled from publicly available data on a community enables PLUM to make news more informative and relevant for all residents of the community.

In order to augment text, PLUM integrates techniques from several fields of computation. The PLUM Parser uses part-of-speech tagging and pattern matching to analyze news wires. Because PLUM concentrates on one domain, disaster news, the Parser extracts with satisfactory accuracy the characteristic facts reported in articles. Relying on a frame-based knowledge representation, the PLUM FactBase cross-indexes three databases serving as background information for augmentations. Using the rules defined in the PLUM RuleBase and template-based language generation, the PLUM Augmenter produces the augmented articles as hyper-text documents for the MIT community on-line newspaper Fishwrap. The strength of PLUM is to combine all these techniques in order to improve a reader's understanding of news.

In Fishwrap, readers click on highlighted words to reveal informative augmentations that place statistics in a familiar context. Augmentations can also be viewed from the perspective communities other than one's own home town. Currently, Fishwrap readers can use PLUM to read news augmented for Boston, Massachusetts, Bellefontaine, Ohio, Buenos Aires and Helsinki. Because PLUM supports adding home communities, more cities can be added if there is sufficient interest. Because PLUM's articles are on the World Wide Web, readers can easily contribute information and feedback. Readers are encouraged to add pointers to web-sites relating to the articles. This way, the otherwise static PLUM FactBase is continuously growing.

Because breaking disaster news is often presented without context, it is important to augment such news. Readers tend to generalize from the drama portrayed by disaster articles. By providing a scale for understanding the scope of a disaster, PLUM contributes to a more realistic image of the disaster-stricken country. PLUM also demonstrates that, within restricted domains, a computer program can expand on and localize news written for a global audience. With the increase in digital archives of information, computer systems that help editors and readers are becoming necessary. While the creation, assembling, sorting, archiving, search, and delivery of unrestricted information cannot be fully automated at this point, a computer program with a limited knowledge of the content or a domain can contribute to these tasks.

References

[Appelt 1985] D. Appelt. *Planning English Sentences.* Cambridge University Press, 1985.

[Blauvelt] Andrew Blauvelt. *Cultures of Design and the Design of Cultures.*

[Benthall 1993] Jonathan Benthall. *Disasters, Relief and the Media.* I.B.Tauris & Co, London, 1993.

[Cable & Broadcasting 1994] Cable & Broadcasting, Issue of October 31, 1994.

[Cate 1993] Fred H. Cate. *Media, Disaster Relief and Images of the Developing World.* Publication of the Annenberg Washington Program, Washington D.C., 1993.

[Chesnais 1995] Pascal Chesnais, Matthew Mucklo, Jonathan Sheena. *The Fishwrap Personalized News System.* Proceedings of the IEEE Second International Workshop on Community Networking, 1995

[Dale 1990] Robert Dale, Chris Mellish, Michael Zock (eds). *Current Research in Natural Language Generation.* Academic Press, 1990.

[DeJong 1979] Gerald DeJong. *Script application: Computer understanding of newspaper stories.* Doctoral Thesis, Yale University, New Haven, 1979.

[Haase 1993] Ken Haase. *Multi-Scale Parsing Using Optimizing Finite State Machines.* ACL-93 Proceedings, 1993.

[Haase 1995] Ken Haase and Sara Elo. *FramerD, The Dtype Frame System.* MIT Media Lab internal report, 1995.

[Haase 1995a] Ken Haase. *Analogy in the Large.* SIGIR'95 Proceedings, 1995.

[Halliday 1985] M. Halliday. *An Introduction to Functional Grammar.* Cambridge University Press, 1985.

[Holahan 1978] Charles J. Holahan. *Environment and Behavior, a Dynamic Perspective.* Plenum press, New York, 1978.

[IDNDR 1994] The Media Round Table, World Conference on Natural Disaster Reduction, Yokohama, Japan, May 1994.

[Ishizaki 1995] Suguru Ishizaki. *Typographic Performance: Continuous Design Solutions as Emergent Behaviours of Active Agents.* PhD Thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, 1995.

[Jacobs 1993] Paul S. Jacobs, Lisa F. Rau. *Innovations in Text Interpretation.* Artificial Intelligence 63, pp. 141-191, 1993.

[Lashkari 1995] Yezdi Lashkari. *Feature Guided Automated Collaborated Filtering.* SM Thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, 1995.

[Lenat 1990] D.B. Lenat and R.V. Guha. *Building Large Knowledge Based Systems.* Addison-Wesley, Reading, MA, 1990.

[McQuail 1987] Denis McQuail. *Mass Communication Theory.* Sage Publications, 1987.

[Miller 1990] George Miller. *WordNet: An On-line Lexical Database.* International Journal of Lexicography, 3(4).

[Morrison 1982] Philip and Phylis Morrison, the Office of Charles and Ray Eames. *Powers of ten: a book about the relative size of things in the universe and the effect of adding another zero.* Redding, CO: Scientific American Library; San Francisco: Distributed by W.H. Freeman, 1982.

[Paris 1993] Cecile L. Paris. *User Modeling in Text Generation.* Pinter Publishers, UK, 1993.

[Perlin 1993] Ken Perlin, David Fox. *Pad, an Alternative Approach to the Computer Interface.* Computer Graphics Proceedings, Annual Conference Series, 1993.

[Rynn 1994] J. Rynn, J. Barr, T. Hatchard, P. May. *National Report 1990-1994, Australia, International Decade for Natural Disaster Reduction.* Pirie Printers Pty. Ltd., Australia, 1994.

[Sack 1994] Warren Sack. *Actor-Role Analysis: Ideology, Point of View, and the News.* SM Thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, 1995.

[Salton 1983] Richard Salton. *An Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[Scarry 1966] Richard Scarry. *Richard Scarry's Storybook Dictionary.* Dean Publishers, 1966.

[Schank 1977] R.C. Schank and R.P. Abelson. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum, New Jersey, 1977.

[Schank 1990] Roger Schank. *Tell Me a Story.* Charles Scribner's Sons, 1990.

[Shapiro 1991] Gregory Piatetsky-Shapiro, William J. Frawley (eds). *Knowledge Discovery in Data Bases*, 1991.

[Weitzman 1994] Louis Weitzman and Kent Wittenburg. *Automatic Representation of Multimedia Documents Using Relational Grammars.* ACM Multimedia'94, San Francisco, 1994.

[Wraith 1994] R. Wraith and Rob Gordon. *Community Responses to Natural Disasters.* Melbourne Royal Children's Hospital article, Melbourne, Australia, 1994.

[Wurman 1989] Richard S. Wurman. *Information Anxiety.* Doubleday, 1989.

[Yan 1995] Tak Woon Yan and Hector Garcia-Molina. *SIFT -- A Tool for Wide-Area Information Dissemination.* Proceedings of the 1195 USENIX Technical Conference, pp. 177-186, 1995.

Annex 1 - Extraction Patterns

```
;;/mas/disks/mu4/users/elo/Aardvark/framerd/templates.lisp

;; THIS FILE DEFINES THE TEMPLATES USED BY THE PARSER TO EXTRACT
;; DISASTER FEATURES FROM AN ARTICLE.
;; If the pattern is a list
;; e.g. (list ‘(#(t :NUMBER t) #(“people” t nil))
;;            ‘(#(“kill” :PASSIVE-VERB nil)))
;; signifies that the pattern matches a number followed by “people” or its synonym, followed within
;; a window of 5 words by “kill” or its synonym
;; a word in text matches a vector #(elmt-1 elmt-2 elmt-3 elmt-4) in the pattern:
;; if elmt-1 is t, any word in text matches, if elmt-1 is a string, word has to be same as string or its synonym
;; if elmt-2 is t, word of any part-of-speech matches, if elmt-2 is tag, part-of-speech of the word has
;; to match elmt-2 or its synonym
;; if elmt-3 is t, word in text will be hyper-linked to augmentation
;; if elmt-4 exists, word in text must NOT match it or any of its synonyms
;; If the pattern is a function, a word in text evaluated by the function must return t to match.
;; Synonyms of words and part-of-speeches are defined in lisp hash-table *syn* at the end of this file.
;;*****************************************************************************
(make-new-frame “PEOPLE-KILLED”
‘patterns (list (list ‘(#(t :NUMBER t) #(“people” t nil))
                      ‘(#(“kill” :PASSIVE-VERB nil)))
           (list ‘(#(“kill” :VERB nil))
                  ‘( #(t :NUMBER t) #(“people” t nil)))
           (list ‘( #(t :NUMBER t) #(“people” t nil))
                  ‘( #(“died” :VERB nil)))
           (list ‘( #(t :NUMBER t) #(“people” t nil))
                  ‘( #(“dead” t nil)))
           (list ‘( #(t :NUMBER t) #(“dead” t nil)))
           (list ‘( #(t :NUMBER t) #(“deaths” t nil))
           (list ‘( #(t :NUMBER t) #(“bodies” :NOUN nil)))
           (list ‘( #(“death” t nil) #(“toll” t nil))
                  ‘( #(t :NUMBER t)))
           (list ‘( #(“kill” :VERB nil))
                  ‘( #(t :NUMBER t) #(“animals” t nil t))) ;;= number should not quantify
           (list ‘( #(“claimed” :VERB nil))
                  ‘( #(t :COUNT-ADJECTIVE t) #(“lives” :NOUN nil)))
           (list ‘( #(t :COUNT-ADJECTIVE t) #(“lives” :NOUN nil)))
           (list ‘( #(t :NUMBER t) #(“animals” t nil t)) ;;= number should not quantify
                  ‘( #(“died” :VERB nil))))
‘function (list “rule.people-quantity”))

(make-new-frame “COUNTRY-AFFECTED”
‘patterns (list (list ‘( #(#’(lambda (y) (let ((x (string-trim “.,” y)))
                        (setq x (or (gethash x *syn*) x))
                        (if (fobject? (name x))
                         (cond ((string= (frame-get (name x) ‘is-a) “country”)
                         (name x))
                         ((string= (frame-get (name x) ‘is-a) “state”)
                         (name x)) ;; RETURNS STATE
                        ((string= (frame-get (name x) ‘is-a) “country-adjective”)
                         (frame-get (name x) ‘country)
                         ((string= (frame-get (name x) ‘is-a) “state-possessive”)
                         (frame-get (name x) ‘state))
                         ((string= (frame-get (name x) ‘is-a) “country-possessive”)
                         (frame-get (name x) ‘country))))))
                        t t))))
‘function (list “rule.area-geography”))

(make-new-frame “TYPE”
‘patterns (list (list (list #(#’(lambda (x) (gethash (string-downcase x) *ht*)) t t))))
‘function (list “rule.disaster-type”))

(make-new-frame “LAND-AFFECTED”
‘patterns (list (list ‘(#(t :NUMBER t) #(“acres” t nil))))
‘function (list “rule.area-quantity”))

(make-new-frame “CROPS-AFFECTED”
‘patterns (list (list (list #(“crops” t t))))
‘function (list “rule.crops-kind”)

(make-new-frame “HOUSES-AFFECTED”
‘patterns (list (list ‘(#(t :NUMBER t) (#(“houses” t nil)))
‘function (list “rule.houses-affected”))

(make-new-frame “ANIMALS-KILLED”
‘patterns (list (list (list #(“cattle” t t))))
‘function (list “rule.animals-kind”))

(make-new-frame “PEOPLE-AFFECTED”
‘patterns (list (list ‘(#(t :NUMBER t) #(“people” t nil)))
  (list ‘( #(“displace” :VERB nil))
                         ‘( #(t :NUMBER t) #(“animals” t nil t))) ;;= not animals
‘function (list “rule.people-quantity”))

(make-new-frame “FAMILIES-AFFECTED”
‘patterns (list (list ‘( #(t :NUMBER t) #(“family” t nil))))
‘function (list “rule.families-quantity”)

(make-new-frame “LOSSES”
‘patterns (list (list ‘( #(“$x” t t))))
‘function (list “rule.money-quantity”))

(make-new-frame “DISTANCE”

‘patterns (list (list ‘(#(t :COUNT-ADJECTIVE t) #(“miles” t nil))))
‘function (list “rule.distance”)


;;********************************************************************************************
;; synonyms for words appearing in templates (defined in hash-table *syn*)


(setq *syn* (make-hash-table :test #’equal))
(setf (gethash “kill” *syn*) (list “perish” “drown” ))
(setf (gethash “damaged” *syn*) ‘(“destroyed” “submerged” “burnt” “burn” “flooded” “leveled”))
(setf (gethash “damage” *syn*) ‘(“destroy” “submerge” “burn” “flood” “level”))
(setf (gethash “people” *syn*) ‘(“children” “men” “women” “residents” “inhabitants” “pedestrians”
                                 “survivors” “refugees” “persons” “homeless”))
(setf (gethash “houses” *syn*) ‘(“homes” “huts” “dwellings” “apartments” “households”))
(setf (gethash “crops” *syn*) ‘(“crop” “farmland”))
(setf (gethash “acres” *syn*) ‘(“square miles” “sq mi” “square kilometers” “sq km” “square” “sq”
                                 “hectares” “ha”))
(setf (gethash “cattle” *syn*) ‘(“livestock”))
(setf (gethash “animals” *syn*) ‘(“heads” “cattle” “sheep” “cows” “dogs” “deer” “birds” “trees” “years”
                                  “buildings” “provinces” “villages” “counties” “countries” “percent”))
(setf (gethash “family” *syn*) ‘(“families”))
(setf (gethash “miles” *syn*) ‘(“kilometers” “mi” “km”))
(setf (gethash “displace” *syn*) ‘(“affect” “evacuate” “injure” “hurt”))
(setf (gethash :NUMBER *syn*) ‘(:COUNT-ADJECTIVE))
(setf (gethash :PASSIVE-VERB *syn*) ‘(:VERB-AS-NOUN))
(setf (gethash :VERB *syn*) ‘(:ING-VERB :INFINITIVAL-VERB))
```

Annex 2 - Definition of a Home Community

Representation of Boston in the FactBase:

OBJ-NAME: Boston
PART-OF: Suffolk

POPULATION: 574283
FAMILIES-TOTAL: 117656
HOUSEHOLDS-TOTAL: 227958
FIRST-ANCESTRY--ARAB-400-415: 4429
FIRST-ANCESTRY--AUSTRIAN-003-004: 1075
FIRST-ANCESTRY--BELGIAN-008-010: 291
FIRST-ANCESTRY--CANADIAN-931-934: 2964
FIRST-ANCESTRY--CZECH-111-114: 700
FIRST-ANCESTRY--DANISH-020: 598
FIRST-ANCESTRY--DUTCH-021: 1764
FIRST-ANCESTRY--ENGLISH-015: 24147
FIRST-ANCESTRY--FINNISH-024-025: 475
FIRST-ANCESTRY--FRENCH-EXCEPT-BASQUE-000-1: 9382
FIRST-ANCESTRY--FRENCH-CANADIAN-935-938: 6579
FIRST-ANCESTRY--GERMAN-032-045: 23724
FIRST-ANCESTRY--GREEK-046-048: 5425
FIRST-ANCESTRY--HUNGARIAN-125-126: 1028
FIRST-ANCESTRY--IRISH-050: 106586
FIRST-ANCESTRY--ITALIAN-030-031: 49160
FIRST-ANCESTRY--LITHUANIAN-129: 3602
FIRST-ANCESTRY--NORWEGIAN-082: 1581
FIRST-ANCESTRY--POLISH-142-143: 11412
FIRST-ANCESTRY--PORTUGUESE-084-086: 4004
FIRST-ANCESTRY--ROMANIAN-144-147: 502
FIRST-ANCESTRY--RUSSIAN-148-151: 10565
FIRST-ANCESTRY--SCOTCH-IRISH-087: 4917
FIRST-ANCESTRY--SCOTTISH-088: 6273
FIRST-ANCESTRY--SLOVAK-153: 781
FIRST-ANCESTRY--SUBSAHARAN-AFRICAN-500-599: 9442
FIRST-ANCESTRY--SWEDISH-089-090: 3387
FIRST-ANCESTRY--SWISS-091-096: 711
FIRST-ANCESTRY--UKRAINIAN-171-174: 1277
FIRST-ANCESTRY--UNITED-STATES-OR-AMERICAN-939-994: 10624
FIRST-ANCESTRY--WELSH-097: 825
FIRST-ANCESTRY--WEST-INDIAN-EXCLUDING-HISPANIC-ORIGIN-GRPS: 25667
FIRST-ANCESTRY--YUGOSLAVIAN-152: 308
FIRST-ANCESTRY--RACE-OR-HISPANIC-ORIGIN-GROUPS: 131529
MEDIAN-HOUSEHOLD-INCOME-IN-1989: 29180

ENGLISH-LANGUAGE-ONLY: 400756
GERMAN-LANGUAGE: 2189
YIDDISH-LANGUAGE: 534
OTHER-WEST-GERMANIC-LANGUAGES: 328
SCANDINAVIAN-LANGUAGE: 423
GREEK-LANGUAGE: 3397
INDIC-LANGUAGE: 1267
ITALIAN-LANGUAGE: 11406
FRENCH-OR-FRENCH-CREOLE-LANGUAGE: 19525
PORTUGUESE-OR-PORTUGUESE-CREOLE-LANGUAGE: 7728
SPANISH-OR-SPANISH-CREOLE: 51233
POLISH-LANGUAGE: 2104
RUSSIAN-LANGUAGE: 3211
SOUTH-SLAVIC-LANGUAGES: 128
OTHER-SLAVIC-LANGUAGES: 462
OTHER-INDOEUROPEAN-LANGUAGES: 4554
ARABIC-LANGUAGE: 2617
TAGALOG-LANGUAGE: 622
CHINESE-LANGUAGE: 14255
HUNGARIAN-LANGUAGE: 222
JAPANESE-LANGUAGE: 1560
MON-KHMER-LANGUAGE: 1028
KOREAN-LANGUAGE: 981
NATIVE-NORTH-AMERICAN-LANGUAGES: 155
VIETNAMESE-LANGUAGE: 4212
OTHER-LANGUAGES: 3614
EMPLOYED-16+-AGRICULTURE: 1440
EMPLOYED-16+-MINING: 142
EMPLOYED-16+-CONSTRUCTION: 11416
EMPLOYED-16+-MANUFACTURING: 15916
EMPLOYED-16+-TRANSPORTATION: 12778
EMPLOYED-16+-COMMUNICATIONS-AND-OTHER-PUBLIC-UTILITIES: 7291
EMPLOYED-16+-WHOLESALE-TRADE: 7810
EMPLOYED-16+-RETAIL-TRADE: 40072
EMPLOYED-16+-FINANCE: 31239
EMPLOYED-16+-BUSINESS-AND-REPAIR-SERVICES: 16709
EMPLOYED-16+-PERSONAL-SERVICES: 11007
EMPLOYED-16+-ENTERTAINMENT-AND-RECREATION-SERVICES: 3948
EMPLOYED-16+-PROFESSIONAL-&-RELATED-SVCS-HEALTH-SERVICES: 38290
EMPLOYED-16+-PROFESSIONAL-&-RELATED-EDUCATIONAL-SERVICES: 29753
EMPLOYED-16+-PROFESSIONAL-&-RELATED-SVCS-OTHER: 32160
EMPLOYED-16+-PUBLIC-ADMINISTRATION: 16047
DISTANCES: ("1-mi-BOS" "2.5-mi-BOS" "5-mi-BOS" "5-mi-BOS-1" "15-mi-BOS" "17-mi-BOS" "40-mi-BOS" "102-mi-BOS" "118-mi-BOS" "137-mi-BOS" "167-mi-BOS" "211-mi-BOS" "463-mi-BOS" "1015-mi-BOS" "1285-mi-BOS" "3088-mi-BOS")
MAPS: ("BOS-4x3.5.gif" "BOS-15x13.gif" "BOS-114x100.gif" "BOS-250x220.gif")

Annex 3 - How to Add a Home Community

This Annex describes the information needed about a city in order to add it as a home community in PLUM. Data should be given at city, county and state level. Create a new frame for the city, the county and the state, if not previously defined. In addition to the slot values such as the ones given in Annex 2, a home community frame MUST have the following slots (example for Boston):

ACCRONYM: "BOS" (string of 3 letters)
FULL-NAME: "Boston, MA, USA" (string)
CENSUS-SOURCE: "US Census Data" (string)
DIR: "Boston" (name of directory where augmentation files will be written, create this directory under /mas/disks/mu4/users/elo/plum/ where existing city directories reside)

To create a new frame, evaluate the following function:

```
(make-new-frame "Boston"
          'ACCRONYM: "BOS"
          'FULL-NAME: "Boston, MA, USA"
          'CENSUS-SOURCE: "US Census Data"
          'POPULATION: 574283
          'FAMILIES-TOTAL: 117656
          'HOUSEHOLDS-TOTAL: 227958
          'MEDIAN-HOUSEHOLD-INCOME-IN-1989: 29180
          '0-TO-15-YEARS: 92479
                  ;;... more slots on age
          'FIRST-ANCESTRY--ARAB-400-415: 4429
                  ;;... more slots on origins
          'GERMAN-LANGUAGE: 2189
                  ;;.. more slots on language
          'EMPLOYED-16+-=AGRICULTURE: 1440
                  ;;... more slots on employment
          'MAPS '("BOS-15x13.gif")
          'DISTANCES '("5-mi-BOS"))
```

Check if languages are previously defined by evaluating e.g. (name "GERMAN-LANGUAGE"). If it evaluates to nil, you must define the language. To define a new language create a new frame:

```
(make-new -frame "FRENCH-OR-FRENCH-CREOLE-LANGUAGE"
          'LANGUAGES (list "Cajun" "Haitian Creole" "Provencal" "Walloon" "French"))
```

or

```
(make-new-frame "GERMAN-LANGUAGE"
          'LANGUAGES (list "German"))
```

For each new origin you enter you must do the same:

```
(make-new-frame "FIRST-ANCESTRY--HUNGARIAN-125-126"
          'COUNTRIES (list "Hungary")
```

```
            ‘print-string “of first ancestry Hungarian”)
```

or

```
(make-new-frame “FISRT-ANCESTRY--SUBSAHARAN-AFRICAN-500-599”
            ‘COUNTRIES (list “Angola” “Benin” ...)
            ‘PRINT-STRING “of first ancestry Sub-Saharan African”)
```

Each map must be placed in gif format in the directory /mas/mu/www/elo/maps/ . In Lisp a map covering an area of 30x26 miles must be defined as:

```
(make-new-frame “BOS-15x13.gif”
            ‘width 15 ;;yes, half of width, just to complicate things...
            ‘height 13 ;;half the height
            ‘x-pixels 466 ;;width of image in pixels
            ‘y-pixels 415) ;;height of image in pixels
```

Adding local distances is optional. A new distance for a city must be defined as:

```
(make-new-frame “5-mi-BOS”
            ‘DISTANCE 5
            ‘UNIT “mi”
            ‘DESCRIPTION “riding the T Blue-line from Government Centre to Revere Beach”)
```

When a city has been completely described in Lispworks, you can set the list of home-communities to the new city by evaluating e.g. (defun home-communities () (list “Chicago”)). This will augment all incoming articles for Chicago only.

Of course, all this assumes that you have loaded Lispworks with the PLUM files. To do so, copy /mas/disks/mu4/users/elo/.lispworks into your home directory. This .lispworks loads files from directory /mas/disks/mu4/users/elo/Aardvark/framerd/ PLUM needs to run properly. It also loads the hashtable which stores pointers to all the objects in the PLUM database. Once Lispworks has finished loading, evaluate (in-package :framerd).

Before adding new frames you may want to make a copy of the PLUM database stored in file /mas/disks/mu4/users/elo/plum in case you create erroneous frames or overwrite existing frames. At this point you are ready to create new frames. After the session, evaluate (save-plum) to save all newly defined frames to the PLUM database.

Annex 4 - Augmenting Rules

Each feature of the disaster is associated with one or more augmentation rules. The rules are defined in Lisp in the file /mas/disks/mu4/users/elo/Aardvark/framerd/rules.lisp. This Annex describes each rule briefly.
Let the value extracted from text be X.

****

(frame-get "PEOPLE-KILLED" 'function) -> "rule.people-quantity"

(frame-get "PEOPLE-AFFECTED" 'function) -> "rule.people-quantity"

"rule.people-quantity"

find-people-quantity-home:

- Compare X with the home town population
- Start at town level. Calculate factor between X and all slots of type 'people-quantity'. Move to county, state, and country and repeat calculation. Choose one with smallest factor. In augmentation, compare X to it.

find-people-quantity:

- If X is greater than 500 or X is the largest number extracted from article, then compare X with the total population of disaster-stricken country.

****

(frame-get "COUNTRY-AFFECTED" 'function) -> "rule.area-geography"

"rule.area-geography":

find-people-origin-home:

- For all origins in slot 'origin-tags of disaster-struck country, check if home community has the origin. If yes, report number of people of this origin in home community. If origin refers to a continent or region, mention that no record exists specifically for the disaster-stricken country.

find-language-home:

- For all languages in slot 'languages of disaster-struck country, check if home community has its language-tag. If yes, report number of people speaking language at home. If language refers to a language group, mention that no record exists specifically for the language.

find-area-geography-home:

- Compare size of disaster-stricken country with home state or country, depending which is closer in size. Compare population densities. Generate comparative statistics from World Fact Book between disaster-stricken country and home country.

find-area-geography:

- Report location and other pre-defined facts about disaster-stricken country.

****

(frame-get "TYPE" 'function) -> "rule.disaster-type"

"rule.disaster-type":

find-disaster-history-home:

- Find most costly or most deadly disaster of this type in history of home country. Generate graph of 20 years of disasters in home country.

find-disaster-history:

- Find most costly or most deadly disaster of this type in history of disaster-stricken country. Generate graph of 20 years of disasters in disaster-stricken country.

****

(frame-get “LAND-AFFECTED” ‘function) -> “rule.area-quantity”

“rule.area-quantity”:
find-area-quantity-home:
- Find smallest suitable map of home town and generate a shadowed circle same size as X. If no map is available, report only the radius of such a circle.
- Calculate the percentage X represents of the total area of disaster-stricken country.

****

(frame-get “CROPS-AFFECTED” ‘function) ->“rule.crops-kind”

“rule.crops-kind”:
find-crops-kind:
- Retrieve slot ‘agriculture of disaster-stricken country.

****

(frame-get “HOUSES-AFFECTED” ‘function) -> “rule.houses-affected”

“rule.houses-affected”:
find-houses-quantity-home:
- Compare X to number of households in home town.

****

(frame-get “FAMILIES-AFFECTED” ‘function) ->“rule.families-quantity”

“rule.families-quantity”:
find-families-quantity-home:
- Compare X to number of families living in home town.

****

(frame-get “LOSSES” ‘function) -> “rule.money-quantity”

“rule.money-quantity”:
find-money-quantity:
- Calculate how much every household in home town would have to pay to cover X. Report median household income in home town.
- Compare national product per capita in disaster-stricken country and home country.
- Compare X to national product of disaster-stricken country.

****

(frame-get “DISTANCE” ‘function) -> “rule.distance”

“rule.distance”:
find-distance-home:
- For slot ‘distances of home community, find distance equivalent to X +- 15%.

****