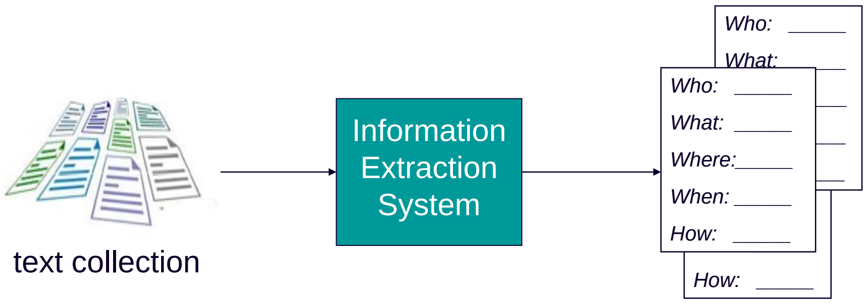
# Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Specifying an IE task**
- **Acquiring extraction patterns**
- **Named entity detection**

# Information extraction



text collection

Information Extraction System

Who: _____

What: _____

Where:_____

When: _____

How: _____

# IE system: terrorism

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE **MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS**.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE **POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THIS ASSASSINATION** TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI **IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.**

# IE system: output

| | |
|---|---|
| 1. DATE | - 15 JAN 90 |
| 2. LOCATION | EL SALVADOR:<br>CENTRAL AMERICAN UNIVERSITY |
| 3. TYPE | MURDER |
| 4. STAGE OF EXECUTION | ACCOMPLISHED |
| 5. INCIDENT CATEGORY | TERRORIST ACT |
| 6. PERP: INDIVIDUAL ID | "FOUR OFFICERS"<br>"ONE COLONEL"<br>"FIVE MEMBERS OF THE ARMED FORCES" |
| 7. PERP: ORGANIZATION ID | "ARMED FORCES", "FMLN" |
| 8. PERP: CONFIDENCE | REPORTED AS FACT |
| 9. HUM TGT: DESCRIPTION | "JESUIT PRIESTS"<br>"WOMEN" |
| 10. HUM TGT: TYPE | CIVILIAN: "JESUIT PRIESTS"<br>CIVILIAN: "WOMEN" |
| 11. HUM TGT: NUMBER | 6: "JESUIT PRIESTS"<br>2: "WOMEN" |
| 12. EFFECT OF INCIDENT | DEATH: "JESUIT PRIESTS"<br>DEATH: "WOMEN" |

# IE system: natural disasters

Disaster Type: earthquake
- location: *Afghanistan*
- date: *today*
- magnitude: *6.9*
- magnitude-confidence: high
- epicenter: *a remote part of the country*
- damage:
  - human-effect:
    - victim: *Thousands of people*
    - number: *Thousands*
    - outcome: dead
    - confidence: medium
    - confidence-marker: *feared*
  - physical-effect:
    - object: *entire villages*
    - outcome: damaged
    - confidence: medium
    - confidence-marker: *Details now hard to come by / reports say*

**PAKISTAN MAY BE PREPARING FOR ANOTHER TEST**

Thousands of people are feared dead following... (voice-over) ...a powerful earthquake that hit Afghanistan today. The quake registered 6.9 on the Richter scale, centered in a remote part of the country. (on camera) Details now hard to come by, but reports say entire villages were buried by the quake.

Document no.: ABC19980530.1830.0342
Date/time: 05/30/1998 18:35:42.49

# IE from semi-structured text

- Job postings:
  - From newsgroups, web pages: Flipdog
- Job resumes:
  - BurningGlass
  - Mohomine
- Seminar announcements
- Company information from the web
- University information from the web
- Apartment rental ads
- …

# Sample job posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 2006 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training.
Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

# Extracted job template

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 2006

# Information extraction  (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Newsgroup messages
  - Classified ads
  - Medical notes

# Template slot types

- Slots in template typically filled by a **substring** from the document.
- Some slots may have a **fixed SETof pre-specified possible fillers** that may not occur in the text itself.
  - Terrorist act: threatened, attempted, accomplished.
  - Job type: clerical, service, custodial, etc.
  - Company type:  SEC code
- Some slots may allow **multiple fillers**.
  - Programming language
- Some domains may allow **multiple extracted templates per document**.
  - Multiple apartment listings in one ad

# Evaluating IE systems

- Evaluate performance on independent, manually-annotated test data not used during system development.
- Compute average value of metrics adapted from IR:
  - **Recall** = *# correct extractions / # extractions in gold standard*
  - **Precision** = *# correct extractions / # extractions by system*
  - **F-Measure** = Harmonic mean of recall and precision

# State of the art

Unrestricted text:
65-70% R; 70-80% P

Semi-structured text:
90+% R/P

**MUC**
**[1991-94]**

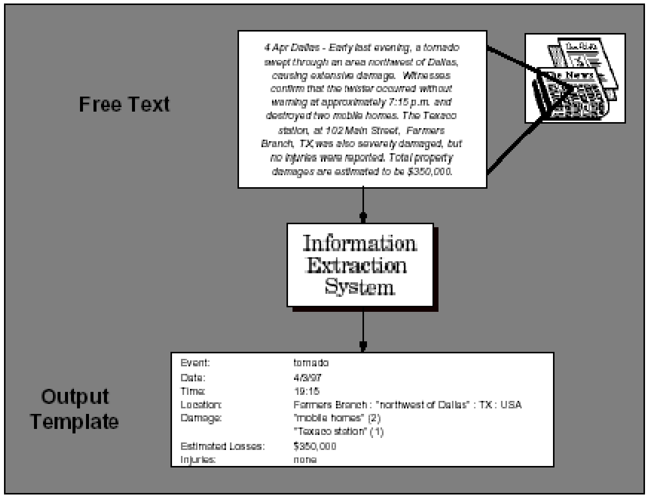
**ACE**
**[1991-94]**

- terrorist activities
- business joint ventures
- microelectronic chip fabrication
- changes in corporate management
- natural disasters
- summarize medical patient records
- create job-listing databases from newsgroups
- bioinformatics

# Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture

# Natural disasters example

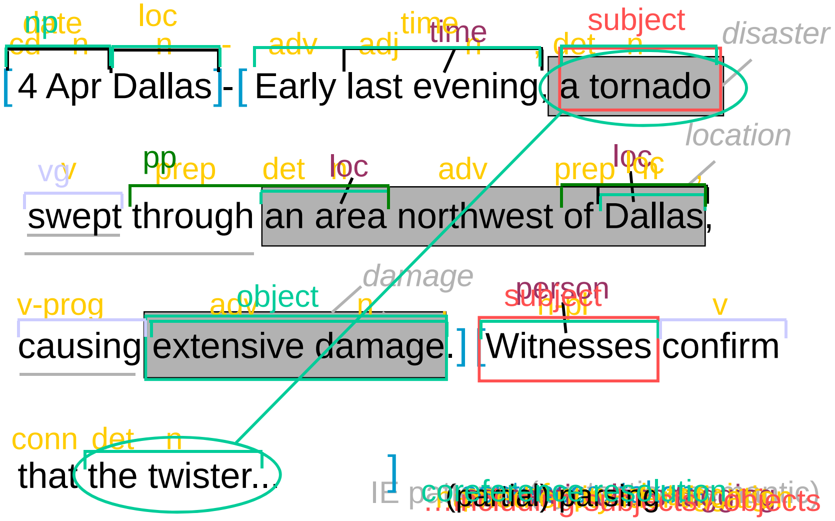


*4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX was also severely damaged, but no injuries were reported. Total property damages are estimated to be $350,000.*

| | |
|---|---|
| Event: | tornado |
| Date: | 4/3/97 |
| Time: | 19:15 |
| Location: | Farmers Branch : "northwest of Dallas" : TX : USA |
| Damage: | "mobile homes" (2)<br>"Texaco station" (1) |
| Estimated Losses: | $350,000 |
| Injuries: | none |

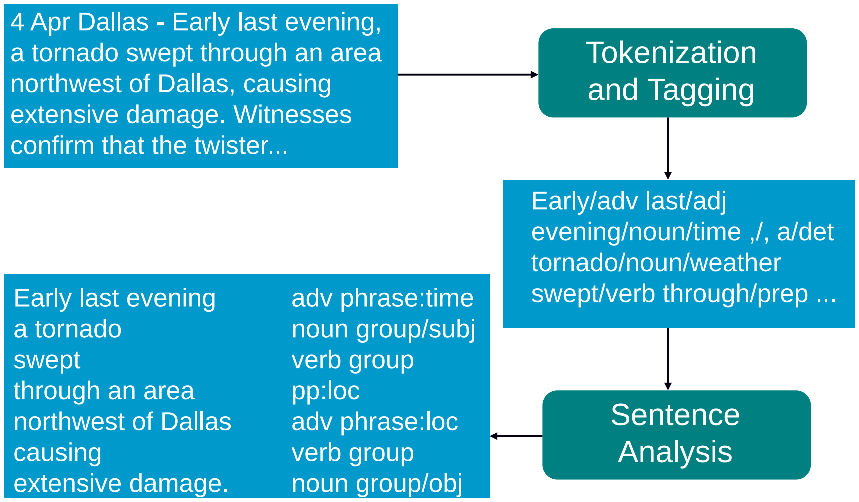
Figure 1: Information Extraction System in the Domain of Natural Disasters.

# IE system components



[4 Apr Dallas]-[ Early last evening, a tornado
swept through an area northwest of Dallas,
causing extensive damage.] [Witnesses confirm
that the twister...]

# Stages of processing



4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister...

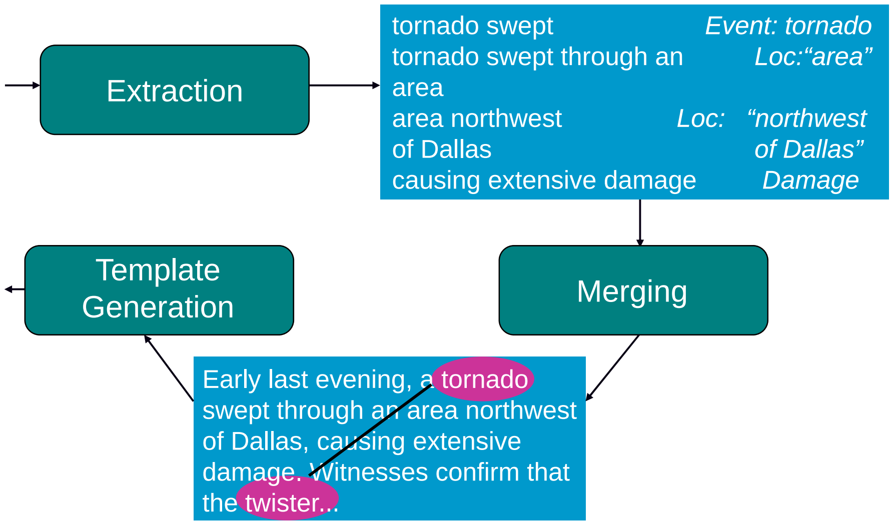
Tokenization and Tagging

Early/adv last/adj evening/noun/time ,/, a/det tornado/noun/weather swept/verb through/prep ...

Sentence Analysis

| | |
|---|---|
| Early last evening | adv phrase:time |
| a tornado | noun group/subj |
| swept | verb group |
| through an area | pp:loc |
| northwest of Dallas | adv phrase:loc |
| causing | verb group |
| extensive damage. | noun group/obj |

# Stages of processing


Extraction
tornado swept
*Event: tornado*
tornado swept through an area
*Loc:"area"*
area northwest of Dallas
*Loc: "northwest of Dallas"*
causing extensive damage
*Damage*
Merging
Template Generation
Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister...

# Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Specifying an IE task**
- **Acquiring extraction patterns**
- **Named entity detection**

# IE: Changes in Management

The company also said its president and former chairman both resigned.

Evergreen said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founders, including Mr. Nelsen.

Excluding these shares, Evergreen Information has more than two million shares or exercisable warrants outstanding, according to a spokeswoman.

# IE: dogs

Cavalier King Charles Spaniel
(Ruby Spaniel) (Blenheim Spaniel)

Height: 12-13 inches (30-33 cm.)
Weight: 10-18 pounds (5-8 kg.)

Prone to syringomyelia, hereditary eye disease, dislocating kneecaps (patella), back troubles, ear infections, early onset of deafness or hearing trouble. Sometime's hip dysplasia. Don't over feed. This breed tends to gain weight easily. Some lines are genetically disposed early onset to a serious heart problem, which sometimes causes early death. When selecting one of these dogs, it is extremely important to check the medical history of several previous generations.

Cavalier King Charles Spaniels are good for apartment life. They are moderately active indoors and a small yard will be sufficient. The Cavalier does not do well in very warm conditions.

Cavalier King Charles Spaniels need a daily walk. Play will take care of a lot of their exercise needs, however, as with all breeds, play will not fulfill their primal instinct to walk. Dogs who do not get to go on daily walks are more likely to display behavior problems. They will also enjoy a good romp in a safe open area off lead, such as a large fenced in yard.

# Specifying the Extraction Task

- **Slots**
  - String fill?
  - Set fill?
  - Normalization?
  - One/multiple fills?
  - Cross-referencing with other slots?

# Acquiring extraction patterns

- **Manually defined**
  - Finite-state methods (as in Fastus)
    - Univ of Sheffield's Gate system
  - Patterns
- **Learning approaches**
  - Fully supervised
  - Semi-supervised methods
  - Not much in the way of unsupervised methods

Will cover specific methods later…

# Annotating sample documents

post post
The company also said its president and former chairman both resigned.

IO-person:out
Evergreen said Barry Nelsen, who had a heart-bypass operation

post post
last week, resigned as president and chief executive. The board

IO-person:out
formally accepted the resignation of Thomas Casey, its former

post post
chairman, who stepped down effective Feb. 2.

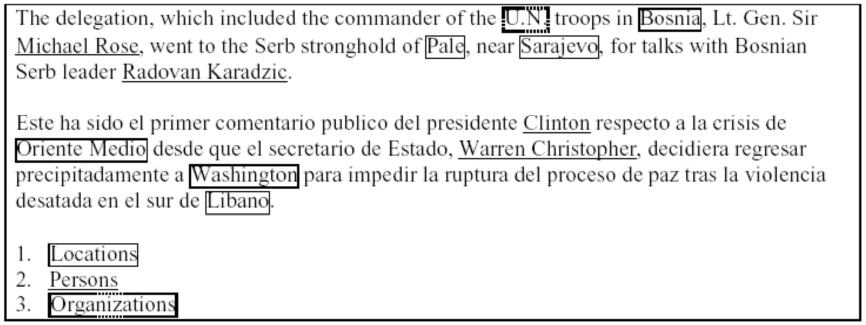
Annotation software exists:
e.g. see Stanford NLP web page

# Manually specified extraction patterns

- **Changes in management**

# NE Identification

- **Identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages.**



The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

Este ha sido el primer comentario publico del presidente Clinton respecto a la crisis de Oriente Medio desde que el secretario de Estado, Warren Christopher, decidiera regresar precipitadamente a Washington para impedir la ruptura del proceso de paz tras la violencia desatada en el sur de Libano.

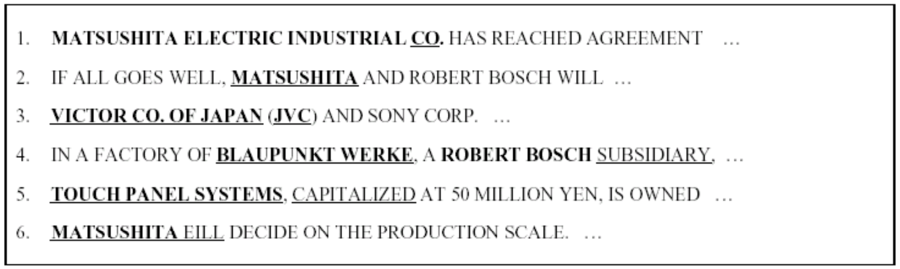
1. Locations
2. Persons
3. Organizations

**Figure 1.1 Examples.** Examples of correct labels for English text and for Spanish text**.**

# Guidelines need to be specified

- ***The Wall Street Journal*** **: artifact or organization?**
- ***White House*** **: organization or location?**
- **Is a street name a location?**
- **Should *yesterday* and *last Tuesday* be labeled as dates?**
- **Is *mid-morning* a time?**

# Examples



1. **MATSUSHITA ELECTRIC INDUSTRIAL CO.** HAS REACHED AGREEMENT …
2. IF ALL GOES WELL, **MATSUSHITA** AND ROBERT BOSCH WILL …
3. **VICTOR CO. OF JAPAN** (**JVC**) AND SONY CORP. …
4. IN A FACTORY OF **BLAUPUNKT WERKE**, A **ROBERT BOSCH** SUBSIDIARY, …
5. **TOUCH PANEL SYSTEMS**, CAPITALIZED AT 50 MILLION YEN, IS OWNED …
6. **MATSUSHITA** EILL DECIDE ON THE PRODUCTION SCALE. …

**Figure 2.1 English Examples.** Finding names ranges from the easy to the challenging. Company names are in boldface. It is crucial for any name-finder to deal with the underlined text.

# Approaches to NE identification

- **Handcrafted finite state patterns**
  - <proper noun>$^+$ <corporate designator> → <corporation>
  - Can't easily capture typical naming conventions
    - "Boston Power & Light" (corporation, electric utility)
  - Time-consuming to define
  - Maintenance is a problem
    - E.g. moving to NYT from WSJ
  - Not generally portable to new languages
- **Machine learning approaches**
  - HMM's (or variants thereof) are the standard

# NE Results Using HMM's

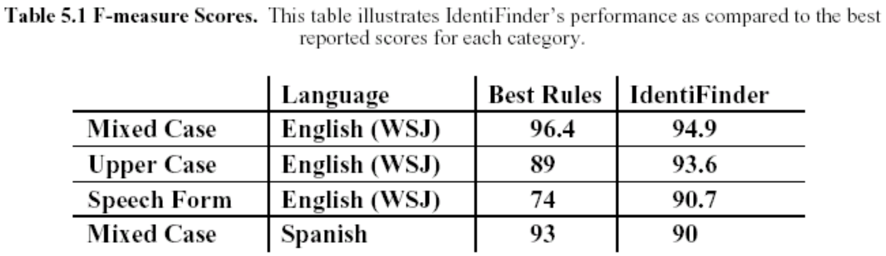


**Table 5.1 F-measure Scores.** This table illustrates IdentiFinder's performance as compared to the best reported scores for each category.

| | Language | Best Rules | IdentiFinder |
|---|---|---|---|
| Mixed Case | English (WSJ) | 96.4 | 94.9 |
| Upper Case | English (WSJ) | 89 | 93.6 |
| Speech Form | English (WSJ) | 74 | 90.7 |
| Mixed Case | Spanish | 93 | 90 |