Appinions Influencer demo

Opinions + WSD = joke

From a clip of Rodney Dangerfield from the Jay Leno show

"I went to my psychiatrist. He said I was crazy. I asked him for a second opinion."

"He said,`OK, you're ugly too'."

Echoes of Power: Language Effects and Power Differences in Social Interaction

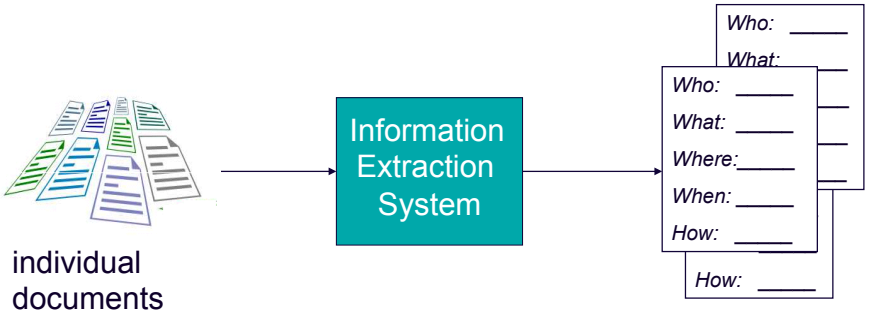
Cristian Danescu-Niculescu-Mizil Cornell University cristian@cs.cornell.edu, Lillian Lee Cornell University llee@cs.cornell.edu, Bo Pang Yahoo! bopang@yahoo-inc.com, Jon Kleinberg Cornell University kleinber@cs.cornell.edu

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Manually defined patterns
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Semi-structured text
- **Named entity detection**
- **Sequence-tagging methods**

Last classes

Information extraction



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*

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 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.**

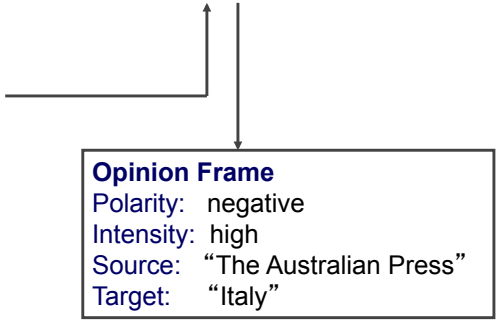
IE system: output

1. DATE	- 15 JAN 90
2. LOCATION	EL SALVADOR: CENTRAL AMERICAN UNIVERSITY
3. TYPE	MURDER
4. STAGE OF EXECUTION	ACCOMPLISHED
5. INCIDENT CATEGORY	TERRORIST ACT
6. PERP: INDIVIDUAL ID	"FOUR OFFICERS" "ONE COLONEL" "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" "WOMEN"
10. HUM TGT: TYPE	CIVILIAN: "JESUIT PRIESTS" CIVILIAN: "WOMEN"
11. HUM TGT: NUMBER	6: "JESUIT PRIESTS" 2: "WOMEN"
12. EFFECT OF INCIDENT	DEATH: "JESUIT PRIESTS" DEATH: "WOMEN"

Fine-grained Opinion Extraction

"The Australian Press **launched a bitter attack** on Italy"

- **Five components**
  - Opinion trigger
  - Polarity
    - positive
    - negative
    - neutral
  - Strength/intensity
    - low..extreme
  - Source (opinion holder)
  - Target (topic)



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
  - Subjective language

Template slot types

- Slots in template typically filled by a **substring** from the document.
- Some slots may have a **fixed set of 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 system performance vs. 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

Specifying the Extraction Task

- **Define the domain**
- **Slots/components in the output template**
  - String fill?
  - Set fill?
  - Normalization?
  - One/multiple fills?
  - Cross-referencing with other slots?
- **Develop manual annotation instructions**

Changes in Management

Evergreen Information 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.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

<TEMPLATE-9303020074-1> :=
DOC_NR: "9303020074"
CONTENT: <SUCCESSION_EVENT-9303020074-1>
<SUCCESSION_EVENT-9303020074-2>
<SUCCESSION_EVENT-9303020074-3>
<SUCCESSION_EVENT-9303020074-4>
<SUCCESSION_EVENT-9303020074-1> :=
SUCCESSION_ORG: <ORGANIZATION-9303020074-1>
POST: "president"
IN_AND_OUT: <IN_AND_OUT-9303020074-1>
<IN_AND_OUT-9303020074-2>
VACANCY_REASON: REASSIGNMENT
COMMENT: "Nelson out, Bell in as pres of Evergreen Info"
/ "This event could be collapsed with SUCCESSION_EVENT-2"

<SUCCESSION_EVENT-9303020074-2> :=
SUCCESSION_ORG: <ORGANIZATION-9303020074-1>
POST: "chief executive" / "CEO"
IN_AND_OUT: <IN_AND_OUT-9303020074-3>
<IN_AND_OUT-9303020074-4>
VACANCY_REASON: REASSIGNMENT
COMMENT: "Nelson out, Bell in as CEO of Evergreen Info"
<SUCCESSION_EVENT-9303020074-3> :=
SUCCESSION_ORG: <ORGANIZATION-9303020074-1>
POST: "chairman"
IN_AND_OUT: <IN_AND_OUT-9303020074-5>
<IN_AND_OUT-9303020074-6>
VACANCY_REASON: REASSIGNMENT
COMMENT: "Casey out, Bell in as chmn of Evergreen Info"
<SUCCESSION_EVENT-9303020074-4> :=
SUCCESSION_ORG: <ORGANIZATION-9303020074-1>
POST: "chief financial officer"
IN_AND_OUT: <IN_AND_OUT-9303020074-7>
VACANCY_REASON: OTH_UNK
COMMENT: "Bell in as CFO at Evergreen Info 'since the fall'"

<IN_AND_OUT-9303020074-1> :=
IO_PERSON: <PERSON-9303020074-1>
NEW_STATUS: OUT
ON_THE_JOB: UNCLEAR
COMMENT: "Nelson out as pres"
/ "ON_THE_JOB: 'resign' (headline), 'resigned'"
<IN_AND_OUT-9303020074-2> :=
IO_PERSON: <PERSON-9303020074-3>
NEW_STATUS: IN
ON_THE_JOB: UNCLEAR
OTHER_ORG: <ORGANIZATION-9303020074-1>
REL_OTHER_ORG: SAME_ORG
COMMENT: "Bell in as pres -- was already CFO at same org"
/ "ON_THE_JOB: 'was named'"
<IN_AND_OUT-9303020074-3> :=
IO_PERSON: <PERSON-9303020074-1>
NEW_STATUS: OUT
ON_THE_JOB: UNCLEAR
COMMENT: "Nelson out as CEO"
/ "This obj identical to IN_AND_OUT-1"

<IN_AND_OUT-9303020074-4> :=
IO_PERSON: <PERSON-9303020074-3>
NEW_STATUS: IN
ON_THE_JOB: UNCLEAR
OTHER_ORG: <ORGANIZATION-9303020074-1>
REL_OTHER_ORG: SAME_ORG
COMMENT: "Bell in as CEO"
/ "This obj identical to IN_AND_OUT-2"
<IN_AND_OUT-9303020074-5> :=
IO_PERSON: <PERSON-9303020074-2>
NEW_STATUS: OUT
ON_THE_JOB: NO
COMMENT: "Casey out"
/ "ON_THE_JOB: 'stepped down effective Feb. 2'"
<IN_AND_OUT-9303020074-6> :=
IO_PERSON: <PERSON-9303020074-3>
NEW_STATUS: IN
ON_THE_JOB: UNCLEAR
OTHER_ORG: <ORGANIZATION-9303020074-1>
REL_OTHER_ORG: SAME_ORG
COMMENT: "Bell in as chmn"
/ "This obj identical to IN_AND_OUT-2"

<IN_AND_OUT-9303020074-7> :=
IO_PERSON: <PERSON-9303020074-3>
NEW_STATUS: IN
ON_THE_JOB: YES
COMMENT: "Bell in"
/ "ON_THE_JOB: has been CFO 'since the fall'"
<ORGANIZATION-9303020074-1> :=
ORG_NAME: "Evergreen Information Technologies Inc."
ORG_ALIAS: "Evergreen Information Technologies"
"Evergreen"
"Evergreen Information"
ORG_DESCRIPTOR: "The computer products and services concern"
ORG_TYPE: COMPANY
ORG_LOCALE: McLean CITY
ORG_COUNTRY: United States

<PERSON-9303020074-1> :=
PER_NAME: "Barry Nelsen"
PER_ALIAS: "Nelsen"
PER_TITLE: "Mr."
<PERSON-9303020074-2> :=
PER_NAME: "Thomas Casey"
<PERSON-9303020074-3> :=
PER_NAME: "Martin Bell"
PER_ALIAS: "Bell"
PER_TITLE: "Mr."

IE: dogs

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.

Information extraction

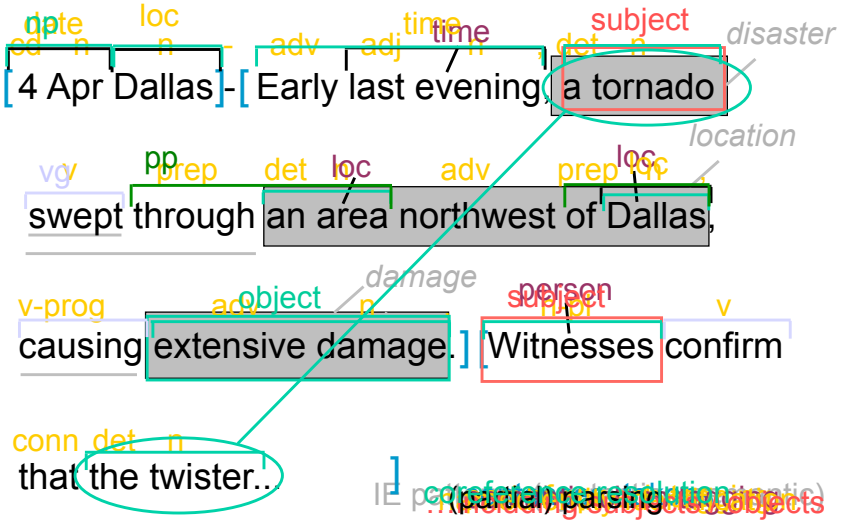
- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Manually defined patterns
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Semi-structured text
- **Named entity detection**
- **Sequence-tagging methods for IE**

Natural disasters example

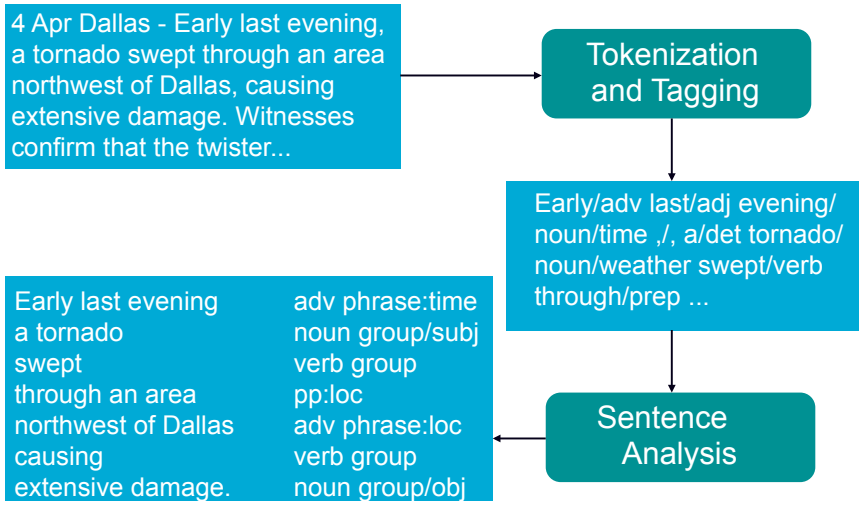
4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15pm and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX, was also severely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

Event:	tornado
Date:	4/3/2008
Time:	19:15
Location:	Farmers Branch: "northwest of Dallas": TX: USA
Damage:	"mobile homes" (2) "Texaco station" (1)
Estimated Losses:	$350,000
Injuries:	none

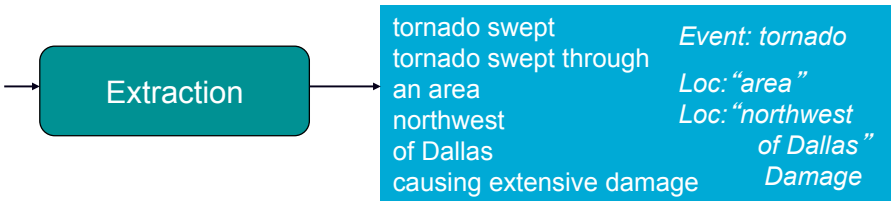
IE system components



Pre-processing



Learning



Post-processing