### Information Extraction

- **Today**
  - Intro to IE
  - IE system architecture
  - Acquiring extraction patterns

### Subjective Language

- **Subjective sentences express private states**, i.e. *internal mental or emotional states*
  - speculations, beliefs, emotions, evaluations, goals, opinions, judgments, ...
    - (1) Jill said, "I hate Bill."
    - (2) John *thought* he won the race.
    - (3) Claire *hoped* her lecture would go well.

### Subjectivity vs. Sentiment

- **Sentiment expressions are a type of subjective expression**
  - expressions of *positive* and *negative* emotions, judgments, evaluations, ...
    - (1) Jill said, "I hate Bill."
    - (2) John *thought* he won the race.
    - (3) Claire *hoped* her lecture would go well.
The Australian press has launched a bitter attack on Italy after seeing their beloved Socceroos eliminated on a controversial late penalty. Italian coach Lippi has been blasted for his comments after the game.

In the opposite camp, Lippi is preparing his side for the upcoming game with Ukraine. He hailed 10-man Italy's determination to beat Australia and said their winning penalty was rightly given.

“"The Australian Press launched a bitter attack on Italy"
Example – fine-grained opinions

Australian press has launched a bitter attack on Italy after seeing their beloved Socceroos lose on a controversial late penalty. Italian coach Lippi has also been blasted for his comments after the game.

In the opposite camp Lippi is preparing his side for the upcoming match against Ukraine. He hailed 10-man Italy's determination to beat Australia and said the penalty was rightly given.

Example – Opinion Summary

- **Australian Press**
- **penalty**
- **Socceroos**
- **Marcello Lippi**
- **Italy**

Information Extraction

- **Today**
  - Intro to IE
  - 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
SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- 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 vs. IR vs. full NLU**

- IE requires more **text-understanding** capabilities than the bag-of-words approaches provided by IR techniques
- IE systems often presume that a **text categorization** system has identified documents relevant to the extraction domain
- IE requires more than **document classification**
- IE requires a more **shallow understanding** of the text than a natural language understanding system attempting full/deep semantic analysis.

**IR, TC < IE < NLU**
Information extraction

- Introduction
  - Task definition
  - Evaluation
  - IE system architecture
- Acquiring extraction patterns

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...

IE system components

Stages of processing

Tokenization and Tagging

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

Extraction:
- Event: tornado
- Loc: "area"
- Damage

Template Generation:
- tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister...

Merging:

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

Syntactico-semantic patterns

The twister occurred without warning at approximately 7:15p.m. and destroyed two mobile homes.

Pattern:
- Trigger: “destroyed”
  - condition: active voice verb?
- Slot: Damaged-Object
- Position: direct-object
  - condition: DO is a physical-object?

from Cardie [1997]
Learning IE patterns from examples

- **Goal**
  - Given a training set of *annotated* documents [answer keys],
  - Learn extraction patterns for each slot using an appropriate machine learning algorithm.

- **Options**
  - Memorize the fillers of each slot
  - Generalize the fillers using
    - p-o-s tags?
    - phrase structure (NP, V) and grammatical roles (SUBJ, OBJ)?
    - semantic categories?

Learning IE patterns

- **Methods vary with respect to**
  - The **class of pattern** learned (e.g. lexically-based regular expression, syntactico-semantic pattern)
  - **Training corpus** requirements
  - Amount and type of human feedback required
  - Degree of **pre-processing** necessary
  - **Other resources/knowledge bases** presumed

Learning syntactico-semantic patterns

The twister occurred without warning at approximately 7:15p.m. and *destroyed two mobile homes.*

**Pattern:**
- **Trigger:** "destroyed"
  - condition: active voice verb?
- **Slot:** Damaged-Object
- **Position:** direct-object
  - condition: DO is a physical-object?

*Autoslog* (Riloff & Lehnert, 1993)

Pattern templates

**Noun phrase extraction only**

```
<subject> <passive-verb> <object>
<subject> <active-verb> <object>
<subject> <infinitival-verb> <object>
<subject> <auxiliary-verb>+<noun>
*<passive-verb> <dobj>
<active-verb> <dobj>
<infinitive> <dobj>
<verb>+<infinitive> <dobj>
<gerund> <obj>
<noun>+ <auxiliary> <dobj>
<noun>+<prep> <np>
<active-verb>+<prep> <np>
<passive-verb>+<prep> <np>
```

```
<victim> was murdered
<perpetrator> bombed
<perpetrator> attempted to kill
<victim> was victim
killed <victim>
bombed <target>
to kill <victim>
threatened to attack <target>
killing <victim>
fatality was <victim>
```

```
bomb against <target>
killed with <instrument>
was aimed at <target>
```