Information Extraction

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

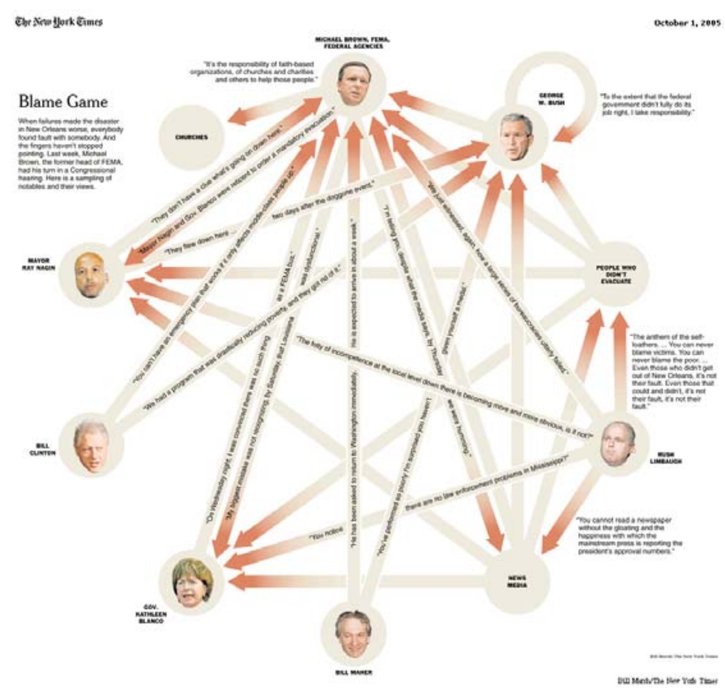
Subjective Language

- 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.

+



Fine-grained Opinion Extraction

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.

Fine-grained Opinion Extraction

Australian press has launched a bitter attack on Italy after seeing their beloved Socceroos eliminated 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 game with Ukraine. He hailed 10-man Italy's determination to beat Australia and said the penalty was rightly given.

(An Aside)

- **Rely on human judgments to identify subjective language**
- **Definitions and many examples provided**
  - See Wiebe, Wilson, & Cardie [LRE, 2004]
- **Trained annotators**
- **Inter-annotator agreement measured**

Fine-grained Opinions

"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)

Opinion Frame
Polarity: negative sentiment
Intensity: high
Source: "The Australian Press"
Target: "Italy"

Example – fine-grained opinions

opinion frame
opinion frame
opinion frame
opinion frame

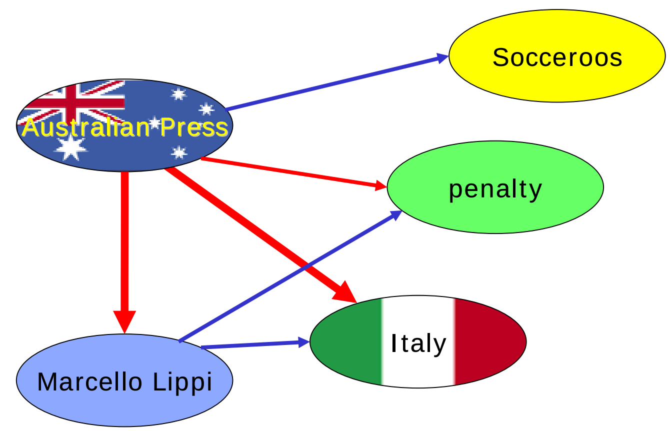
Australian press has launched a bitter attack on Italy after seeing their beloved Socceroos lose on a controversial late penalty. Italian coach Lippi has also been blasted for his comments after the game.

opinion frame
opinion frame

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.

opinion frame

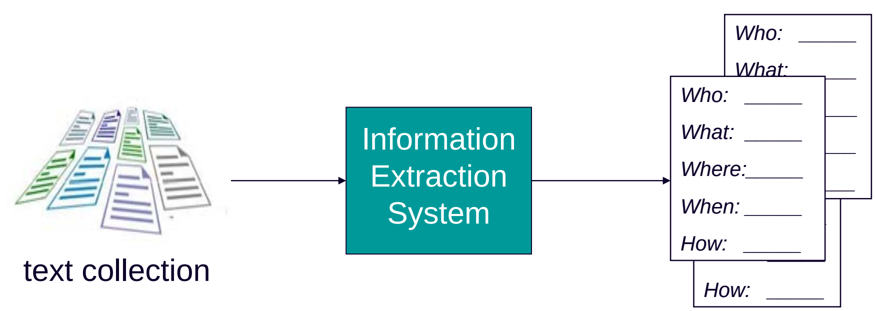
Example – Opinion Summary



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

Information extraction



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"

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

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**

Information extraction

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

Natural disasters example

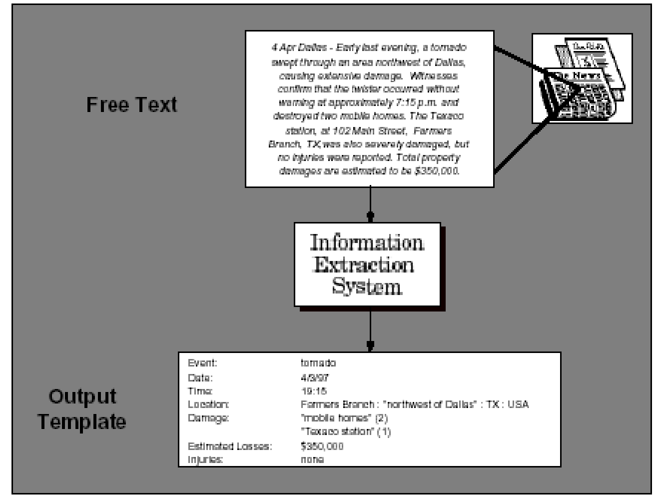
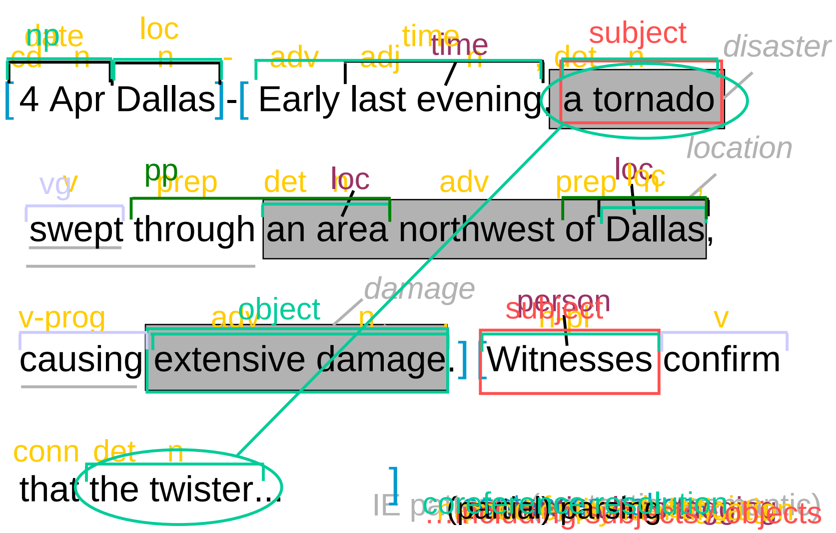
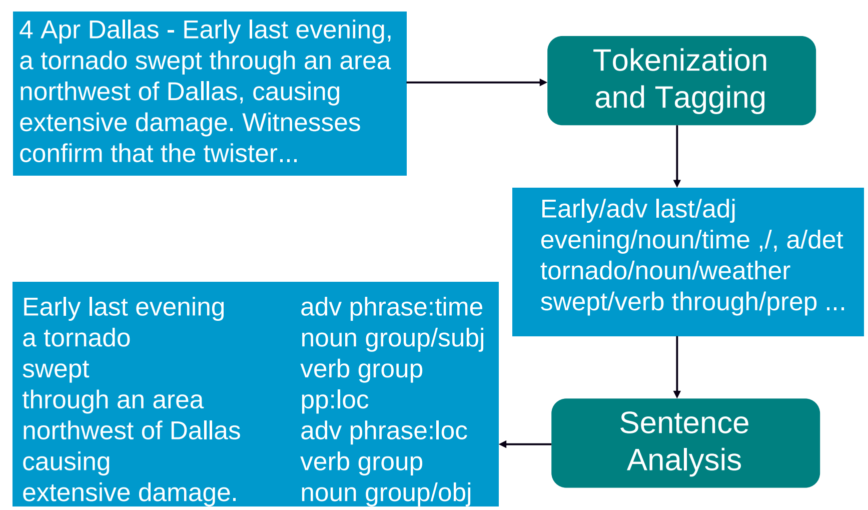


Figure 1: Information Extraction System in the Domain of Natural Disasters.

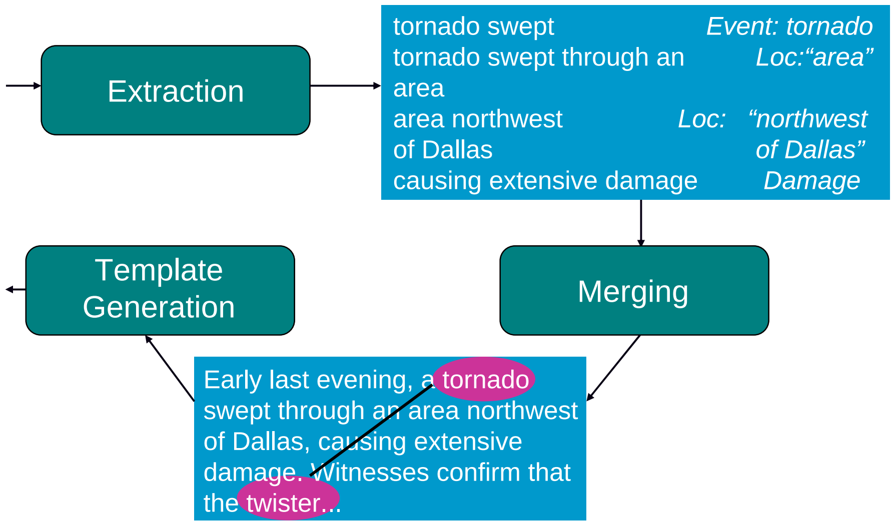
IE system components



Stages of processing



Stages of processing



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** **<u>two mobile homes</u>***.

Pattern:

Trigger: "destroyed"

condition: active voice verb?

Slot: Damaged-Object

Position: direct-object

condition: DO is a physical-object?

from Cardie [1997]

Issues for learning extraction patterns

- **Training data is difficult to obtain**
  - IE "answer keys" provide supervisory information --- string to be extracted and its label
  - Not always supervisory information for learning "set fills"
  - Application of standard "off-the-shelf" learning algorithms is not always straightforward
  - Training examples must encode the output of earlier levels of syntactic and semantic analysis
    - No standard training set available
    - When earlier components change, examples must be regenerated

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>	<victim> was **murdered**
<subject> <active-verb>	<perpetrator> **bombed**
<subject> <infinitival-verb>	<perpetrator> attempted to **kill**
<subject> <auxiliary-verb>+<noun>	<victim> was **victim**
***<passive-verb> <dobj>**	**killed** <victim>
<active-verb> <dobj>	**bombed** <target>
<infinitive> <dobj>	**to kill** <victim>
<verb>+<infinitive> <dobj>	threatened to **attack** <target>
<gerund> <obj>	**killing** <victim>
<noun>+ <auxiliary> <dobj>	**fatality** was <victim>
<noun>+<prep> <np>	**bomb** against <target>
<active-verb>+<prep> <np>	**killed** with <instrument>
<passive-verb>+<prep> <np>	was **aimed** at <target>