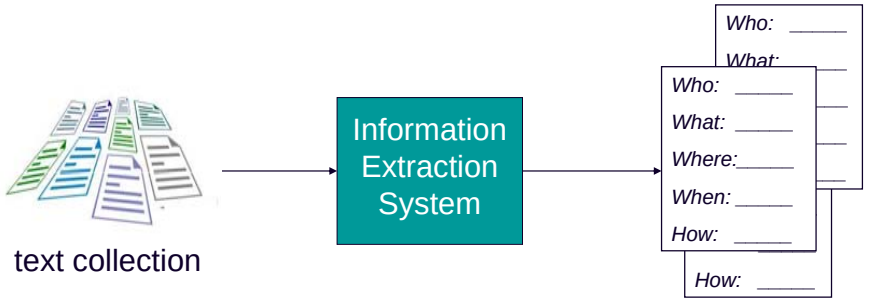
Information Extraction

- **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**.

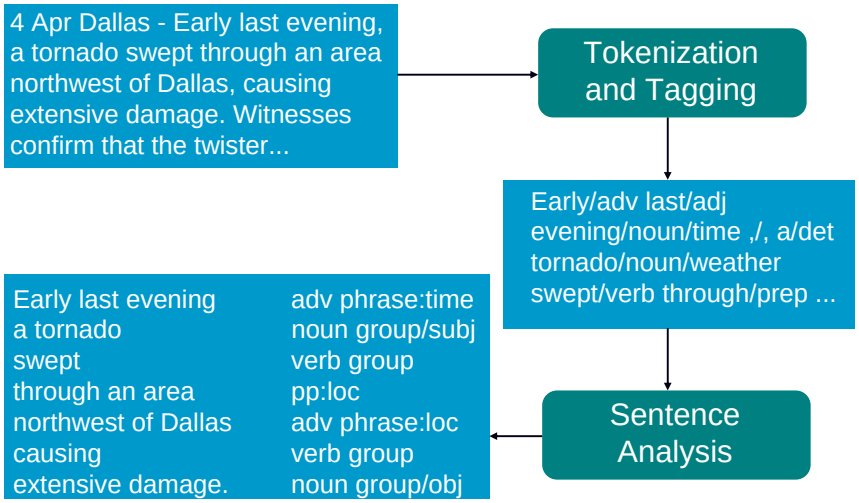
"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE **POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THIS ASSASSINATION** TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

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.**

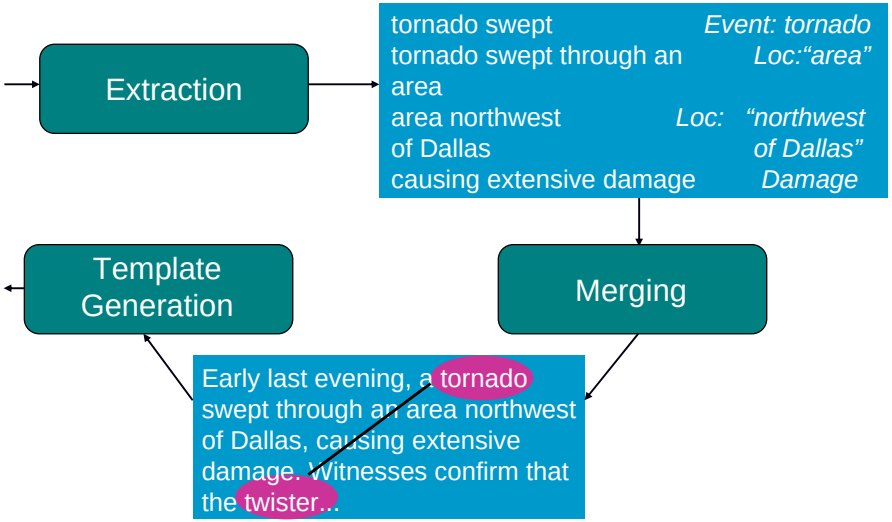
IE system: output

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"

Stages of processing



Stages of processing



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."

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: 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 type 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: physical-object?

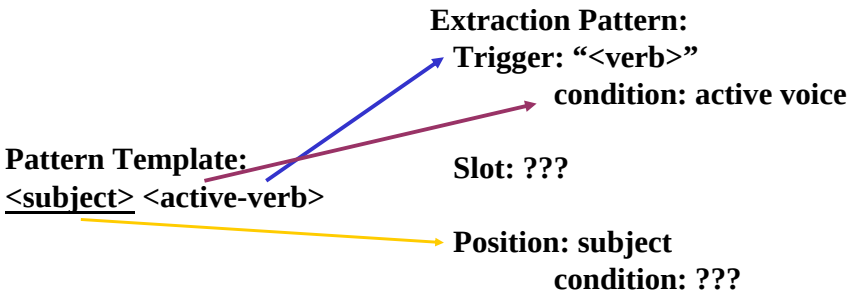
Autoslog (Riloff & Lehnert, 1993)

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>

Template → Extraction Pattern



Semantic restrictions

- **Perpetrator**
  - Person, government, terrorist organization
- **Target (damaged-object)**
  - Building, vehicle, physical-object
- **Victim**
  - Person
- **Location**
  - Location
- **Date**
  - Date
- **Instrument**
  - Weapon

Template → Extraction Pattern

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

damaged-object

Pattern:

Trigger: "<verb>"

condition: active voice

Slot: <slot-type> of <target-np>

Position: direct-object

condition: <<semantic class> of <slot-type>>

Semantic restrictions table

Autoslog algorithm

- **For each annotated "string fill", *s*, in the training data**
  - (Shallow) parse the sentence that contains *s.*
  - Apply the syntactic pattern templates in order. Execute the first one that applies to determine:
    - the *trigger* word
    - the triggering *constraints* (syntactic)
    - the *position* of phrase to be extracted (grammatical role)
  - Determine *slot type*
    - The annotated slot type for *s* in the training corpus
  - Determine the *semantic constraints*
    - Defined a priori based on typical semantic class of fillers
  - Create and save the extraction pattern

Example

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

damaged-object

Pattern:
Trigger: "<verb>"
condition: active voice
Slot: <slot-type> of <target-np>
Position: direct-object
condition: <<semantic class> of <slot-type>>

Instantiation:
Trigger: "destroyed"
condition: active voice verb?
Slot: Damaged-Object
Position: direct-object
condition: physical-object?

Autoslog algorithm

- **Domain-independent**
  - So require little modification when switching domains
- **Requires (minimally) a partial parser**
- **Assumes semantic category(ies) for each slot are known, and all potential slot fillers can be tested w.r.t. them**

Exercise: changes in management

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 date
chairman, who stepped down effective Feb. 2.

Learned terrorism patterns

- **<victim> was murdered**
- **<perpetrator> bombed**
- **<perpetrator> attempted to kill**
- **was aimed at <target>**

Bad patterns are possible

- **took <victim>**

victim

They took 2-year-old Gilberto Molasco, son of Patricio Rodriquez, and 17-year-old Andres Argueta, son of Ernesto Argueta.

Natural disasters patterns

- Yesterday's earthquake registered 6.9 on the Richter scale.
  - <subject> = disaster-event (earthquake) registered (active)
  - registered (active) <direct obj> = magnitude
- measuring 6.9 …
  - measuring (gerund) <direct obj> = magnitude
- …sending medical aid to Afghanistan…
- …sending medical aid to earthquake victims
  - aid (noun)…in/to/for (prep) <obj> = disaster-event-location/victim

Advantages and Disadvantages

- **Learns bad patterns as well as good patterns**
  - Too general (e.g. triggered by "is" or "are" or by verbs not tied to the domain)
  - Too specific
  - Just plain wrong
    - Parsing errors
    - Target NPs occur in a prepositional phrase and Autoslog can't determine the trigger (e.g. is it the preceding verb or the preceding NP?)
- **Requires that a person review the proposed extraction patterns, discarding bad ones**
- **No computational linguist needed (?)**
- **Reduced human effort from 1200-1500 hours to ~4.5 hours**

Results

- **1500 texts, 1258 answer keys**
- **4780 slots (6 types)**
- **Autoslog generated 1237 patterns**
- **After human filtering: 450 patterns**
- **Compare to manually built patterns**

System/Data Set	Recall	Precision	F-measure
Manual/TST3	46	56	50.51
Autoslog/TST3	43	56	48.65
Manual/TST4	44	40	41.90
Autoslog/TST4	39	45	41.79