Information Extraction

- **Today**
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification

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]

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>

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

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

IE Example: Input Text

SANTIAGO, 10 NOV 88 (QUE PASA) -- [TEXT] [CONTINUED]

… THE PLENUM OF THE SOCIALIST PARTY [PS]-ALMEYDA WAS, OF COURSE, THE MOST EAGERLY ANTICIPATED…THEY AMBITIOUSLY FELT THAT THIS WAS THE OPPORTUNITY TO REMOVE SOME STRATEGIC OBSTACLES, SORT OF LIKE **KILLING TWO BIRDS WITH ONE STONE:** REGISTRATION AND THE SOUGHT-AFTER SOCIALIST UNITY...

IE Example: Output Template

1. DATE	10 NOV 88
2. LOCATION	CHILE: SANTIAGO (CITY)
3. TYPE	MURDER
4. STAGE OF EXECUTION	ACCOMPLISHED
5. INCIDENT CATEGORY	TERRORIST ACT
6. PERP: INDIVIDUAL ID	"THEY"
7. PERP: CONFIDENCE	REPORTED AS FACT
9. HUM TGT: DESCRIPTION	"BIRDS"
10. HUM TGT: TYPE	CIVILIAN: "BIRDS"
11. HUM TGT: NUMBER	2: "BIRDS"
12. EFFECT OF INCIDENT	DEATH: "BIRDS”
13. INSTRUMENT	STONE

Information Extraction

- **Today**
  - Learning approaches
    - ➡ Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification

Autoslog-TS

- **Largely unsupervised**
- **Two sets of documents: relevant, not relevant**
- **Apply pattern templates to extract every NP in the texts**
- **Compute *relevance rate* for each pattern *i* :**

Pr (relevant text | text contains i) =
freq of *i* in relevant texts / frequency of *i* in corpus

- **Sort patterns according to relevance rate and frequency**

relevance rate * log (freq)

Autoslog-TS

- Human review of learned patterns is still required
- Also requires, for each pattern, the manual labeling of the semantic category of the extracted slot filler

Information Extraction

- **Today**
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification

Covering algorithms

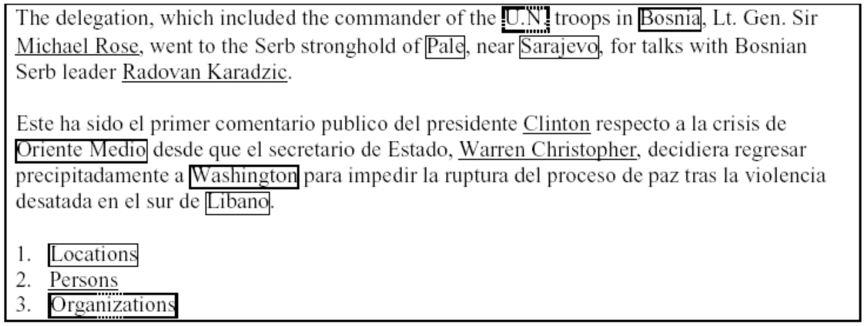
- **E.g. Crystal** [Soderland et al. 1995]
  - Allows for more complicated patterns
    - Can test target NP or any constituent in its context for
      - presence of any word or sequence of words
      - semantic class of heads or modifiers
- **Crystal is a “covering” algorithm**
- **Successively generalizes the patterns derived from input examples until the generalization produces errors**

Information Extraction

- **Today**
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity extraction

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.

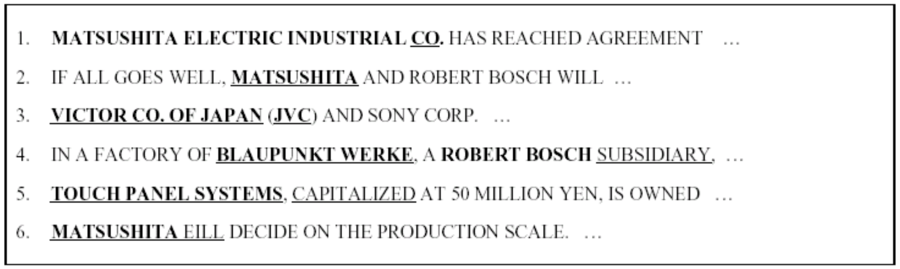
1. Locations
2. Persons
3. Organizations

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.

NE Results Using HMM's

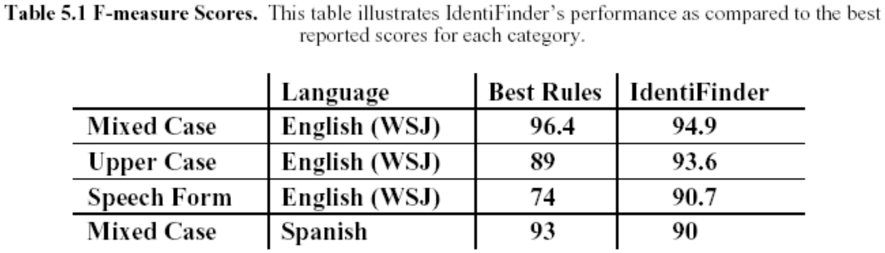


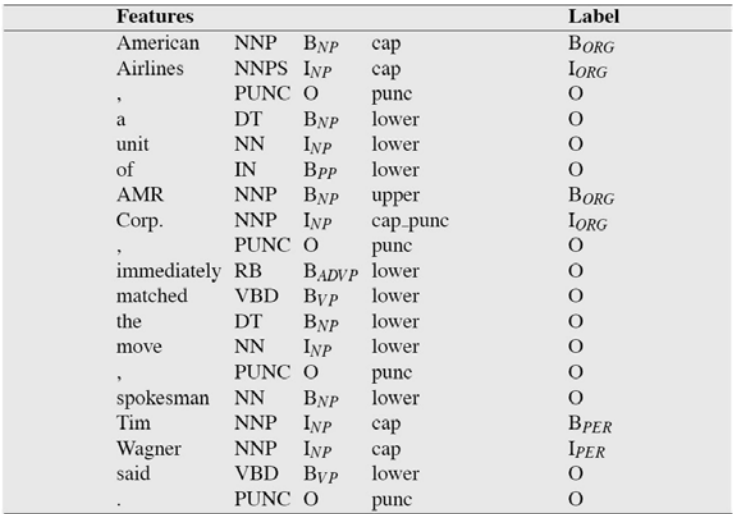
Table 5.1 F-measure Scores. This table illustrates IdentiFinder's performance as compared to the best reported scores for each category.

	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

Class/Tag Values

- **B – begins a PersonName, Loc, etc.**
- **I – inside a PersonName, Loc, etc.**
- **O – outside a PersonName, Loc, etc.**

HMMs for entity detection



Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Figure, copyright J&M 2nd ed