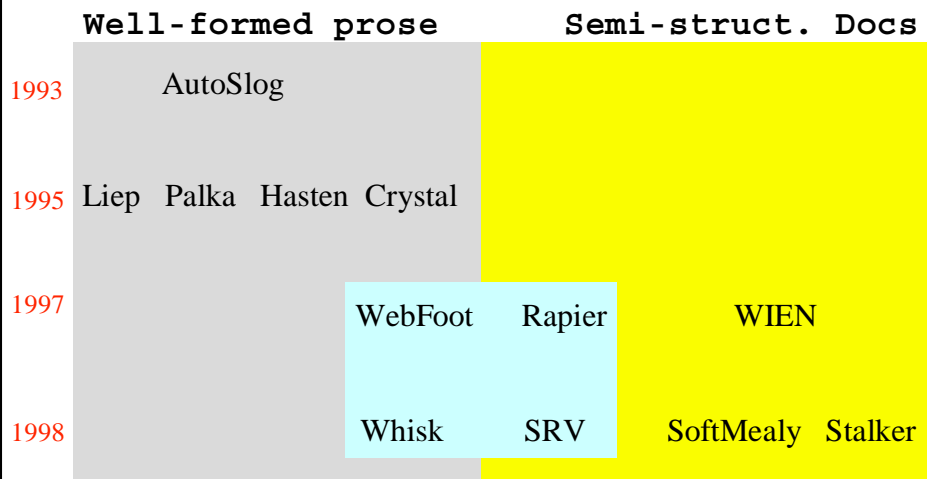

Machine Learning for IE

Ion Muslea
muslea@isi.edu

Introduction

- Information Extraction:
 - GIVEN collection of documents,
IDENTIFY relevant data in each document
 - Examples:
 - extract perpetrators & targets of terrorist attacks from a set of newspaper articles
 - extract titles and prices of Tom Clancy books from Amazon & BN
- systems that LEARN the extraction rules
- compare & contrast salient features

Typical Picture: Extraction Domains

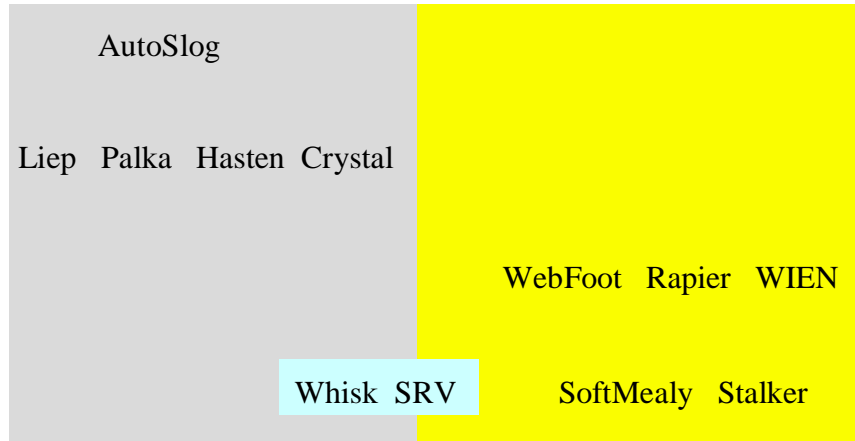


System Categorization

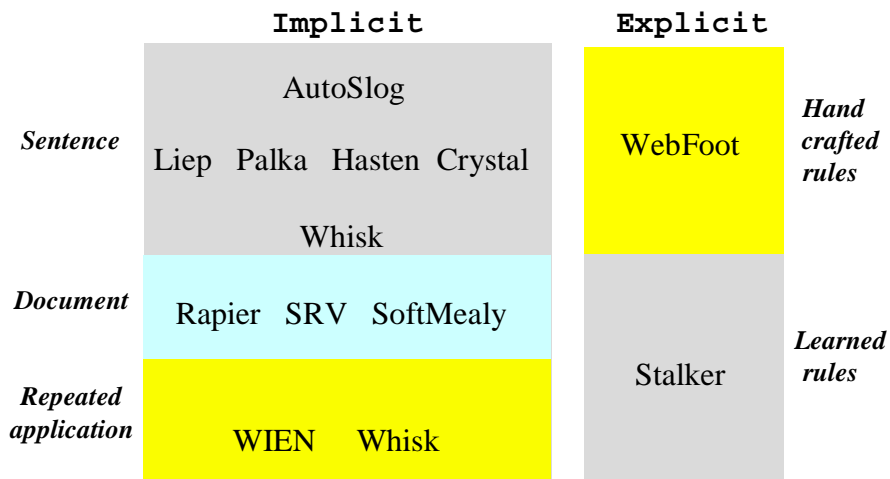
- **What** features are used in the rules ?
- **Where / how** are the rules applied ?

Syntactic Constraints

Use English Syntax **Learn** Domain-specific Syntactic Constructs



Document Segmentation



MUC Systems

- AutoSlog
- Liep
- Palka
- Hasten
- Crystal
 - WebFoot

AutoSlog [Riloff 1993]

The Parliament building was bombed by Carlos.

- CONCEPT NODE:
 - Name: target-subject-passive-verb-bombed
 - Trigger: bombed
 - Variable Slots: (target (*S* 1))
 - Constraints: (class phys-target *S*)
 - Constant Slots: (type bombing)
 - Enabling Conditions: ((passive))

LIEP [Huffman 1995]

The Parliament was bombed by the guerrillas.

- TARGET-was-bombed-by-PERPETRATOR:
noun-group(**TRGT**, head(is-a(physical-target))),
noun-group(**PERP**, head(is-a(perpetrator))),
verb-group(VG, type(passive), head(bombed)),
preposition(PREP, head(by)),
subject(**TRGT**, VG),
post-verbal-prep(VG, PREP),
prep-object(PREP, **PERP**),
⇒ bombing-event(BE, target(**TRGT**), agent(**PERP**)).

PALKA & HASTEN

The Parliament was bombed by the guerrillas.

PALKA [Kim&Moldovan, 1995]:

(BOMBING target: *phys-object* agent: *perp*
pattern: ((*target*) was bombed by (*perp*)))

HASTEN [Krupka 1995]:

BOMBING: **TARGET:** NP "semantic = *phys-object*"
ANCHOR: VG "root = BOMB"
PERP: NP "semantic = *terrorist-group*"

CRYSTAL [Soderland 1995]

The Parliament building was bombed by *Carlos*.

CONCEPT TYPE: BUILDING BOMBING

SUBJECT: Classes include: <PhysicalTarget>

Terms include: BUILDING

Extract: target

VERB: Root: BOMB

Mode: passive

PREPOS-PHRASE: Preposition: BY

Classes include: <PersonName>

Extract: perpetrator

WebFoot [Soderland 1995]

SEGMENTED DOCUMENT:

<segm> field1: <HEAD> LA Forecast </HEAD> </segm>

<segm> field1: .MONDAY... field2: CLOUDY </segm>

<segm> field1: .TUESDAY... field2: SUNNY </segm>

Concept type: FORECAST

Constraints:

FIELD: Classes include: <Day>

Terms include: ``.''' , ``...''

Extract: day

FIELD: Classes include: <WeatherCondition>

Extract: conditions

Post-MUC: WHISK [Soderland '99]

DOCUMENT:

<p> Capitol Hill- 1 br twnhme. D/W W/D. Pkg incl
\$675. 3 BR upper flr no gar. \$995. (206)999-9999

WHISK pattern: * (<Digit>) 'BR' * '\$' (<Nmb>)

OUTPUT: Rental {Bedrooms @1} {Price @2}

Extracted Data: <Bedrooms: *1* Price: *675*>
<Bedrooms:*3* Price: *995*>

WHISK [Soderland 1999]

The Parliament building was bombed by *Carlos*.

WHISK Rule:

* (*PhysObj*) * @Passive *F 'bombed' * {PP 'by' *F (*Person*) }

Comparison

	Exact phrase extract.	Can Ignore Syntactic Constraints	Beyond-Slot Semantic Constraints	Multi-slot
AutoSlog	-	-	-	-
LIEP	✓	✓	-	✓
PALKA	-	-	-	✓
HASTEN	✓	✓	-	✓
CRYSTAL	-	-	✓	✓
WHISK	✓	✓	✓	✓

Granularity of the Extraction

The Parliament building, which was built in 1812 , was bombed ...

Comparison

	Exact phrase extract.	Can Ignore Syntactic Constraints	Beyond-Slot Semantic Constraints	Multi-slot
AutoSlog	-	-	-	-
LIEP	✓	✓	-	✓
PALKA	-	-	-	✓
HASTEN	✓	✓	-	✓
CRYSTAL	-	-	✓	✓
WHISK	✓	✓	✓	✓

Post-MUC Systems

- Whisk
- Rapier
- SRV
- WIEN
- SoftMealy
- Stalker

RAPIER (Califf & Mooney 1997)

DOCUMENT:

C Programmer. 38-44K. Leading *data mining* firm in need of an energetic individual to fill the following position: ...

AREA extraction pattern:

Pre-filler pattern: *word:* leading
Filler pattern: *list:* len: 2
tags: [nn, nns]
Post-filler pattern: *word:* [firm, company]

SRV [Freitag 1998]

DOC-1: ... to purchase 4.5 mln *Trilogy* shares at ...

DOC-2: ... acquire another 2.4 mln *Roach* shares ...

Acquisition:

```
- length( = , 1 ),  
  some(?A [] capitalized true),  
  some(?A [next-token] all-lower-case true),  
  some(?A [right-AN] wn-word 'stock').
```

Wien [Kushmerick, 97] & SoftMealy [Hsu+Dung, 98]

D1: 1. Joe's: (313) 323-5545 2. Li's: (406) 545-2020

D2: 1. KFC: 818-224-4000 2. Rome: (656) 987-1212

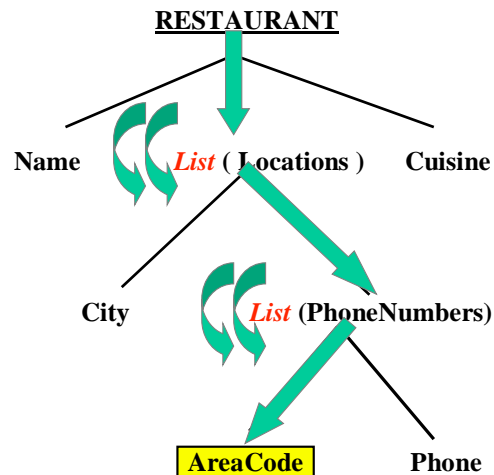
WIEN rule: * '.' (*) ':' * '(' (*) ')'

SoftMealy rule: * '.' (*) EITHER ':' (<Nmb>) '-'
OR ':' * '(' (<Nmb>) ')'

Output: Restaurant {Name @1} {AreaCode @2}

STALKER: *explicit output schema*

Name:	Gino's
Cuisine:	Italian
Locations:	
Venice	(310) 123-4567, (800) 888-4412.
L.A.	(213) 987-6543.
Encino	(818) 999-4567, (888) 727-3131.



Comparison

	Nested Data	Single slot	Beyond slots	Disj. rules	Orthogr. Constr.	English Syntax	Semantic Constrs.	Other
Wien	☑	-	-	-	-	-	-	-
Rapier	-	☑	-	☑	-	-	<i>WordNet</i>	<i>POS, length</i>
SRV	-	☑	☑	☑	☑	☑	<i>WordNet</i>	<i>length</i>
Whisk	-	-	☑	☑	☑	☑	user-def	-
SoftM	☑	-	-	☑	☑	-	user-def	-
Stalker	☑	☑	☑	☑	☑	-	user-def	-

Trends

- **combining several types of rules ([Freitag 1998])**

VS

extraction rules for all genres (WHISK)

- **explicit output schema + single-slot rules**

VS

implicit schema + multi-slot rules

- **“symbolic learners”**

VS

“probabilistic learners” (HMMs) e.g. [McCallum 2000, 2001]

CLAIM

Convergence towards similar sets of
extraction features.

STALKER (Muslea et al., 1998)

Document:

<p>Name: ROMA <p><hr> Cuisine: ITALIAN <p>

STALKER rule:

* 'Name:' (AllCaps) '<p>' * 'Cuisine:' *

WIEN (Kushmerick, 1997)

D1:

<hr> <p>Name: ROMA <p><hr> Cuisine: ITALIAN

<hr> <p>Name: KFC <p><hr> Cuisine: FAST FOOD

WIEN rule:

* 'Name:' (*) '<p>' * ':' (*) '
'

Output:

Restaurant {Name @1} {Cuisine @2}

