

Information Retrieval

Lecture 5: Information Extraction

Computer Science Tripos Part II



UNIVERSITY OF
CAMBRIDGE

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk

Information Extraction: the task

2

-
- Identify instances of a particular class of events or relationships in a natural language text
 - Limited semantic range of events/relationships (domain-dependence)
 - Extract the relevant arguments of the event or relationship into pre-existing “templates” (tabular data structures)
 - MUC (Message Understanding Conference; NIST) 1986-97 competitive evaluation

-
- 1980s and before: lexico-semantic patterns written by hand (FRUMP, satellite reports, patient discharge summaries...)
 - 1987 **First MUC** (Message Understanding Conference); domain: naval sightings
 - 1989 **Second MUC**; domain: naval sightings
 - 1991 **Third MUC**; domain: terrorist acts
 - Winner (SRI) used partial parsing
 - 1992 **Fourth MUC**; domain: terrorist acts
 - 1993 **Fifth MUC**; domain: joint ventures/electronic circuit fabrication
 - Performance of best systems ~ 40% R, 50% P (Humans in 60-80% range)
 - Lehnert et al.: first bootstrapping method

-
- 1995 **Sixth MUC**; domain: labour unit contract negotiations/changes in corporate executive management personnel
 - Encourage more portability and deeper understanding
 - Separate tasks into
 - * **NE**: Named Entity
 - * **CO**: Coreference
 - * **TE**: Template Element
 - * **ST**: Scenario Templates
 - 1995: IE for summarisation (Radev and McKeown)
 - 1998: **Seventh MUC**; domain: satellite rocket launch events
 - Mikheev et al., hybrid methods for NE
 - 2003: **CoNLL** NE recognition task; similar training data to MUC

- Participants get a description of the scenario and a training corpus (a set of documents and the templates to be extracted from these)
- 1-6 months time to adapt systems to the new scenario
- NIST analysts manually fill templates of test corpus (“answer key”)
- Test corpus delivered; systems run at home
- Automatic comparison of system response with answer key
- Primary scores: precision and recall
- Participants present paper at conference in spring after competition
- Show system’s workings on predefined “walk through” example

Template example (MUC-3)

0	MESSAGE ID	TST1-MUC3-0080
1	TEMPLATE ID	1
2	DATE OF INCIDENT	03 APR 90
3	TYPE OF INCIDENT	KIDNAPPING
4	CATEGORY OF INCIDENT	TERRORIST ACT
5	PERPETRATOR: ID OF INDIV(S)	“THREE HEAVILY ARMED MEN”
6	PERPETRATOR: ID OF ORG(S)	“THE EXTRADITABLES”
7	PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: “THE EXTRADITABLES”
8	PHYSICAL TARGET: ID(S)	*
9	PHYSICAL TARGET: TOTAL NUM	*
10	PHYSICAL TARGET: TYPE(S)	*
11	HUMAN TARGET: ID(S)	“FEDERICO ESTRADA VELEZ” (“LIBERAL SENATOR”)
12	HUMAN TARGET: TOTAL NUM	1
13	HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL: “FEDERICO ESTRADA VELEZ”
14	TARGET: FOREIGN NATION(S)	–
15	INSTRUMENT: TYPE(S)	*
16	LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17	EFFECT ON PHYSICAL TARGETS	*
18	EFFECT ON HUMAN TARGETS	*

TST-1-MUC3-0080

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) - [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS WE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND INTO A BLUE RENAULT.

HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT. LAST WEEK, FEDERICO ESTRADA HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

Text Example (MUC-5)

<DOC>

<DOCNO> 0592 </DOCNO>

<DD> NOVEMBER 24, 1989, FRIDAY </DD>

<SO>Copyright (c) 1989 Jiji Press Ltd.;</SO>

<TXT>

BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO JAPAN.

THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO., CAPITALIZED AT 20 MILLION NEW TAIWAN DOLLARS, WILL START PRODUCTION IN JANUARY 1990 WITH PRODUCTION OF 20,000 IRON AND "METAL WOOD" CLUBS A MONTH. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 55,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID.

THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID.

BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUBS PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.

WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.

</TXT>

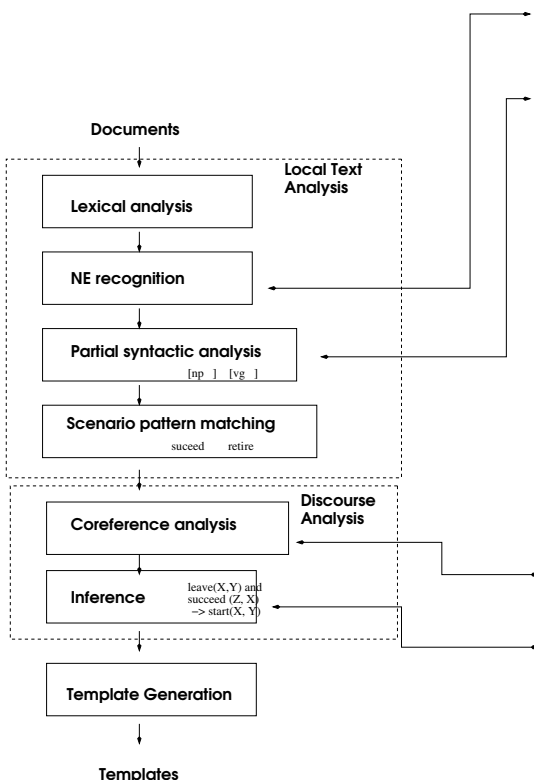
</DOC>

```

<TEMPLATE-0592-1> :=
  DOC NR: 0592
  DOC DATE: 241189
  DOCUMENT SOURCE: "Jiji Press Ltd."
  CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
<TIE_UP_RELATIONSHIP-0592-1> :=
  TIE-UP STATUS: EXISTING
  ENTITY:<ENTITY-0592-1>
    <ENTITY-0592-2>
    <ENTITY-0592-3>
  JOINT VENTURE CO:<ENTITY-0592-4>
  OWNERSHIP: <OWNERSHIP-0592-1>
  ACTIVITY:<ACTIVITY-0592-1>
<ENTITY-0592-1>:=
  NAME: BRIDGESTONE SPORTS CO
  ALIASES: "BRIDGESTONE SPORTS"
    "BRIDGESTON SPORTS"
  NATIONALITY: Japan (COUNTRY)
  TYPE: COMPANY
  ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-2>:=
  NAME: UNION PRECISION CASTING CO
  ALIASES: "UNION PRECISION CASTING"
    "BRIDGESTON SPORTS"
  LOCATION: Taiwan (COUNTRY)
  NATIONALITY: Taiwan (COUNTRY)
  TYPE: COMPANY
  ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-3>:=
  NAME: TAGA CO
  NATIONALITY: Japan (COUNTRY)
  TYPE: COMPANY
  ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>

<ENTITY-0592-4>:=
  NAME: BRIDGESTONE SPORTS TAIWAN CO
  ALIASES: "UNION PRECISION CASTING"
    "BRIDGESTON SPORTS"
  LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
  TYPE: COMPANY
  ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<INDUSTRY-0592-1>:=
  INDUSTRY-TYPE: PRODUCTION
  PRODUCT/SERVICE: (CODE 39 "20,000 IRON AND 'METAL WOOD")
[CLUBS])
<ENTITY_RELATIONSHIP-0592-1>:=
  ENTITY1: <ENTITY-0592-1>
    <ENTITY-0592-2>
    <ENTITY-0592-3>
  ENTITY2: <ENTITY-0592-4>
  REL OF ENTITY2 TO ENTITY1: CHILD
  STATUS: CURRENT
<ACTIVITY-0592-1>:=
  INDUSTRY: <INDUSTRY-0592-1>
  ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
  START TIME: <TIME-0592-1>
<TIME-0592-1>:=
  DURING: 0190
<OWNERSHIP-0592-1>:=
  OWNED: <ENTITY-0592-4>
  TOTAL-CAPITALIZATION: 20000000 TWD
  OWNERSHIP-%: (<ENTITY-0592-3> 10)
    (<ENTITY-0592-2> 15)
    (<ENTITY-0592-1> 75)
  
```

NYU's IE system



Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb.

Sam Schwartz (person) retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc (organisation). He will be succeeded by Harry Himmelfarb(person).

[np: e1 Sam Schwartz (person)] [vg retired] as [np: e2 executive vice president] of [np: e3 the famous hot dog manufacturer], [np:e4 Hupplewhite Inc (organisation)]. [np: e5 He] [vg will be succeeded] by [np: e6 Harry Himmelfarb(person)].

e1	type:person	name:"Sam Schwartz"
e2	type:position	value:"executive vice president"
e3	type:manufacturer	
e4	type:company	name: "Hupplewhite Inc."
e5	type:person	name: "Harry Himmelfarb"
e6	type:person	name: "Harry Himmelfarb"

e2	type:position	value:"executive vice president" company:e3
e3 = e4		

e7	leave-job	person:e1 position:e2
e8	succeed	person1:e6 person2:e5

e5 = e1		
e9	start-job	person:e6 position e2

EVENT:	leave job
PERSON:	Sam Schwartz
POSITION:	executive vice president
COMPANY:	Hupplewhite Inc.

EVENT:	start job
PERSON:	Harry Himmelfarb
POSITION:	executive vice president
COMPANY:	Hupplewhite Inc.

- NE types:
 - ENAMEX (type= person, organisation, location)
 - TIMEX (type= time, date)
 - NUMEX (type= money, percent)
- Allowed to use **gazetteers** (fixed list containing names of a certain type, e.g. countries, last names, titles, state names, rivers...)
- ENAMEX is harder, more context dependent than TIMEX and NUMEX:
 - Is **Granada** a COMPANY or a LOCATION?
 - Is **Washington** a PERSON or a LOCATION?
 - Is **Arthur Anderson** a PERSON or an ORGANISATION?

- NE markup with subtypes:
<ENAMEX TYPE='PERSON'>Flavel Donne</ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends</ENAMEX>, which has been based in <ENAMEX TYPE='LOCATION'>Little Spring</ENAMEX> since <TIMEX>July 1998</TIMEX>.
- Most systems use manually written regular expressions
 - Rules about mid initials, postfixes, titles
 - Gazetteers of common first names
 - Acronyms: Hewlett Packard Inc. → HP

PATTERN: “president of <company>” matches

executive vice president of Hupplewhite

- Gazetteer of full names impossible and not useful, as both first and last names can occur on their own
- **Last name** gazetteer impractical
 - Almost infinite set of name patterns possible: last names are productive (1.5M surnames in US alone)
 - Overlap with common nouns/verbs/adjectives
 - * First 2 pages of Cambridge phone book include 237 names
 - * Of those, 6 (2.5%) are common nouns: Abbey, Abbot, Acres, Afford, Airs, Alabaster
- **First name** gazetteer less impractical, but still not foolproof
 - First names can be surprising, eg. MUC-7 walk-through example: “Llennel Evangelista”
 - First names are productive, eg. Moonunit Zappa, Apple Paltrow . . .

- – Overlap with common nouns:
 - * River and Rain Phoenix, Moon Unit Zappa, Apple Paltrow
 - * “Virtue names”: Grace (134), Joy (390), Charity (480), Chastity (983), Constance, Destiny
 - * “Month names”: June, April, May
 - * “Flower names”: Rose, Daisy, Lily, Erica, Iris . . .
 - * From US Social Security Administration’s list of most popular girls’ names in 1990, with rank:
 - Amber (16), Crystal (41), Jordan (59), Jade (224), Summer (291), Ruby (300), Diamond (450), Infant (455), Precious (472), Genesis (528), Paris (573), Princess (771), Heaven (902), Baby (924) . . .
- Additional problem: non-English names alliterated into English; variant spellings
- Complicated name patterns with titles: Sammy Davis Jr, HRH The Prince of Wales, Dr. John T. Maxwell III

- Ambiguity of name types: *Columbia* (Org.) vs. (British) *Columbia* (Location) vs. *Columbia* (Space shuttle)
- Company names often use common nouns (“Next”, “Boots”, “Thinking Machines”...) and can occur in variations (“Peter Stuyvesant”, “Stuyvesant”)
- Coordination problems/ left boundary problems:
 - One or two entities in *China International Trust and Investment Corp invests \$2m in... ?*
 - Unknown word at beginning of potential name: in or out?
Suspended Ceilings Inc vs *Yesterday Ceilings Inc*
Mason, Daily and Partners vs. *Unfortunately, Daily and Partners*
- Experiments show: simple gazetteers fine for locations (90%P/80%R) but not for person and organisations (80%P/50%R)

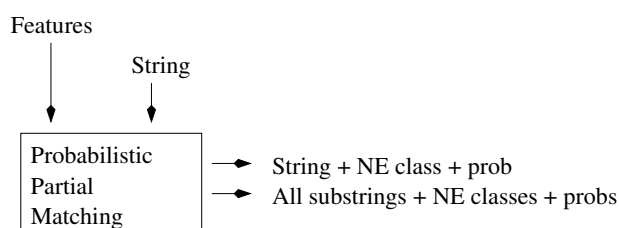
- Staged combination of rule-based system with probabilistic partial matching
- Use machine learning to decide type of NE
- Use internal phrase structure of name
- Make high-precision decisions first
- Keep off decision about unsure items until all evidence has been seen
- Assume: one name type per discourse (article)
 - unless signalled by writer with additional context information

Rule	Assign	Example
(Xx+)+ (is ,) a? JJ* PROF	PERS	Yuri Gromov, a former director
(Xx+)+ is? a? JJ* REL	PERS	John White is beloved brother
(Xx+)+ himself	PERS	White himself
(Xx+)+, DD+ ,	PERS	White, 33,
share in (Xx+)+	ORG	shares in Trinity Motors
(Xx+)+ Inc.	ORG	Hummingbird Inc.
PROF (of at with) (Xx+)+	ORG	director of Trinity Motors
(Xx+)+ (region area)	LOC	Lower Beribidjan area

Mikheev et al. (1998): ML; External and internal evidence 18

External:

- Position in sentence (sentence initial)
- Word exists in lexicon in lowercase
- Word seen in lowercase in document



Internal:

- Contains any non-alpha characters
- Number of words it consists of
- Suffix, Prefix
- Adjectives ending in “an” or “ese” + whose root is in Gazetteer

1. **Apply Grammar Rule Set 1** (“Sure fire” rules)
→ tag as definite NEs of given type
2. **Use ML for variants** (probabilistic partial match)
 - Generate all possible substrings of sure-fire tagged NEs:
 - *Adam Kluver Ltd* → *Adam Kluver, Adam Ltd, Kluver Ltd*
 - ME model gives probability for possible string and NE type
 - Tag all occurrences of NE in text (over prob. threshold) with type
3. **Apply Grammar Rule Set 2** (Relaxed rules)
 - Mark anything that looks like a PERSON (using name grammar)
 - Resolve coordination, genitives, sentence initial capitalized modifiers
 - Coordinated or possessive name parts, or rest of sentence initial coordinated name seen on their own? If not, assume one name (*Murdoch’s News Corp, Daily, Bridge and Mason*)
4. **Apply ML again (for new variants)**
 - X and Y are of same type → resolved type ‘ ‘Un7ited States and Russia’ ’
5. **Apply specialised ME model to title** (capitalisation, different syntax).

MURDOCH SATELLITE CRASH UNDER FBI INVESTIGATION

London and Tomsk. The crash of Rupert Murdoch Inc’s news satellite yesterday is now under investigation by Murdoch and by the Siberian state police. Clarity J. White, vice president of Hot Start, the company which produced the satellite’s ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. Investigator Robin Black, 33, who investigates the crash for the FBI, recently arrived by train at the crash site in the Tomsk region. Neither White nor Black were available for comment today; Murdoch have announced a press conference for tomorrow.

LONDON and TOMSK	Org
Rupert Murdoch	Person
Murdoch	Person
Neither White	Person
Investigator Robin Black	Person

Additional problem: `Black` and `White` have last names which overlap with adjectives and first names which overlap with common nouns (`Robin` and `Clarity`), thus they cannot be in a gazetteer.

MURDOCH SATELLITE CRASH UNDER FBI INVESTIGATION

London and Tomsk. The crash of [Rupert Murdoch Inc](#)(ORG)'s news satellite yesterday is now under investigation by Murdoch and by the Siberian state police. Clarity J. White, vice president of [Hot Start](#)(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. [Investigator Robin Black](#)(PERSON), 33, who investigates the crash for the FBI, recently arrived by train at the crash site in the [Tomsk](#)(LOC) region. Neither White nor Black were available for comment today; Murdoch have announced a press conference for tomorrow.

Underlined instances: newly suggested in this round

- Sure fire rules applied
- But exact extend of name not known yet: Investigator Robin Black? Black?

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and [Tomsk](#)(LOC?). The crash of [Rupert Murdoch Inc](#)(ORG?)'s news satellite yesterday is now under investigation by [Murdoch](#)(ORG?) and by the Siberian state police. Clarity J. White, vice president of [Hot Start](#)(ORG?), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. [Investigator Robin Black](#)(PERSON?), 33, recently arrived by train at the crash site in the [Tomsk](#)(LOC?) region. Neither White nor [Black](#)(PERSON?) were available for comment today; [Murdoch](#)(ORG?) have announced a press conference for tomorrow.

Green instances: around from last round

- All instances from last round and their substrings are now hypothesized; they and their context are now subjected to ML

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and [Tomsk](#)(LOC✓). The crash of [Rupert Murdoch Inc](#)(ORG✓)'s news satellite yesterday is now under investigation by [Murdoch](#)(ORG✓) and by the Siberian state police. Clarity J. White, vice president of [Hot Start](#)(ORG✓), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator [Robin Black](#)(PERS✓), 33, recently arrived by train at the crash site in the [Tomsk](#)(LOC✓) region. Neither White nor [Black](#)(PERS✓) were available for comment today; [Murdoch](#)(ORG✓) have announced a press conference for tomorrow.

- ML has reconfirmed some instances (Robin Black) and discarded others (Investigator Robin Black)

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and [Tomsk](#)(LOC). The crash of [Rupert Murdoch Inc](#)(ORG)'s news satellite yesterday is now under investigation by [Murdoch](#)(ORG) and by the Siberian state police. [Clarity J. White](#)(PERS?), vice president of [Hot Start](#)(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator [Robin Black](#)(PERS), 33, recently arrived by train at the crash site in the [Tomsk](#)(LOC) region. [Neither White](#)(PERS?) nor [Black](#)(PERS) were available for comment today; [Murdoch](#)(ORG) have announced a press conference for tomorrow.

- Relaxed rules: Mark everything as a possibility which roughly follows Name shape (blue, underlined)
- (plus confirmed NEs from last round in green)

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

[London](#)(LOC \checkmark) and [Tomsk](#)(LOC). The crash of [Rupert Murdoch Inc](#)(ORG)'s news satellite yesterday is now under investigation by [Murdoch](#)(ORG) and by the Siberian state police. [Clarity J. White](#)(PERS \checkmark), vice president of [Hot Start](#)(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator [Robin Black](#)(PERS), 33, recently arrived by train at the crash site in the [Tomsk](#)(LOC) region. Neither [White](#)(PERS \checkmark) nor [Black](#)(PERS) were available for comment today; [Murdoch](#)(ORG) have announced a press conference for tomorrow.

- Some of these possibilities reconfirmed by ML, others discarded
- "London" found by X and Y rule.
- Missing step: different segmentation and ML for title; 'Murdoch' is found there.

93.39% combined P and R – best and statistically different from next contender

		ORG		PERSON		LOC	
		R	P	R	P	R	P
1	Sure fire rules	42	98	40	99	36	96
2	Partial Match 1	75	98	80	99	69	93
3	Relaxed Rules	83	96	90	98	86	93
4	Partial Match 2	85	96	93	97	88	93
5	Title Assignment	91	95	95	97	95	93

- System design: Keep precision high at all stages, raise recall if possible
- Gazetteers improve performance, but system can determine persons and organizations reasonably well even without any gazetteer (ORG: P86/R85; PERSON: P90/R95), but not locations (P46/R59)

Summary of today

- IE consists of different tasks (as defined by MUC): NE, CO, TE, ST
- Today: NE
 - Principal problems with NE
 - NE with manual rules
 - Mikheev et al. (1998)
 - * Use internal and external evidence
 - * Cascaded design: commit in order of confidence/supportive evidence from text, not in text order!

-
- Mikheev, Moens and Grover (1998). Description of the LTG system. MUC-7 Proceedings.
 - Mikheev, Moens and Grover (1999). Named Entity Recognition without Gazetteers. EACL'99
 - R. Grishman (1997): Information Extraction: Techniques and challenges, in: Information Extraction, Springer Verlag, 1997.