# Information Retrieval

# Lecture 5: Information Extraction

Computer Science Tripos Part II
Lent Term 2004

UNIVERSITY OF
CAMBRIDGE

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk

## Information Extraction: the task

- 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 preexisting "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
- 1889 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 $\sim$ 40% R, 50% P (Humans in 60-80% range)
  - Lehnert et al.: first bootstrapping method

## History of IE, ctd.

- 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.

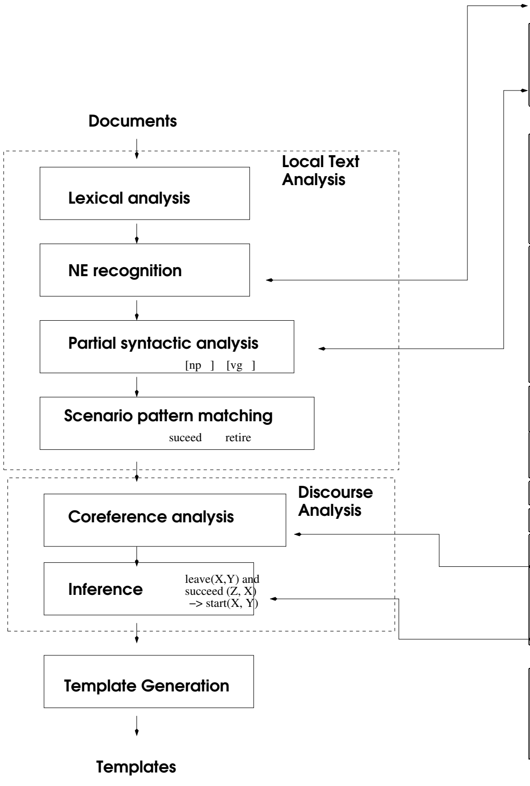
LAST WEEK, FEDERICO ESTRADA HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

## 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 | |
| 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?

## Person names – evidence against gazetteers

- 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 are productive, cf. River and Rain Phoenix, Moonunit Zappa, . . .
  - Overlap with common nouns:
    - ∗ “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) . . .
  - In MUC-7 walk-through example: “Llennel Evangelista”
- Additional problems are variant spellings from non-English names alliterated into English
- Complicated name patterns with titles etc.
  - Sammy Davis Jr
  - HRH The Prince of Wales
  - Dr. John T. Maxwell III
- However, titles are safest of places to pick up new names

## Named Entity recognition – common approaches

- 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>.
- 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*

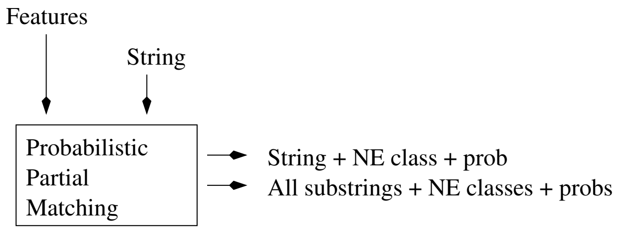
- Ambiguity of name types: *Columbia* (Org.) vs. (British) *Columbia* (Location) vs. *Columbia* (Space shuttle)
- Company names often use common nouns (“Next”, “Boots”, “Thinking Machines”. . . ) and can occur in variations (“Peter Stuyvesant”, “Stuyvesant’)
- Experiments show: simple gazetteers fine for locations (90%P/80%R) but not for person and organisations (80%P/50%R)
- 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*

## Mikheev et al. (1998): Cascading NE

- 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

## External:

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


Features

String

Probabilistic Partial Matching

String + NE class + prob

All substrings + NE classes + probs

## 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

# Mikheev et al. (1998): Algorithm

1. Apply Grammar Rule Set 1 (“Sure fire” rules)
   → tag as definite NEs of given type
2. Use ML for variants (probabilistic partial match)
   - Find all NEs marked up at this stage
   - Generate all possible substrings of words of NEs
     → tag as possible NEs of same type
     – *Adam Kluver Ltd* → *Adam Kluver*, *Adam Ltd*, *Kluver Ltd*
   - Pretrained Maximum Entropy model gives probability for possible NE to actually be that NE
3. Apply Grammar Rule Set 2 (Relaxed rules)
   - Mark anything that looks like a PERSON (using name grammar and list of names)
   - Resolve coordination – have parts been used on their own? If not, assume coordinated name

*. . . , “Suspended Ceiling Contractors Ltd”*)

- Resolve genitive ambiguity: *Murdoch’s News Corp*

4. Apply ML again (for new variants)
   - He worked for X and Y → X and Y are of same type
   - Only system that resolved typo `‘‘Un7ited States and Russia’’` in MUC-7
5. Resolve NEs in title (all words in title are capitalised). Use different ME model trained on titles.

## Examples of sure fire rules

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

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 Sibirian 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.

## Mikheev et al – potential NE hypotheses (all incorrect)

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

Additional problem: Black and White have names which cannot be in gazetteer: first names are common nouns (Robin and Clarity), last names are adjectives.

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 Sibirian 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.

## Mikheev et al – After Step 2 (Partial Match)

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 Sibirian 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.

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 Sibirian 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.

## Mikheev et al – After Step 3

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 Sibirian 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.

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London(LOC√) and Tomsk(LOC). The crash of Rupert Murdoch Inc(ORG)'s news satellite yesterday is now under investigation by Murdoch(ORG) and by the Sibirian 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.

## Mikheev et al: Results in MUC-7

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)

- IR 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!

## Literature

- 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.