Université de Montréal

Information Extraction

Guy Lapalme
Université de Montréal

rali

Information Extraction

- Wide scope definition
  - Any information filtering method from large quantity of text (include IR and text mining)
- More precise definition (*à la* MUC)
  - Identification of instances of specific classes of events or relations in a text and extraction of their arguments
  - Its goal is to *feed* a structured data base from information gathered in the text

rali

Information extraction advantages

- Well defined tasks
- Deals with *real* texts
- Raises difficult and interesting NLP issues
- Can compare system performances with the one of humans

rali

Message Understanding Conferences (MUCs)

- 7 conferences between 1987-1998 by DARPA (MUCK{1,2}, MUC[3-7])
- Development of formal evaluation methods
- Competitions in which participants compare their results between them and against human prepared references
- Short preparation time (~1 month) to stimulate portability

rali

Types of events and relations

MUC	Events Relations	Examples of arguments
3 (1991)	Terrorist acts	`Incident_Type, Date, Location, Perpretator, Physical_Target, Human_target, Effects, Instruments`
6	Change of personal in organizations	`Post, Company, In_Person, Out_Person, Vacancy_Reason, Old_organization, New_Organization`
7 (training)	Flight accidents	`Aircraft_Model, Location, Origin, Destination, Casualties`
7 (competition)	Events linked to space vehicles and their launching	`Vehicle_Type, Vehicle_Owner, Vehicle_Manufacturer, Payload_Type, Payload_Func, Payload_Owner, payload_Target, Launch_Date, Launch_Site, Mission_Type, Mission_Function`

rali

5 Tasks at MUC-7

- Named Entity
  - name of organizations, persons, locations. Mentions of dates, times, money et percentage.
- Template Element
  - organizations, persons et vehicles mentioned in the text, relevant or not to the scenario
- Template Relation
  - domain independent relations: `location_of, employee_of, product_of`
- Coreference
  - identify all entity references (pronouns)
- Scenario Template

rali

Scenario Template Task

- Identification of event instances and relations
- Argument extraction
- Structured representation creation

rali

Examples of entities

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Persons:	Fletcher Maddox, Dr. Maddox, Oliver, Ambrose, Maddox
Organizations:	UCSD Business School, La Jolla Genomatics, La Jolla Genomatics, L.J.G.
Locations:	La Jolla, CA
Artifacts:	Geninfo, Geninfo
Dates:	June 1999

Examples of attributes

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Name	Descriptor	Category
Fletcher Maddox Maddox	Former Dean of the UCSD Business School his father the firm’s CEO	PERSON
Oliver	His son Chief Scientist	PERSON
UCSD Business School		ORGANIZATION

rali

Examples of facts

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Fletcher Maddox Fletcher Maddox Oliver Ambrose	Employee_of Employee_of Employee_of	UCSD Business School La Jolla Genomatics La Jolla Genomatics
Geninfo	Product_of	La Jolla Genomatics
La Jolla CA	Location_of Location_of	La Jolla Genomatics La Jolla Genomatics

rali

Examples of events

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Company-Formation-Event	
COMPANY	La Jolla Genomatics
PRINCIPALS	Fletcher Maddox Oliver Ambrose
DATE	
CAPITAL	

Release-Event	
COMPANY	La Jolla Genomatics
PRODUCT	Geninfo
DATE	June 1999
COST	

rali

Evaluation measures

C : corrects slots

R : filled slots

R_{Ref} : filled slots in the reference

Recall : C / R_{Ref}

Precision : C / R

F-Score : $(2 \times Precision \times Recall)/(Precision+Recall)$

rali

Results on MUC tasks

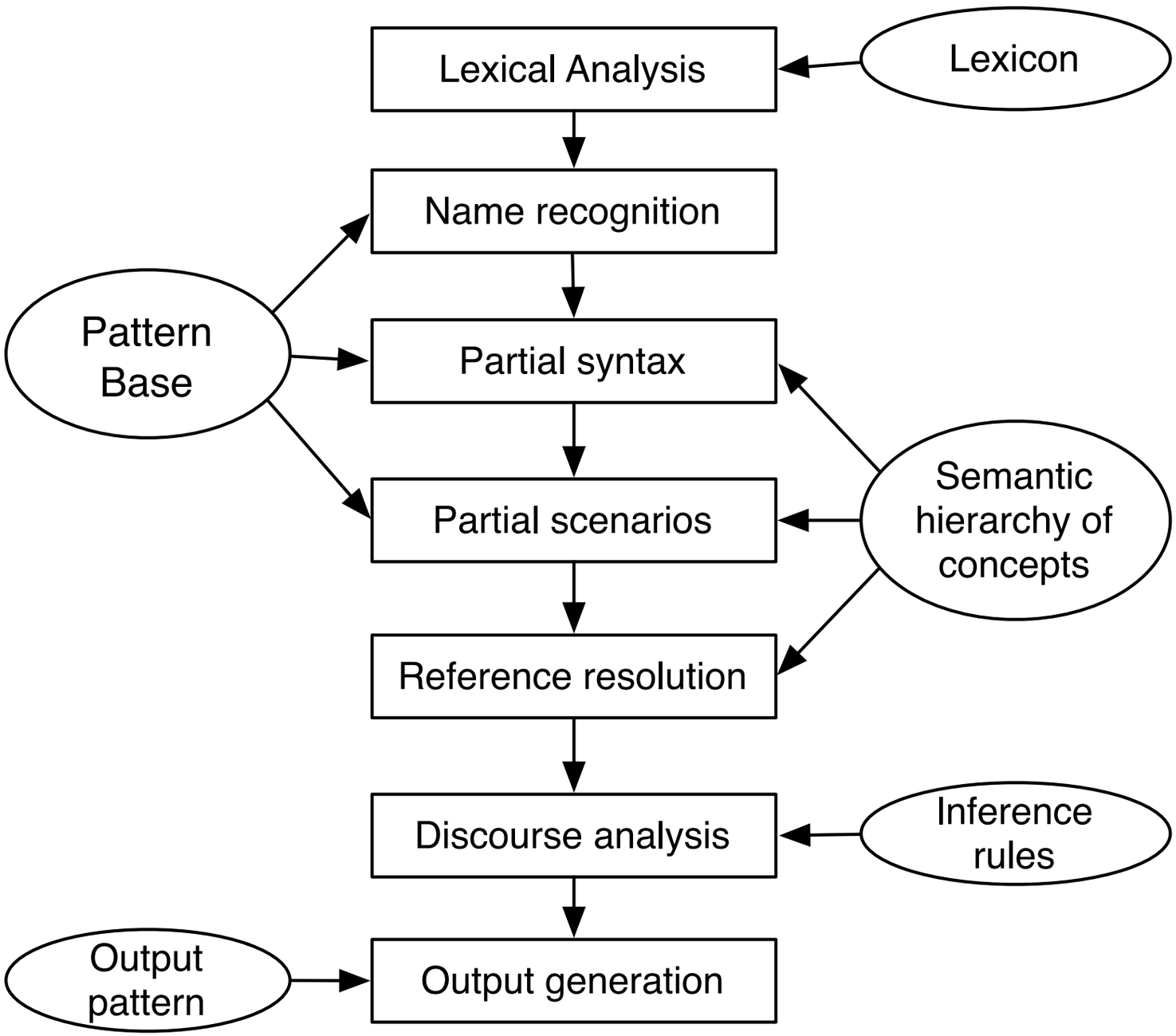
Tasks	Named Entity	Co-reference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					R < 50% P < 70%	
MUC-4					F < 56%	
MUC-5					EJV F < 53% EME F < 50%	JJV F < 64% JME F < 57%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%	
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%	

Human
60-80%

E: English - J:Japanese
JV:Joint Venture
ME: Microelectronics

rali

Typical Architecture of an information extraction system



rali

Name identification and classification

- Seldom dealt with in classical linguistic processing
- Even though it is very important to determine proper name, addresses, quantities
- Results often as XML annotations

Capt Andrew Ahab was appointed vice-president of the Great White Whale Company of Salem, Massachussets.

Capt <NAME TYPE="PERSON">Andrew Ahab</NAME> was appointed vice-president of the <NAME TYPE="ORGANIZATION">Great White Whale Company</NAME> of <NAME TYPE="LOCATION">Salem</NAME>, <NAME TYPE=" LOCATION">Massachussets</NAME>.

rali

Tagger Creation

- many regular expressions
- priority rules when many expressions match
- expression corresponding to a proper noun

WordInCaps+ `Corp.`
`Mr.` *WordInCaps*+
WordInCaps+ `,` *NumberLessThan100* `,`

- Need for list of proper names
- alias : `Fred Smith` vs `Mr. Smith`
- ambiguity: `Robert Smith Park` vs `Mr Park`

rali

Machine Learning Approaches

- Needed for portability between texts and languages
- Example for proper noun determination
  - label each noun according to
    - start, middle or end of a noun
    - start and end of a single word noun
    - nothing
  - learning with trigrams but this requires a great deal of labelled corpus
- It is often necessary to combine more than one method

rali

Use of recognized names

- Event elements extraction
- Translation of names by names
- Information retrieval based on terms that requires that words appear contiguously
- Automatic indexing of books and documents

rali

Event extraction with patterns

WordInCaps+$_1$ `appointed` *WordInCaps*+$_2$ `as president`$_3$

PERSON	2
POSITION	3
COMPANY	1
START/LEAVE JOB	`start job`

`Ford appointed Harriet Smith as president`

rali

Variations with the same pattern - 1

- Corporation names

  `Abercrombie and Fitch appointed Harriet Smith as president`
- Corporation designator

  `IBM, the famous computer manufacturer, appointed Harriet Smith as president`
- Modifier

  `IBM unexpectedly appointed Harriet Smith as president`
- Nominalization

  `IBM announced the appointment Harriet Smith as president`

Variations with the same pattern - 2

- Position names

  `IBM appointed Harriet Smith as vice president for networking`

- Conjunction

  `IBM declared a special dividend and appointed Harriet Smith as president`

- Need for inference

  `Thomas J. Watson resigned as president of IBM, and Harriet Smith succeeded him`

- Anaphoric reference

  `IBM has made a major management shuffle; the company appointed Harriet Smith as president this week`

Shallow Parsing

- Try to find certain type of constituents
- Series of simple deterministic recognizers
  - proper nouns
  - noun phrases
  - verb phrases
  - events
- this is a type of *bottom-up* parsing

rali

Shallow Parsing example

Ford Motor Company has appointed Harriet Smith, 45, as president

- Proper noun recognition

Ford Motor Company has appointed **Harriet Smith**, 45, as president
name type=org *name type=person*

- Noun phrase recognition

Ford Motor Company has appointed **Harriet Smith, 45, as president**
np head=org *np head=person* *np head=pres*

- Verb phrase recognition

Ford Motor Company has appointed Harriet Smith, 45, as president
np head=org *vg head=appoint np head=person* *np head=pres*

- Event recognition

Ford Motor Company has appointed Harriet Smith, 45, as president

PERSON	Harriet Smith
POSITION	president
COMPANY	Ford Motor Company
START/LEAVE JOB	start job

rali

Shallow Parsing
What have we achieved?

- Complex parsing methods using semantic models have given only lukewarm success
- Shallow parsing is well adapted to the simpler problem of IE
- Inter domains porting problems
- Some learning approaches have also been used with some success

rali

Some IE Applications

- Equipment failure reports
- International events news articles
- Road accident reports analysis
- Medical reports analysis

rali

References

- N. A. Chincor, Overview of MUC-7/MET-2, `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/`
- J. Cowie, W. Lehnert, Information Extraction, CACM, Volume 39 , Issue 1 (January 1996), p 80-91.
- I Eriksson, Information Extraction Walk-Through, `http://www.sics.se/~jarmo/kurser/iuiuu99/slides/IEWalkIUI99/`
- R. Grishman, Information Extraction, chap 30 of The Oxford Handbook of Computational Linguistics

rali