# Processing of large document collections

Part 10 (Information extraction: multilingual IE, IE from web, IE from semi-structured data) Helena Ahonen-Myka Spring 2006

#### **Multilingual IE**

- assume we have documents in two languages (English/French), and the user requires templates to be filled in one of the languages (English) from documents in either language
  - "Gianluigi Ferrero a assisté à la réunion annuelle de Vercom Corp à Londres."
    - "Gianluigi Ferrero attended the annual
  - meeting of Vercom Corp in London."

### Both texts should produce the same template fill:

- <meeting-event-01> :=
  - organisation: 'Vercom Corp'
  - location: 'London'
  - type: 'annual meeting'
  - present: <person-01>
- <person-01> :=
  - name: 'Gianluigi Ferrero'
  - organisation: UNCLEAR

## Multilingual IE: three ways of addressing the problem

- 1. solution
  - a full French-English machine translation (MT) system translates all the French texts to English
  - an English IE system then processes both the translated and the English texts to extract English template structures
  - in general (n languages): the solution requires a separate full IE system for each target language (here: for English) and a full MT system for each language pair

#### Multilingual IE: three ways of addressing the problem

• 2. solution

- separate IE systems process the French and English texts, producing templates in the original source language
- a 'mini' French-English MT system then translates the lexical items occurring in the French templates
- in general: the solution requires a separate full IE system for each language and a mini-MT system for each language pair

Multilingual IE: three ways of addressing the problem

3. solution

- a general IE system, with separate French and English front ends
- French and English texts are analyzed (syntax/semantics)
- a language-independent representation of the input text (discourse model) is produced
- the discourse model is a part of a domain model (ontology)
  - knowledge of the domain (entities, events,...) is described as concepts and relations
     concepts are related via mannings to lexical items in multiple
  - concepts are related via mappings to lexical items in multiple language-specific lexicons

6

# Multilingual IE: three ways of addressing the problem

- 3. solution continues...
  - the required information is extracted from the discourse model and the mappings from concepts to the English lexicon are used to produce templates with English lexical items
  - in general: the solution requires a separate syntactic/semantic analyser for each language, and the construction of mappings between the domain model and a lexicon for each language

#### IE from web

- problem setting: data is extracted from a web site and transformed into structured format (database records, XML documents)
- the resulting structured data can then be used to build new applications without having to deal with heterogenous structures
  - e.g., price comparisons
- challenges:
  - thousands of changing heterogeneous sources
     scalability: speed is important -> no complex
    - processing possible

#### IE from web

- a **wrapper** is a piece of software that can translate an HTML document into a structured form
- critical problem:
  - How to define a set of extraction rules that precisely define how to locate the information on the page?
- for any item to be extracted, one needs an extraction rule to locate both the beginning and end of the item
  - extraction rules should work for all of the pages in the source
  - both HTML markup and text content can be used
  - linguistic analysis is secondary

#### Example: country codes

<HTML><TITLE>Some Country Codes</TITLE> <BODY>

- <B>Congo</B> <I>242</I><BR>
- <B>Egypt</B> <I>20</I><BR>
- <B>Belize</B> <I>501</I><BR>
- <B>Spain</B> <I>34</I><BR>
- <HR></BODY></HTML>

Extract: {<Congo, 242>, <Egypt, 20>, <Belize, 501>, <Spain, 34>}

### Learning from examples

- input: a set of web pages, in which the data to be extracted is annotated
  - the user provides the initial set of annotated examples
  - the system can suggest additional pages for the user to annotate
- output: a set of extraction rules that describe how to locate the desired information on a web page

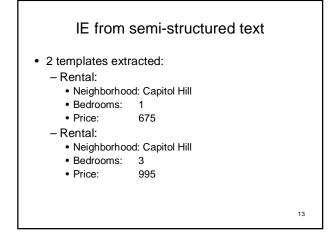
11

q

### IE from semi-structured text

Capitol Hill - 1 br twnhme. Fplc D/W W/D. Undrgrnd Pkg incl \$675. 3 BR, upper flr of turn of ctry HOME. incl gar, grt N. Hill loc \$995. (206) 999-9999 <br> <i> <font size=2> (This ad last ran on 08/03/97.) </font> </i> <hr> 8

10



#### IE from semi-structured text

- the sample text (rental ad) is not grammatical nor has a rigid structure
  - we cannot use a natural language parser as we did before
  - simple rules that might work for structured text do not work here

14

16

# Rule for neighborhood, number of bedrooms and associated price

- Pattern:: \*( Nghbr) \*( Digit) ' ' Bdrm \* '\$' ( Number)
  Output:: Rental {Neighborhood \$1} {Bedrooms \$2} {Price \$3}
  - assuming the semantic classes Nghbr (neighborhood names for the city) and Bdrm (BR, Br, Br, bedrooms,...)

### Other trends in IE

- semi- and unsupervised methods
- ACE (Automatic Content Extraction) evaluations

   extraction of general relations from text: person in a location; person has some social relation to another person, etc.
- cross-document processing
  - e.g. error correction, when the slot values for all templates are known
- backtracking in the process

   now errors on the earlier levels propagate into later levels
  - could one backtrack and correct errors made earlier, and start then again?

15