Processing of large document collections

Part 9 (Information extraction: portability)
Helena Ahonen-Myka
Spring 2005

Portability

- one of the barriers to making IE a practical technology is the cost of adapting an extraction system to a new scenario
- in general, each application of extraction will involve a different scenario
- implementing a scenario should not require too much time and not the skills of the extraction system designers

Portability

- the basic question in developing a customization tool is the form and level of the information to be obtained from the user
- goal: the customization is performed directly by the user (rather than by an expert system developer)

Portability

- if we are using a pattern matching system, most work will probably be focused on the development of the set of patterns
- also changes
  - to the dictionaries
  - to the semantic hierarchy
  - to the set of inference rules
  - to the rules for creating the output templates

Portability

- we cannot expect the user to have experience with writing patterns (regular expressions with associated actions) and familiarity with formal syntactic structure
- one possibility is to provide a graphical representation of the patterns but still too many details of the patterns are shown
- possible solution: learning from examples

Portability

- learning of patterns
  - information is obtained from examples of sentences of interest and the information to be extracted
- for instance, in a system ”AutoSlog” patterns are created semiautomatically from the templates of the training corpus

Portability

- in AutoSlog
  - given a template slot which is filled with words from the text (e.g. a name), the program would search for these words in the text and would hypothesize a pattern based on the immediate context of these words
  - the patterns are presented to a system developer, who can accept or reject the pattern

7

Portability

- the earlier MUC conferences involved large training corpora (over 1000 documents and their templates)
- however, the preparation of large, consistent training corpora is expensive
  - large corpora would not be available for most real tasks
  - users are willing to prepare a few examples (20-30?) only

8

Next time...

- we will talk about the ways to automatize the phases of the IE process, i.e. the ways to make systems more portable and faster to implement

9