Information extraction for expert systems

Good reference:
http://www.cs.helsinki.fi/u/hahonen/tet03/material/index.html
IE Mission

To create a structured form (e.g. database) from unstructured text

example:
Incident: explosion
date: 07.07.2005
Location: London
IE, IR and Full Text Understanding

- **IR**: Get relevant documents
- **IE**: Fill predefined data structure
- **FTU**: Map all text knowledge into a complex knowledge base
Typical Routines

• Lexical analysis (e.g. part-of-speech tagging)
• Name recognition (names, dates, currency…)
• Parsing (probably, partial :) )
• Coreference analysis
Syntax & Semantic

Current parsers are not perfect (see e.g. http://www.codeproject.com/csharp/englishparsing.asp)

Parsers mainly use Chomsky grammars, and the results are based on syntactic structures (debatable)

A semantic analyzer should produce a full-valued parse tree or a knowledge base
FASTUS
Lexical analysis + domain patterns + transducers

e.g.: bomb was placed by <NounGroup1> on <NounGroup2>
(has predefined descriptions of noun group, verb group, …)

regexps can be used also:

```
Subj VerbGr {NounGr / Other}* Conj VerbGr NounGr
```

Example: Salvadoran President-elect Alfredo Cristiani condemned the
terrorist killing of Attorney General Roberto Garcia Alvarado and accused
the Farabundo Marti National Liberation Front (FMLN) of the crime.

Regexp info: . | () [abc] [^abc] \d + *
FASTUS Metachars

• Part-of-speech:
  – VerbGr, NounGr, Conj, Adj, Noun, Verb…

• Part of a sentence:
  – Subj, Pred, Obj
Concept nodes in CIRCUS

Concept node **KIDNAP**
trigger words: kidnapped, abducted
variables: perpetrator, victim
constraints: perpetrator, victim $\in$ Organization or PersonalName
pattern: victim VerbGroup "by" perpetrator

(some predefined classes in constraints)

The leader of the party was kidnapped by the terrorists.
Semi-structured Data

- Much easier to find patterns
- Still very important

Example 1 – rental ad:

C-Hill, 1 br, W/F, W/P, $675
-> (district, bedrooms, furniture, price)

Example 2 – international country codes:

<table><tr><td>Afghanistan</td><td>93</td></tr>
<tr><td>Albania</td><td>355</td></tr>
<tr><td>Algeria</td><td>213</td></tr></table>

*Regexp can be used*
Self-Learning Algorithms

Simple techniques like pattern-aided recognition + human-supervised learning on predefined text corpus

(used for automatic classes extraction for specialized texts)
random graphs are used to simulate user’s behavior
Exercises

1) Write a regexp to extract information about paper submission:

2) Make a good FASTUS template to extract terrorist data (use http://www.cs.helsinki.fi/u/hahonen/tet03/material/sample1.txt text)

3) Make a CIRCUS concept node for recipes
http://www.online-cookbook.com/
Find a recipe with the components of your choice (maybe without some of them)