Multilingual databases for named entity recognition

Thierry Poibeau
Institut National des Langues et Civilisations Orientales,
Paris
The need for multilingual systems

- More and more systems are facing multilingualism
  - Knowledge management (KM) in multinational organizations
  - Information retrieval (IR) and question answering (QA) from the Internet

- Need of multilingual linguistic tools
  - Named entity is a good starting point (useful for KM, IR, QA...)

Langues’O
The multilingual named entity framework
System overview

- More than 10 languages implemented or under implementation
  - Arabic, Chinese, English, French, German, Japanese, Finnish, Malagasy, Persian, Polish, Russian, Spanish and Swedish

- A large disparity between these languages
  - Resources already available or not
  - Corpora available or not
  - Original development strategies (resource sharing, rapid grammar derivation, etc.)
General architecture

Rule-based Systems

(1) Lexical analysis

(2) Grammar application

dictionary

(3) Dynamic acquisition from the text

grammar

(4) Revision mechanisms

Annotated Text
Implementation issues

- Preference for *de facto* standards:
  - Unicode (Unicode Little-Endian)
  - XML (XSL and CSS for visualisation)
  - Java (multi-platform)

- Development based on finite-state transducers
  - Intex/Unitex toolbox for finite-state technology management
  - Except for Japanese (Chasen morphological analyser + Unicode perl scripts)
The Japanese system
System components

- Inspired form previous NE recognizers developed for Japanese (cf. IREX)
- Morpho-syntactic analysis
  - Chasen system (Asahara and Matsumoto)
  - JIS format
- Named entity recognition
  - Perl scripts
  - Unicode format
  - Same DTD as for other languages
Japanese system architecture

Text

Chasen morpho-syntactic analyzer

Analyzed text

Perl scripts

Annotated text

Resources

Text (JIS format)

Chasen morpho-syntactic analyzer (JIS format)

Analyzed text (UNICODE format)

Perl scripts (UNICODE format)

Annotated text (UNICODE format)
Named entity recognition in Japanese

- Linguistic knowledge is defined by:
  - Trigger words
  - Dictionaries of Named Entities
  - Grammars to group NE elements together

- This analysis is done on top of the analysis from Chasen
# Dictionaries

<table>
<thead>
<tr>
<th>Language</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ait-Hamadouche</td>
<td>Jayalath</td>
</tr>
<tr>
<td>Ryo</td>
<td>Yoshiko</td>
</tr>
</tbody>
</table>

International Herald Tribune, 

- ATTAC
- International Herald Tribune,
Morpho syntactic analysis (Chasen)
An annotated text
The XML resulting file
Evaluation and conclusion
Evaluation and applications

- French and English systems achieved state-of-the-art performances
- Other systems still need to be improved

Applications:
- Some modules are already used in Information Extraction systems (French, English...)
- Deployment of other modules for scientific and technological intelligence, etc.
Japanese module

- Relatively good coverage given that the analysis is made on top of Chasen
- Dictionaries must be extended
  - Analysis of new corpora
  - IREX resources?
- Integration of machine learning techniques
- Technical problems (JIS format, need Japanese Windows platforms)
Conclusion and future work

- Basic systems for more than 10 languages
  - Good coverage of newspaper corpora
  - Easy adaptation to specialized corpora (technical and scientific papers)

- Perspective: Integration to Papillon?
Further information

- See the website:
  http://crimlangueso.asso.fr/realisations/entites/frameEN.html

- Reports are available on the website

- Resources are available for research purposes