1. Introduction

HeiNER - Heidelberg Named Entity Resource

Multilingual resource for Named Entity Disambiguation, Translation and Transliteration, which is freely available at http://heiner.cl.uni-heidelberg.de

Contains:
- 1.547,586 disambiguated English NEs
- Translations into 253 languages
- Context Sets in 16 languages
- A dictionary that maps ambiguous proper names to sets of unique and disambiguated NEs

The resource is automatically constructed from Wikipedia by exploiting its internal link structure.

Method

- Based on Wikipedia
- Extraction of disambiguated NEs
- Translation of monolingual NE seeds into all languages available in Wikipedia
- Context Extraction in all languages

Advantages

- Avoids manual annotation
- Viable for resource-poor languages
- Yields large NE resource for multiple languages

Current State of the project

- Capitalisation-based heuristic for NER in English as source language, with a precision of 0.95
- NE translation to 15 target languages

Usage

HeiNER provides data for supervised training in:
- Multilingual Information Extraction
- Machine Translation
- NER and NE Disambiguation
- Text summarisation

2. Named Entity Acquisition

Monolingual

Only article titles are considered as candidates, which circumvents the need for NE boundary detection.

Heuristics:

- Based on (Bunescu and Pasca, 2006)
- Three steps in recognition:
  1. Multiword title “United_Nations”
  2. CamelCase “YouTube”
  3. Capitalisation in text

Capitalisation:

- Candidate is frequently capitalised (> 75%)
- Occurrences in sentence initial position are not counted

Disambiguation:

- Extraction of different - disambiguated - surface forms, by exploitation of Disambiguation and Redirect Pages in Wikipedia
- Disambiguation Pages distinguish between different NE readings and Redirect Pages unify different surface forms in a single representation
- Mappings between surface and disambiguated forms are stored in the Disambiguation Dictionary

Multilingual

Translation/Transliteration

- Utilisation of cross-language links in Wikipedia
- Translation of NEs from a single source language into all other languages in Wikipedia
- Triangulation between languages to increase coverage, by exploiting the fact that links should obey the principle of translitility
- All extracted translations constitute the Translation Dictionary

3. Context Acquisition

Context Set

<table>
<thead>
<tr>
<th>Context Set</th>
<th>NUMBER OF CONTEXTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSOLUTE</td>
<td>AMBIG. NE</td>
</tr>
<tr>
<td>en</td>
<td>43,165,448</td>
</tr>
<tr>
<td>fr</td>
<td>7,621,202</td>
</tr>
<tr>
<td>ja</td>
<td>6,321,906</td>
</tr>
<tr>
<td>zh</td>
<td>17,214,399</td>
</tr>
<tr>
<td>ru</td>
<td>1,945,096</td>
</tr>
<tr>
<td>pt</td>
<td>1,322,408</td>
</tr>
<tr>
<td>de</td>
<td>2,632,568</td>
</tr>
<tr>
<td>es</td>
<td>1,213,563</td>
</tr>
<tr>
<td>it</td>
<td>1,806,103</td>
</tr>
<tr>
<td>nl</td>
<td>625,572</td>
</tr>
<tr>
<td>MEAN</td>
<td>6,063,035</td>
</tr>
</tbody>
</table>

Table 1: Number of contexts extracted for different languages

Disambiguation Dictionary

<table>
<thead>
<tr>
<th>Languages</th>
<th>NUMBER OF CONTEXTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSOLUTE</td>
<td>AMBIG. NE</td>
</tr>
<tr>
<td>en</td>
<td>43,165,448</td>
</tr>
<tr>
<td>fr</td>
<td>7,621,202</td>
</tr>
<tr>
<td>ja</td>
<td>6,321,906</td>
</tr>
<tr>
<td>zh</td>
<td>17,214,399</td>
</tr>
<tr>
<td>ru</td>
<td>1,945,096</td>
</tr>
<tr>
<td>pt</td>
<td>1,322,408</td>
</tr>
<tr>
<td>de</td>
<td>2,632,568</td>
</tr>
<tr>
<td>es</td>
<td>1,213,563</td>
</tr>
<tr>
<td>it</td>
<td>1,806,103</td>
</tr>
<tr>
<td>nl</td>
<td>625,572</td>
</tr>
<tr>
<td>MEAN</td>
<td>6,063,035</td>
</tr>
</tbody>
</table>

Table 2: Increase in coverage by means of triangulation

Context Distribution

- Contexts are unevenly distributed
- 1.5% of NEs in English cover 44% of the contexts
- 23k NEs with more than 200 contexts
- Four NEs with 100,000+ contexts:
  1. United States
  2. England
  3. United Kingdom
  4. Germany

4. Results and Evaluation

Evaluation:

- 2 evaluation sets with 2000 Markables with two and three annotators respectively
- Fleiss Kappa Agreement of 0.774 and 0.771
- Average NER Precision of 0.95

Triangulation:

- Average boost in coverage of 4.47%
- Maximum boost in Turkish with 7.73%
- 77,694 additional NE translations acquired

5. Further Work

- Providing the data as database dumps instead of XML
- Semantic Class labels for all NEs
- NER with proper NER systems in several languages
- Topic signatures for all NEs
- Context alignment