Named Entity WordNet

*Lantonio Toral
^Rafael Muñoz
*Monica Monachini

*Istituto di Linguistica Computazionale (Pisa, Italy)
^University of Alicante (Spain)

LREC 2008
O12 - Named Entity Recognition

Marrakech, 2008-05-28
Outline

Intro
- Named Entities (NEs)
- Language Resources (LRs)
- Why NEs in LRs?
- How to enrich LRs with NEs?

Named Entity WordNet
- Mapping & Disambiguation
- Article extraction
- NE identification
- NE repository

Conclusions & Future
Usually refer to

- Proper nouns: names of people, locations, organizations, ...
- Numerical expressions: time, amounts, ...

Important for NLP tasks

- NEs: 10% of text + carry important semantic info

Different sets of NE categories

- ConLL -> flat, 4 types (per, org, loc, misc)
- Sekine -> hierarchy, +100 subtypes
Manually created by expert lexicographers
Broad-coverage resources
  - Common nouns, adjectives, verbs, adverbs
Rich Semantic Info (relations, roles, etc)
WordNet
  - +100k word senses
Manually created by expert lexicographers
Broad-coverage resources
  - Common nouns, adjectives, verbs, adverbs
Rich Semantic Info (relations, roles, etc)
WordNet
  - +100k word senses
LRs lack info about NEs
  - “building a proper noun ontology is more difficult than building a common noun ontology as the set of proper nouns grows more rapidly” (Mann, 2002)
Stored Knowledge can be applied to NLP tasks
E.g. Question Answering
  Question (CLEF 2006)
    Who is Vigdis Finnbogadottir?
  QA system
    Linguistic analysis of text [S. Ferrandez et al. 06]
      “[... presidents: Vigdis Finnbogadottir (Iceland), [...]”
    Solution (wrong): Iceland
Why NEs in LRs?

- Stored Knowledge can be applied to NLP tasks
- E.g. Question Answering
  - Question (CLEF 2006)
    - Who is Vigdis Finnbogadottir?
  - QA system
    - Linguistic analysis of text [S. Ferrandez et al. 06]
      - “[...] presidents: Vigdis Finnbogadottir (Iceland), [...]”
    - Solution (wrong): Iceland
  - Possible related knowledge in LR
    - “Vigdis Finnbogadottir” instance_of: “president”, “icelandic”, “female head of state”
  - LR can be useful within QA, for example to:
    - Find answers
    - Validate answers
How to enrich LRs with NEs?

- NEs should be acquired & introduced automatically
- **Ideal Source**
  - Up-to-date
  - High Coverage
  - Allow a Good Quality Extraction
How to enrich LRs with NEs?

NEs should be acquired & introduced automatically

Ideal Source
- Up-to-date
- High Coverage
- Allow a Good Quality Extraction

Wikipedia
- Dynamic source
- Huge amount of NEs
- Some degree of structure
Automatically Extend WordNet with NEs extracted from Wikipedia
Mapping

Map lemmas
- WordNet: noun classes (instantiated)
- Wikipedia: categories

Results

<table>
<thead>
<tr>
<th>Synsets</th>
<th>Total</th>
<th>200704</th>
<th>200711</th>
<th>200801</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>893</td>
<td>513</td>
<td>536</td>
<td>541</td>
</tr>
<tr>
<td>Mapped</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>57.44%</td>
<td>60.02%</td>
<td>60.58%</td>
<td></td>
</tr>
</tbody>
</table>

Analysis (non mapped)
- 75% no matching category but matching article
- 13% no matching category nor matching article
- 10% matching category but PoS error
Disambiguation

WordNet polysemous nouns to Wikipedia categories

Intersection of instances

WN obelisk

Obelisk1: stone pillar
Obelisk2: character used in printing

Mapping

WK Obelisks
Disambiguation

WordNet polysemous nouns to Wikipedia categories

- Intersection of instances

WN obelisk

- Obelisk1: stone pillar
- Obelisk2: character used in printing

Mapping

WK Obelisks

Washington Monument

has_instance
Disambiguation

WordNet polysemous nouns to Wikipedia categories

Intersection of instances

WN obelisk

Obelisk1: stone pillar

Obelisk2: character used in printing

Washington Monument

Mapping

WK Obelisks

contains

Washington Monument
Disambiguation

WordNet polysemous nouns to Wikipedia categories

- Intersection of instances

Results (262 words): 100% precision, 39% recall

Analysis non disambiguated words:
- 78% no common instance found
- 22% no sense corresponds to category
For each category mapped (and its hyponyms*) fetch:

- Titles
- Abstracts
- Variants

*Hyponym identification (subcategories)

- ^ category (" by " | " of " | " in " | " stubs$")
  - Obelisks in Argentina
- ^ (JJ|JJR|NN|NP)+ (CC(JJ|JJR|NN|NP)+)* " " category$
  - Ancient obelisks
An extracted article might be a NE or a common noun

Look for occurrences of its title in its body text & check capitalisation (Bunescu & Pasca 2006)

Not only in the English Wikipedia, but in 10 Wikipedias for langs that follow these caps. norms

Text size to look for occurrences bigger -> results more representative

Language independent -> whatever the language we obtain the article equivalent in these languages
An extracted article might be a NE or a common noun
- Look for occurrences of its title in its body text & check capitalisation (Bunescu & Pasca 2006)
- Not only in the English Wikipedia, but in 10 Wikipedias for langs that follow these caps. norms
  - Text size to look for occurrences bigger -> results more representative
  - Language independent -> whatever the language we obtain the article equivalent in these languages

Results
- Only English -> F 78.06%, P 73.91%, R 87.93%
- 10 languages -> F 82.26%, P 79.69%, R 87.93%
## Extracted NEs

### General
- 310,742 Nes, 452,017 variants, 381,043 instance rels

### Detailed (per lexicographic file)

<table>
<thead>
<tr>
<th>Lex File</th>
<th>Nes</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>act</td>
<td>4,214</td>
<td>Project_Pluto instanceOf project0_4</td>
</tr>
<tr>
<td>artifact</td>
<td>23,878</td>
<td>Akinada_Bridge instanceOf suspension_bridge0_6</td>
</tr>
<tr>
<td>communication</td>
<td>1,973</td>
<td>Flower_of_Scotland instanceOf national_antherm0_10</td>
</tr>
<tr>
<td>event</td>
<td>58</td>
<td>Sino-Soviet_split instanceOf schism0_11</td>
</tr>
<tr>
<td>group</td>
<td>1,216</td>
<td>Medici instanceOf family0_14</td>
</tr>
<tr>
<td>location</td>
<td>43,582</td>
<td>Incense_Route instanceOf trade_route0_15</td>
</tr>
<tr>
<td>object</td>
<td>28,180</td>
<td>Pyxis instanceOf constellation=17</td>
</tr>
<tr>
<td>person</td>
<td>277,941</td>
<td>Vladimir_Kotelnikov instanceOf electrical_engineer0_18</td>
</tr>
</tbody>
</table>
Elements: NEs, classes, relations, variants, definitions
LMF compliant: ISO standard for lexicons
  Independent from specific LRs
Web test & download
  dlsi.ua.es/~atoral/#Resources
  www2.ilc.cnr.it/ne-repository
## Named Entity WordNet demo

Introduce Named Entity [ ] get info

### LexicalEntry

<table>
<thead>
<tr>
<th>le id</th>
<th>PoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>le_Tim_Robbins</td>
<td>PN</td>
</tr>
</tbody>
</table>

### FormRepresentation

<table>
<thead>
<tr>
<th>written form</th>
<th>variant type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timothy_Francis_Robbins</td>
<td>alias</td>
</tr>
<tr>
<td>Timothy_Robbins</td>
<td>alias</td>
</tr>
<tr>
<td>Tim_Robbins</td>
<td>full</td>
</tr>
</tbody>
</table>

### Sense

<table>
<thead>
<tr>
<th>sense id</th>
<th>resource</th>
<th>id in resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_Tim_Robbins</td>
<td>Wikipedia</td>
<td>269416</td>
</tr>
</tbody>
</table>

### SenseRelation

<table>
<thead>
<tr>
<th>source sense id</th>
<th>relation type</th>
<th>target sense id</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_Tim_Robbins</td>
<td>instanceof</td>
<td>s_film_director0_18</td>
</tr>
<tr>
<td>s_Tim_Robbins</td>
<td>instanceof</td>
<td>s_screenwriter0_18</td>
</tr>
</tbody>
</table>
Conclusions & Future

High Quality & Large NE extension of WordNet
- +310k Nes (it had 7k), +380k relations
- Standard-compliant output

Future
- Apply to other LRs for different languages
  - Empirically demonstrate generality of the approach
  - Derive a Multilingual NE repository
- Exploit Textual Entailment to disambiguate mapping
Thanks for your attention!

Questions?