Towards a Named Entity WordNet

Abstract

- automatically add named entities to WordNet extracted from Wikipedia
- motivate that knowledge acquisition can benefit from exploiting wiki texts due to some characteristics of these resources that provide advantages over other commonly used resources such as corpora or MRDs
- the extraction approach enriches the English WordNet with more than 55,000 new instances (i.e. more than seven times the amount of instances that WordNet contains) with a precision over 93%
- a valuable resource for NLP tasks and especially for Named Entity Recognition

NewText

New forms of communication: NewText
Among them wikis: collaborative editing
- Corpora: relations tend to be subjective judgements, same instance in different manners
e.g. Bill Clinton and William Clinton, acquired twice?
- MRDs: limited size

WordNet

- nouns, verbs, adjectives and adverbs organised into sets of synonyms (synsets)
- several semantic relations among nodes
- nouns: classes (common) and instances (proper)
  - high coverage of classes but low of instances
  - “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)

Mapping of WordNet synsets to Wikipedia categories
A synset is associated to a category if the lemma of any of its words is identical to the lemma of the category

Synset ‘screenwriter, film_writer’ linked to category ‘screenwriters’

Evaluation (test set)

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Precision</th>
<th>Recall</th>
<th>F1β=0.5</th>
<th>F1β=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>90.90</td>
<td>66.97</td>
<td>78.43</td>
<td>82.19</td>
</tr>
<tr>
<td>0.89</td>
<td>90.90</td>
<td>68.97</td>
<td>78.43</td>
<td>82.19</td>
</tr>
<tr>
<td>0.91</td>
<td>93.02</td>
<td>68.97</td>
<td>79.20</td>
<td>83.33</td>
</tr>
<tr>
<td>0.93</td>
<td>92.86</td>
<td>67.24</td>
<td>78.00</td>
<td>82.39</td>
</tr>
</tbody>
</table>

Tim Robbins begins by caps 96.67% (>91%)