Weak Named Entities Recognition using morphology and syntax

Antonio Toral, Rafael Muñoz and Andres Montoyo
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{atoral, rafael, montoyo}@dlsi.ua.es

Abstract

A lot of research has been done regarding strong named entities recognition, both following the rule-based and the learning approaches. However, weak named entities have been not treated in detail yet. We propose a system for the detection and classification of these kind of entities that uses morphology (PoS) and syntax (shallow parsing) features. Our starting point is the text correctly tagged with strong entities. The results are encouraging as we obtain a F-score better than 60% without learning (corpus) nor knowledge (gazetteers or grammars) resources.

Introduction

NER is the task consisting of detecting and classifying entities that refer to people, locations, organizations, time, etc.

Types of entities:
- **Strong:**
  - proper names
  - accuracy achieved: 90% detection, 80% classification
  - example: "Portugal"

- **Weak:**
  - syntagmas (trigger word + proper name)
  - not researched into detail yet
  - example: "el presidente de Portugal"

Motivation: morphologic and syntactic features can help to detect and classify weak entities
- Morphology: trigger words are common nouns
- Syntax: weak entities take the form of a noun phrase

Procedure

**Input**
- words + strong entities

**PoS tagger**
- morphologic analysis (gets lemma + PoS)

**Shallow Parser**
- syntactic analysis (gets type of syntagma + number of syntagma in sentence)

**Weak NER**
- detects and classifies weak entities
  a. recognizes weak entities
  b. applies heuristics to smooth entities boundaries:
    - takes out prepositions (p)
    - joins syntagmas (s)

**Output**
- words + strong entities + weak entities

Example

<table>
<thead>
<tr>
<th>Text + Strong entities</th>
<th>PoS tagger</th>
<th>Shallow Parser</th>
<th>Weak NER</th>
<th>Text + Weak entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>el el DAOMS0 sn 4 O B-PER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alcalde alcalde NCMS000 sn 4 O I-PER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de de SP500 sn 4 O I-PER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arévalo Arévalo NP00000 sn 4 B-LOC I-PER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Francisco Francisco NP00000 sn 6 B-PER O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>León León NP00000 sn 6 I-PER O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>42.00</td>
<td>33.87</td>
<td>37.50</td>
</tr>
<tr>
<td>NER+p</td>
<td>66.00</td>
<td>53.23</td>
<td>58.93</td>
</tr>
<tr>
<td>NER+s</td>
<td>44.83</td>
<td>41.94</td>
<td>43.33</td>
</tr>
<tr>
<td>NER+p+s</td>
<td>65.52</td>
<td>61.29</td>
<td>63.33</td>
</tr>
</tbody>
</table>

Conclusions

We have developed a novel approach to treat weak entities
We achieve acceptable results without linguistic resources
We have demonstrated that there are morphologic and syntactic features that help the treatment of weak entities

Future work

Apply a customized PoS tagger (need to distinguish between only 3 categories)
Take into account deeper levels of syntactic analysis
Build a NER system to treat both strong and weak entities