An Experimental Study of a Hybrid Entity Recognition and Linking System

Julien Plu, Giuseppe Rizzo, Raphaël Troncy
EURECOM, Sophia Antipolis, France
(plu, rizzo, troncy)@eurecom.fr

Problem
Numerous entity linking systems, that are generally tailored and optimized for specific data, scenario and applications
This calls for a self-configurable adaptive system:
• Enable to process heterogeneous documents that vary:
  • in kind (e.g. tweet, news article, blog post, subtitles)
  • or language (e.g. English, French, Spanish, Chinese)
• Enable to personalize the information to extract and disambiguate:
  • common types (e.g. PER, LOC, ORG) versus specific types (Roles)
  • types interpretation (e.g. Facebook can be a Product or an Organization)
  • knowledge base (e.g. DBpedia, Wikidata, Geonames)

Proposed Architecture

Entity Extractor

Wikipedia
Edward Joseph "Ed" Snowden (born June 21, 1983) is an American computer professional, former CIA employee, and former government contractor who leaked classified information from the U.S. National Security Agency (NSA) in 2013.

Twitter
RT: So my favorite part of #Citizenfour is when the @selenagomez video clip "Come & Get It" plays on #edsnowden hotel TV :D:D

Text Normalization
Remove emoticons, other special symbols, the RT symbol for tweets

Overlap Resolution
Date: June 21, 1983
Number extractor: 21 and 1983

Candidate Filtering
Use link references in Wikipedia
NSA: db:National_Security_Agency, db:FC_NSA_Sofia
Edward Joseph "Ed" Snowden: db:Edward_Snowden

Scoring
Each candidate is scored using a linear formula. L for Levenshtein distance, R for set of redirects, D for set of disambiguation and PR for PageRank. a, b and c are weights set with a > b > c and a + b + c = 1
score(db:National_Security_Agency) = (a * L("NSA", "National Security Agency") + b * ("NSA" in Redirects) + c * "NSA" in Disambiguation) * PR(db:National_Security_Agency)

Evaluation

Datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Co-reference</th>
<th>NIL entities</th>
<th>Type</th>
<th>Date</th>
<th>Number</th>
<th>Kind of text</th>
<th>Nb of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Micropost2014</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>tweet</td>
<td>1165</td>
</tr>
<tr>
<td>#Micropost2015</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>tweet</td>
<td>2027</td>
</tr>
<tr>
<td>OKE2015</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Article</td>
<td>101</td>
</tr>
</tbody>
</table>

Task                  | Precision | Recall | F-measure |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>68.4</td>
<td>75.2</td>
<td>71.6</td>
</tr>
<tr>
<td>Linking</td>
<td>48.8</td>
<td>47.1</td>
<td>47.9</td>
</tr>
</tbody>
</table>

1. Similar results regardless of the kind of text
2. Performance at extraction stage similar to top state-of-the-art systems (or slightly better)
3. Big drop of performance at linking stage mainly due to an unsupervised approach

Values for each dataset

<table>
<thead>
<tr>
<th>Task</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>69.2</td>
<td>72.5</td>
<td>70.8</td>
</tr>
<tr>
<td>Linking</td>
<td>47.4</td>
<td>45.2</td>
<td>46.3</td>
</tr>
</tbody>
</table>

#Micropost2014  #Micropost2015 OKE2015

https://www.3cixty.com/