Entity Linking with Dexter

2nd HPC Lab. Workshop

Salvatore Trani
Diego Ceccarelli, Claudio Lucchese,
Salvatore Orlando, Raffaele Perego
The goal is to identify relevant **entities** mentioned by **fragments** of text.

- Entities are taken from a given catalogue, e.g. Wikipedia.
Entity Linking

- State-of-the-art approaches run three steps:

1. **Spotting**
   - Given a document, *find fragments of text* potentially referring to entities, a.k.a. *spots*
   - Common approach is to match anchors in Wikipedia
   - Some spots are ambiguous, e.g. “Michael Collins”

2. **Disambiguation**
   - Given a set of spots in a document, *find the correct entity* for each spot.
   - Steps 1 and 2 are sometimes referred to as *Word Sense Disambiguation*

3. **Filtering**
   - Given a document, its spots and their entities, decide *which links to use*:
     - E.g., “…[Bank of London]…” vs “…[the] [Bank of London] …”

- The main *research questions* is:
  - How to improve the *disambiguation* step
Disambiguation

- Disambiguation is achieved by looking at the relatedness among candidate senses in the whole document.

- Relatedness is measured with:

\[ \rho_{MW}^{MW}(a,b) = 1 - \frac{\log(\max(|in(a)|,|in(b)|)) - \log(|in(a) \cap in(b)|)}{\log(|W|) - \log(\min(|in(a)|,|in(b)|))} \]

- inspired to the so-called Normalized Google Distance between words
- nice, it may be applied to any pair of Wikipedia entities

- It is adopted as a building block in all state-of-the-art algorithms
Entity Relatedness

- But it does not always work
  - Example: a) Andronicus of Rhodes b) Chondrichthyes c) Aristotle
  - $\rho^{MW}(a,c) = 0.56$  
    (Andronicus of Rhodes is credited with the production of the first reliable edition of Aristotle's works)
  - $\rho^{MW}(a,b) = 0.54$  
    (‘[…] He separated the aquatic mammals from fish, and knew that sharks and rays were part of the group he called Selachē (selachians)’)
  - Andronicus and Aristotle are related as much as Andronicus and a fish!

- Our research question is:
  - What is the best relatedness function?

- We cast it into a learning to rank problem..
Relatedness as a Ranking Function

The documents are being ranked by their relatedness.

Apollo 11

Michael Collins

The query
Relatedness as a Ranking Function

from good relatedness $\rho$ function we expect this

Apollo 11

The query $\varepsilon_h$

Spot1 Spot2 Spot3 Spot4

1 2 3 4

5 6 7 8

9 10 11 12 13

$\pi^h\rho$
 Entity Relatedness Learning

- The **quality** of a ranked list is measured as:

\[
\text{DCG}(\pi^h) = \sum_{j=1}^{\left|\pi^h\right|} \frac{l_j}{\log(j + 1)}
\]

  where \(l_j\) is a binary label of the \(j\)-th ranked entity identifying the correct entities.

- The **Entity Relatedness Learning Problem** requires to find the **ranking function** \(\rho\) that **maximizes the ranking quality**:

\[
\frac{1}{|D|} \sum_{D \in \mathcal{D}} \frac{1}{|S_D|} \sum_{s_h \in S_D} \text{NDCG}(\pi^h)
\]

\text{Documents in the collection} \quad \text{Spots in document} D

\text{Normalized DCG}
Learning to Rank is not easy

- **Warning!**
  Optimizing NDCG is not easy.

- One simple reason is that NDCG implies sorting, which is not a nicely derivable function
- Therefore we cannot apply gradient descent or similar…

- State-of-the-art approaches:
  - Optimize number of correctly ordered pairs
  - Optimize RMSE on labels (in our case binary labels)
    - See Gradient-Boosted Regression Trees
  - Approach to NDCG with some heuristic
    - See $\lambda$-MART
Benchmark Dataset

• Required for the evaluation and for the learning
  • we minimize the empirical risk on the training set.

• CoNLL 2003
  • 1,494 annotated news stories of the Reuters Corpus V1.
  • For each spot in a news we extract a tuple containing
    • A *reference entity* $\varepsilon_h$ (the query)
    • The *other relevant entities* in the document
    • *Candidate but not relevant entities* in the document
    • With a maximum distance constraint of 50 words

• We extracted a benchmark dataset with
  • 17,040 tuples/queries each with ~97 results
  • a total of 1,649,841 query-result pairs
• Machine-learning approaches allow to easily test a wide set of features

### Singleton Features

| P(a) | probability of a mention to entity a: $P(a) = \frac{\|\text{in}(a)\|}{|W|}$. |
|------|---------------------------------------------------------------------|
| H(a) | entropy of a: $H(a) = -P(a) \log(P(a)) - (1-P(a)) \log(1-P(a))$. |

### Asymmetric Features

| P(a|b) | conditional probability of the entity a given b: $P(a|b) = \frac{|\text{in}(a) \cap \text{in}(b)|}{|\text{in}(b)|}$. |
|------|---------------------------------------------------------------------|
| Link(a→b) | equals 1 if a links to b, and 0 otherwise. |
| P(a→b) | probability that a links to b: equals $1/|\text{out}(a)|$ if a links to b, and 0 otherwise. |
| Friend(a, b) | equals 1 if a links to b, and $|\text{out}(a) \cap \text{in}(b)|/|\text{out}(a)|$ otherwise. |
| KL(a||b) | Kullback-Leibler divergence: $KL(a||b) = \log \frac{P(a)}{P(b)} P(a) + \log \frac{1-P(a)}{1-P(b)} (1 - P(a))$.
### Symmetric Features

<table>
<thead>
<tr>
<th>$\rho^{MW}_{out}(a,b)$</th>
<th>$\rho^{MW}_{in-out}(a,b)$</th>
<th>$J_{out}(a,b)$</th>
<th>$J_{in-out}(a,b)$</th>
<th>$\chi^2(a,b)$</th>
<th>$\chi^2_{out}(a,b)$</th>
<th>$\chi^2_{in-out}(a,b)$</th>
<th>PMI$(a,b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J(a,b)$</td>
<td>Jaccard similarity: $J(a,b) = \frac{\text{in}(a) \cap \text{in}(b)}{\text{in}(a) \cup \text{in}(b)}$.</td>
<td>$\rho^{MW}$ considering outgoing links.</td>
<td>$\rho^{MW}$ considering the union of the incoming and outgoing links.</td>
<td>$\chi^2$ statistic: $\chi^2(a,b) = (</td>
<td>\text{in}(b) \cap \text{in}(a)</td>
<td>\cdot (</td>
<td>W</td>
</tr>
</tbody>
</table>
Performance of the Relatedness function

<table>
<thead>
<tr>
<th></th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{MW}$</td>
<td>0.59</td>
<td>0.63</td>
<td>0.62</td>
<td>0.42</td>
<td>0.31</td>
<td>0.72</td>
</tr>
<tr>
<td>$\rho_{\lambda\text{MART}}$</td>
<td>0.75</td>
<td>0.79</td>
<td>0.80</td>
<td>0.51</td>
<td>0.36</td>
<td>0.87</td>
</tr>
<tr>
<td>$\rho_{\text{GBRT}}$</td>
<td>0.75</td>
<td>0.78</td>
<td>0.80</td>
<td>0.51</td>
<td>0.35</td>
<td>0.86</td>
</tr>
</tbody>
</table>

**Example:** a) Andronicus of Rhodes b) Chondrichthyes c) Aristotle

- $\rho_{\lambda\text{MART}}(a,c) = 0.66$ (Andronicus of Rhodes is credited with the production of the first reliable edition of Aristotle's works)
- $\rho_{\lambda\text{MART}}(a,b) = 0.0015$ (‘[…] He separated the aquatic mammals from fish, and knew that sharks and rays were part of the group he called Selachē (selachians)’)

**Andronicus very related to Aristotle but not to the fish!**
Impact on Entity Linking Algorithms

WIKIMINER
Learning to link with wikipedia,
Milne and Witten, CIKM2008

TAGME
TAGME: on-the-fly annotation of short text
fragments (by wikipedia entities)
Ferragina and Scaiella CIKM2010

REFERENT GRAPH
Collective entity linking in web text:
a graph-based method
X. Han, L. Sun, and J. Zhao. SIGIR2011
Best results were with WIKIMINER

<table>
<thead>
<tr>
<th></th>
<th>$\rho_{MW}$</th>
<th>$\rho_{\lambda\text{MART}}$</th>
<th>$\rho_{GBRT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P@1$</td>
<td>0.78</td>
<td><strong>0.86</strong> +10%</td>
<td>0.83 +6%</td>
</tr>
<tr>
<td>$P@5$</td>
<td>0.64</td>
<td>0.68 +6%</td>
<td>0.69 +8%</td>
</tr>
<tr>
<td>$P@10$</td>
<td>0.50</td>
<td>0.51 +2%</td>
<td>0.53 +6%</td>
</tr>
<tr>
<td>$iP_{r=0.10}$</td>
<td>0.88</td>
<td><strong>0.92</strong> +5%</td>
<td>0.91 +3%</td>
</tr>
<tr>
<td>$iP_{r=0.50}$</td>
<td>0.66</td>
<td>0.73 +11%</td>
<td>0.77 +17%</td>
</tr>
<tr>
<td>NDCG</td>
<td>0.66</td>
<td>0.72 +9%</td>
<td>0.75 +14%</td>
</tr>
<tr>
<td>MRR</td>
<td>0.87</td>
<td><strong>0.92</strong> +6%</td>
<td>0.90 +3%</td>
</tr>
<tr>
<td>NDCG@5</td>
<td>0.71</td>
<td>0.76 +7%</td>
<td>0.77 +8%</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.69</td>
<td>0.73 +6%</td>
<td>0.75 +9%</td>
</tr>
<tr>
<td>Recall</td>
<td>0.64</td>
<td>0.70 +9%</td>
<td>0.75 +17%</td>
</tr>
<tr>
<td>Rprec</td>
<td>0.56</td>
<td>0.60 +7%</td>
<td><strong>0.64</strong> +14%</td>
</tr>
</tbody>
</table>
Preliminary feature analysis

- Not all the features are necessary

Figure 2: Multidimensional mapping of feature similarity computed using Kendall’s τ coefficient. The size of each circle is proportional to the single-feature model score.
Preliminary feature analysis

- $\rho^{MW}$ has the 4-th best performance, but it is ranked 19-th

<table>
<thead>
<tr>
<th>Features</th>
<th>Rank</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>P@5</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(c</td>
<td>e)</td>
<td>1</td>
<td>0.68</td>
<td>0.72</td>
<td>0.47</td>
<td>0.33</td>
</tr>
<tr>
<td>J(e, c)</td>
<td>2</td>
<td>0.62</td>
<td>0.66</td>
<td>0.44</td>
<td>0.31</td>
<td>0.75</td>
</tr>
<tr>
<td>Friend(e,c)</td>
<td>24</td>
<td>0.59</td>
<td>0.64</td>
<td>0.42</td>
<td>0.31</td>
<td>0.71</td>
</tr>
<tr>
<td>$\rho^{MW}$ (e,c)</td>
<td>19</td>
<td>0.59</td>
<td>0.63</td>
<td>0.42</td>
<td>0.31</td>
<td>0.72</td>
</tr>
<tr>
<td>$J_{in-out}(e,c)$</td>
<td>26</td>
<td>0.60</td>
<td>0.63</td>
<td>0.42</td>
<td>0.30</td>
<td>0.74</td>
</tr>
<tr>
<td>AvgFr(e,c)</td>
<td>3</td>
<td>0.57</td>
<td>0.62</td>
<td>0.40</td>
<td>0.30</td>
<td>0.69</td>
</tr>
<tr>
<td>P(e,c)</td>
<td>27</td>
<td>0.56</td>
<td>0.60</td>
<td>0.39</td>
<td>0.28</td>
<td>0.70</td>
</tr>
<tr>
<td>$\rho^{MW}_{in-out}(a,b)$</td>
<td>9</td>
<td>0.56</td>
<td>0.60</td>
<td>0.40</td>
<td>0.29</td>
<td>0.71</td>
</tr>
<tr>
<td>$J_{in-out}(e,c)$</td>
<td>4</td>
<td>0.54</td>
<td>0.58</td>
<td>0.39</td>
<td>0.28</td>
<td>0.67</td>
</tr>
<tr>
<td>$\rho^{MW}_{out}(a,b)$</td>
<td>17</td>
<td>0.52</td>
<td>0.55</td>
<td>0.37</td>
<td>0.27</td>
<td>0.65</td>
</tr>
<tr>
<td>$\chi^2_{in-out}(e,c)$</td>
<td>25</td>
<td>0.51</td>
<td>0.55</td>
<td>0.37</td>
<td>0.27</td>
<td>0.64</td>
</tr>
<tr>
<td>P(e</td>
<td>c)</td>
<td>22</td>
<td>0.48</td>
<td>0.54</td>
<td>0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>H(c)</td>
<td>5</td>
<td>0.48</td>
<td>0.51</td>
<td>0.30</td>
<td>0.20</td>
<td>0.68</td>
</tr>
<tr>
<td>$\chi^2_{out}(e,c)$</td>
<td>16</td>
<td>0.47</td>
<td>0.50</td>
<td>0.34</td>
<td>0.24</td>
<td>0.61</td>
</tr>
<tr>
<td>AvgFr(c,e)</td>
<td>21</td>
<td>0.44</td>
<td>0.49</td>
<td>0.33</td>
<td>0.25</td>
<td>0.56</td>
</tr>
<tr>
<td>P(c)</td>
<td>13</td>
<td>0.47</td>
<td>0.49</td>
<td>0.29</td>
<td>0.19</td>
<td>0.66</td>
</tr>
<tr>
<td>PMI(e,c)</td>
<td>23</td>
<td>0.42</td>
<td>0.48</td>
<td>0.32</td>
<td>0.25</td>
<td>0.53</td>
</tr>
<tr>
<td>$\chi^2_{in-out}(e,c)$</td>
<td>11</td>
<td>0.44</td>
<td>0.46</td>
<td>0.33</td>
<td>0.23</td>
<td>0.58</td>
</tr>
<tr>
<td>$P(e\rightarrow c)$</td>
<td>18</td>
<td>0.37</td>
<td>0.38</td>
<td>0.24</td>
<td>0.15</td>
<td>0.55</td>
</tr>
<tr>
<td>Link(e→c)</td>
<td>20</td>
<td>0.37</td>
<td>0.38</td>
<td>0.24</td>
<td>0.15</td>
<td>0.55</td>
</tr>
<tr>
<td>$P(c\rightarrow e)$</td>
<td>12</td>
<td>0.35</td>
<td>0.36</td>
<td>0.22</td>
<td>0.14</td>
<td>0.52</td>
</tr>
<tr>
<td>Link(c→e)</td>
<td>15</td>
<td>0.31</td>
<td>0.33</td>
<td>0.21</td>
<td>0.14</td>
<td>0.46</td>
</tr>
<tr>
<td>$KL(c</td>
<td></td>
<td>e)$</td>
<td>10</td>
<td>0.32</td>
<td>0.32</td>
<td>0.19</td>
</tr>
<tr>
<td>Link(c↔e)</td>
<td>14</td>
<td>0.28</td>
<td>0.29</td>
<td>0.17</td>
<td>0.11</td>
<td>0.45</td>
</tr>
<tr>
<td>$KL(e</td>
<td></td>
<td>c)$</td>
<td>8</td>
<td>0.26</td>
<td>0.28</td>
<td>0.17</td>
</tr>
<tr>
<td>P(e)</td>
<td>6</td>
<td>0.08</td>
<td>0.11</td>
<td>0.06</td>
<td>0.06</td>
<td>0.17</td>
</tr>
<tr>
<td>H(e)</td>
<td>7</td>
<td>0.08</td>
<td>0.11</td>
<td>0.06</td>
<td>0.06</td>
<td>0.17</td>
</tr>
</tbody>
</table>
We developed an Entity Linking framework:
- Open source software (github)
- Designed to use commodity hardware
- Modular and easily pluggable
- Provide both REST APIs and webapp

Current Status:
- Spotting based on a dictionary
- Several features available for entities and spots
- Several relatedness measures available
- Wikiminer, TAGME and Referent Graph implemented

Available at: [http://dexter.isti.cnr.it/](http://dexter.isti.cnr.it/)
What’s Next

- Capture the “aboutness” of a document
  - Find the most central topics of a document
  - Assign a centrality score to the entities of a given document
  - Solutions proposed in literature use words/phrases as topics
    - Tips: use the entities provided by an Entity Linking approach

- Endless Applications
  - News stream analysis
    - Summarization
    - Recommendation
    - Identification of news facets
  - Web queries and web documents annotation
Dexter Annotation Framework

**Problem:** freely available manually annotated datasets are:
- often quite noisy and not very coherent
- they miss of the centrality score
- old datasets

**Solution:** we developed an Annotation framework:
- Open source (both the source code and the produced dataset)
- Simple and powerful to use
- Smart and elegant UI
- Split the annotation task in two steps:
  - **Step 1:** annotate a document with its relevant entities
    - **Spots:** create new spots, enlarge/restrict spots, delete existing spots
    - **Entities:** select relevant entity for a spot, add a not listed entity
  - **Step 2:** assign to the previously selected entities a centrality score (0 to 3)
Dexter Annotation Framework

• **Idea**: “fuzzy” annotations
  • Takes into account that the task is often quite subjective
  • Could help to improve the disambiguation

• **Current Status**:
  • Opened to the HPC lab only from 3 weeks
  • Documents taken from the same reuters subset of CONLL
  • 90 annotated documents (~30 uniques)
  • ~2500 annotated spots (~27 spots/doc on average)
  • …the dataset is still small to make some kind of statistical analysis but…
    • It seems that the agreement between different users is on average low

• **Available at**: [http://bruciato.isti.cnr.it:8888/dexter-annotate/](http://bruciato.isti.cnr.it:8888/dexter-annotate/)
Thank you

THANKS FOR YOUR ATTENTION

AND

ANY QUESTIONS?