WebTLab: A co-occurrence-based approach to KBP 2010 Entity-Linking task

Prof. Luis Sanchez-Fernandez
Web Technologies Laboratory
University Carlos III of Madrid
http://webtlab.it.uc3m.es
Outline

- Introduction
- Strategy
- Outline of the system
- Results
- Conclusions
Introduction

Some name conventions!

Entity: text + type

Instance: a particular person, location (GPE), organization, ...

ESSENCE.COM: How’s your book tour going? What has the feedback been from women who have read it thus far?

NENE LEAKES: The turnout at all the signings has been amazing. I’ve been averaging about 300 people at most of them. In my hometown of Athens, Georgia, I received a very warm welcome—over 400 people came out to the signing with a first-time author, but I’ve heard that’s really good. Women have said they know me better from reading the book. I’ve gotten nothing but love.

ESSENCE.COM: This season on “The Real Housewives of Atlanta” you reunited with Kim. What do you have to say to those viewers who think you guys are fake because you weren’t cool when the show ended last season?

LEAKES: Well, it’s not fake. We’re just like any other group of girls that hang out. We have our ups and downs, we get back together—that’s real life. There’s nothing on the show that is fake. It is what it is.

ESSENCE.COM: You directed an alter-ego photo shoot in an episode this season which we all loved. Do you have the bug to do more creative directing down the road?

LEAKES: I’m a reality star, I’m an author, I’m a businesswoman and I’m a director. I really like the entertainment industry. Directing? I love that kind of stuff. In tonight’s episode, I will reveal the final photos from the shoot to all of the ladies. It is difficult to direct people because everyone doesn’t see your vision, but at the end I think they all appreciated what I did for them. The job was definitely meant for me.
Strategy

ESSENCE.COM: How's your book tour going? What has the feedback been from women who have read it thus far?

NENE LEAKES: The turnout at all the signings has been amazing. I've been averaging about 300 people at most of them. In my hometown of Athens, Georgia, I received a very warm welcome—over 400 people came out for the signing. I'm a first-time author, but I've heard that's really good. Women have said they know me better from reading the book. I've gotten nothing but love.

ESSENCE.COM: This season on “The Real Housewives of Atlanta” you reunited with Kim. What do you have to say to those viewers who think you guys are fake because you weren't cool when the show ended last season?

LEAKES: Well, it's not fake. We're just like any other group of girls that hang out. We have our ups and downs, we get back together—that's real life. There's nothing on the show that is fake. It is what it is.

ESSENCE.COM: You directed an alter-ego photo shoot in an episode this season which we all loved. Do you have the bug to do more creative directing down the road?

LEAKES: I'm a reality star, I'm an author, I'm a businesswoman and I'm a director. I really like the entertainment industry. Directing? I love that kind of stuff. In tonight's episode, I will reveal the final photos from the shoot to all of the ladies. It is difficult to direct people because everyone doesn’t see your vision, but at the end I think they all appreciated what I did for them. The job was definitely meant for me.
Approach

- Find entities in document
- For each entity, identify candidate instances that are compatible with the entity name
- Assign a ranking value to each candidate instance: $0 \leq r \leq 1$
  - Greater ranking values indicate greater likelihood of occurrence
Semantic coherence (in terms of ranking)

“An instance would have a high ranking value if the instances that typically co-occur with it also have high ranking values”

\[
r(I_i) = \sum_{j \in C_i} \text{Cooc}(I_i, I_j) r(I_j)
\]
We can add a vector $E$ that accounts for other context information

Equation similar to Google PageRank

$$\vec{R} = (1 - \alpha) \vec{A} \vec{R} + \alpha \vec{E}$$
Outline of the system
Use 3 open source tools

- GATE (ANNIE): University of Sheffield
- Named Entity Recognizer (NER): Stanford University
- LbjNerTagger: University of Illinois at Urbana-Champaign

Decide by majority

Limit the number of entities to 100
Instance finder & filter

- Alternative instance names extracted by processing a Wikipedia dump
  - Page titles, redirects, disambiguation pages, anchors
    - Indexed by Lucene
- Candidate instances are obtained by querying Lucene
- Candidate instances weighted by combining Lucene scores and PageRank values
- Filtering limits the maximum number of candidates

http://webtlab.it.uc3m.es
Instance ranker

\[ R = (k_L A_L \vec{R} + k_C A_C \vec{R}) + k_E \vec{E} \]

\( A_L \): based on direct links

\( A_C \): based on instance co-occurrence in Wikipedia pages

\( A_L \): candidate instance weights passed by Wikipedia pages
Instance selector

- If no candidate found: NIL
- If the first ranked candidate is not in the KB: NIL
- Otherwise, decide by comparing the two candidates with highest rank
  - Different threshold if exact match found

\[
\text{Plausibility} = \frac{\text{TopInstanceRanking}}{\text{SecondInstanceRanking}}
\]
## Results I

<table>
<thead>
<tr>
<th>Run</th>
<th>I. finder</th>
<th>I. ranker</th>
<th>I. selector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_L$</td>
<td>$\alpha_{PR}$</td>
<td>$k_L$</td>
</tr>
<tr>
<td>WebTLab1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>WebTLab2</td>
<td>0.8</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>WebTLab3</td>
<td>0.8</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>2250 queries</th>
<th>1020 non-NIL</th>
<th>1230 NIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>WebTLab1</td>
<td>0.7698</td>
<td>0.6647</td>
<td>0.8569</td>
</tr>
<tr>
<td>WebTLab2</td>
<td>0.7636</td>
<td>0.6098</td>
<td>0.8911</td>
</tr>
<tr>
<td>WebTLab3</td>
<td>0.7596</td>
<td>0.6049</td>
<td>0.8878</td>
</tr>
</tbody>
</table>

$$\widehat{R} = \left( k_L A_L \widehat{R} + k_C A_C \widehat{R} \right) + k_E \widehat{E}$$
Results II

<table>
<thead>
<tr>
<th>Run</th>
<th>I. finder</th>
<th>I. ranker</th>
<th>I. selector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_L$</td>
<td>$\alpha_{PR}$</td>
<td>$k_L$</td>
</tr>
<tr>
<td>WebTLab1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>WebTLab2</td>
<td>0.8</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>WebTLab3</td>
<td>0.8</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

$$\vec{R} = \left( k_L A_L \vec{R} + k_C A_C \vec{R} \right) + k_E \vec{E}$$
### Results III: ablation tests

#### Run I. finder I. ranker I. selector

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_L$</td>
<td>$\alpha_{PR}$</td>
<td>$k_L$</td>
<td>$k_C$</td>
</tr>
<tr>
<td>Lucene</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Lucene-PR</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>WebTLab1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.55</td>
<td>0.25</td>
</tr>
</tbody>
</table>

#### Run 2250 queries 1020 non-NIL 1230 NIL

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucene</td>
<td>0.6364</td>
<td>0.2627</td>
<td>0.9463</td>
</tr>
<tr>
<td>Lucene-PR</td>
<td>0.6658</td>
<td>0.3363</td>
<td>0.9390</td>
</tr>
<tr>
<td>WebTLab1</td>
<td><strong>0.7698</strong></td>
<td>0.6647</td>
<td>0.8569</td>
</tr>
</tbody>
</table>

\[
\vec{R} = \left( k_L A_L \vec{R} + k_C A_C \vec{R} \right) + k_E \vec{E}
\]
Approach based on instance co-occurrence

Text from Wikipedia restricted to: titles, anchors

Results considered promising
  Should improve for GPE
Thank You!

Questions?