ENTITY-BASED QUERY INTERPRETATION

Bachelor’s Defence

Marcel Gohsen

Bauhaus-Universität Weimar

04 July 2018
Problem of Query Interpretation

new york times square dance
PROBLEM OF QUERY INTERPRETATION

19 U.S. States Have Designated It As Their Official State Dance

Square Dance

new york times square dance
Problem of Query Interpretation

new york  times square  dance
Entities in Queries

- Named Entity
  - object from the real world with a proper name
  - e.g., person, location, organization

- Entities in Queries
  - Definitions differ
  - May be limited to proper nouns \(^1\)
  - May include general concepts \(^2\)

\(^1\) [Hasibi et al., 2015]
\(^2\) [Cornolti et al., 2016]
**Used Entity Taxonomy**

- Based on “Extended Named Entity Hierarchy” [Sekine et al., 2002]
- 8 main classes
- 108 specialized subclasses

- for example: removed class *Units* (e.g., kilogram)
Traditional Problem Statements
Linking an entity in a query to the most likely candidate in some knowledge base.

\[
\begin{align*}
\text{obama mother} & \rightarrow \ ("obama", \text{Barack Obama}) \\
\text{new york pizza manhattan} & \rightarrow \ ("new york", \text{New York City}) \\
& \quad \text{("manhattan", Manhattan)}
\end{align*}
\]

Issues:

- Non-overlapping entities only
Finding subsets of semantic compatible non-overlapping linked entities

- obama mother $\rightarrow \{\text{Barack Obama}\}$
- new york pizza manhattan $\rightarrow \{\text{New York City}, \text{Manhattan}\}$
  $\{\text{New York-Style Pizza}, \text{Manhattan}\}$

Issues:
- Imprecise interpretations
- Explicit mentioned entities only
Finding subsets of semantic compatible non-overlapping linked entities

- obama mother → {Barack Obama} mother?
- new york pizza manhattan → {New York City, Manhattan} pizza?
  {New York-Style Pizza, Manhattan}

Issues:
- Imprecise interpretations
- Explicit mentioned entities only
REDEFINED PROBLEMS
Explicit Entity Recognition

**Given:** - Query

**Task:** - Identifying explicit mentioned entities in a query
- Segment is an entity’s name or surface form

- obama mother → (“obama”, Barack Obama)
  (“obama”, Michelle Obama)
  (“obama”, Natsuki Obama)...

- new york pizza manhattan → (“new york”, New York City)
  (“new york”, New York (state))
  (“manhattan”, Manhattan)
  (“manhattan”, Manhattan (film))...
Implicit Entity Recognition

**Given:** - Query

**Task:** - Identifying implicitly referenced entities in a query
- Segment is a description of an entity

- *obama mother* → (“obama mother”, Ann Dunham)
  (“obama mother”, Marian Shields)...
- *new york pizza manhattan* → ∅
- *president of usa* → (“president of usa”, Donald Trump)
  (“president of usa”, Barack Obama)
  (“president of usa”, George W. Bush)
**Entity-Based Query Interpretation**

**Given:**
- Query
- Explicit entities in query
- Implicit entities in query

**Task:**
- Semantically segmentation of query
- Replacing explicit and implicit entity-mentions with entities

\[
\text{obama mother} \rightarrow \{\text{Barack Obama, Ann Dunham}\} \\
\{\text{Michelle Obama, Marian Shields}\} \\
\ldots
\]

\[
\text{new york pizza manhattan} \rightarrow \{\text{New York City, “pizza”, Manhattan}\} \\
\ldots
\]
CORPORA
Dataset of the ERD’14 Challenge
- 91 queries
  - 45 queries having annotated entities
- Provides query interpretation

obama family tree → \{Barack Obama\}
east ridge high school → \{East Ridge High School (FL)\}
  \{East Ridge High School (MN)\}
  \{East Ridge High School (KY)\}
YSQLE DATASET [YAHOO, 2010]

- “Yahoo Search Query Log to Entities”
- 2635 queries
  - 2583 queries having annotated entities
- No query interpretations

france 1998 final → France National Football Team, France, Fifa World Cup 1998 Final

obama mother → Barack Obama, Ann Dunham
Collection for Entity Search

467 queries

No query interpretations

Introduced relevance levels

- 2: highly relevant
- 1: relevant
- 0: irrelevant

john lennon, parents → \{Julia Lennon : 2,
                          Alfred Lennon : 1
                          ... : 0\}
Query Interpretation Corpus

- Queries from the three existing corpora
- Manually (re-)annotated:
  - Query difficulty judgments \{easy | moderate | hard\}
  - Explicit entities with relevance judgments \{relevant | plausible\}
  - Implicit entities with relevance judgments
  - Entity-based query interpretations with relevance judgments
- 2068 queries
  - 1578 queries with explicit entities
  - 131 queries with implicit entities
  - 1597 queries with query interpretations
ALGORITHMIC APPROACHES
Entity Linking Steps

Typical steps for entity linking frameworks

(i) **Candidate Generation**
(ii) **Scoring**
(iii) **Selecting**
(1) **Candidate Generation**

- DBpedia Ontology [DBpedia, 2017] used for classification
  - Digital representation of our entity taxonomy
- Index all Wikipedia articles that represent entities
- Retrieve the top 100 articles from the index containing a segment from the query
- Retrieve for each segment of the query
(ii) Scoring

\[ Jaccard(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \]

\[ norm = \frac{|segment|}{|query|} \]
(iii) Selection

- Precision vs. Recall
- Threshold vs. Fixed number of retrieved entities
- Take the top 20 entities by score
EVALUATION
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>rec</th>
<th>prec</th>
<th>$F_1$</th>
<th>$rec^*$</th>
<th>$F_1^*$</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nordlys EL</td>
<td>.55</td>
<td>.69</td>
<td>.58</td>
<td>.50</td>
<td>.52</td>
<td>4400 ms</td>
</tr>
<tr>
<td>Explicit Entity Approach</td>
<td>.40</td>
<td>.16</td>
<td>.18</td>
<td>.35</td>
<td>.16</td>
<td>270 ms</td>
</tr>
<tr>
<td>Smaph</td>
<td>.38</td>
<td>.45</td>
<td>.37</td>
<td>.32</td>
<td>.31</td>
<td>117000 ms</td>
</tr>
<tr>
<td>TagMe</td>
<td>.37</td>
<td>.39</td>
<td>.33</td>
<td>.31</td>
<td>.28</td>
<td>40 ms</td>
</tr>
<tr>
<td>Nordlys ER</td>
<td>.33</td>
<td>.05</td>
<td>.07</td>
<td>.29</td>
<td>.06</td>
<td>1900 ms</td>
</tr>
</tbody>
</table>
Conclusion

- Refined problem statements for entity linking
  - Ambiguous explicit and implicit entities
  - More precise and diverse query interpretations

- Query Interpretation Corpus
  - Comparatively large corpus
  - Explicit and implicit entities
  - Query interpretations

- Algorithmic Approaches
  - Efficient explicit entity recognition
  - Implicit entity recognition prototype

Thank you for the attention!
ERD’14: ENTITY RECOGNITION AND DISAMBIGUATION CHALLENGE.

A PIGGYBACK SYSTEM FOR JOINT ENTITY MENTION DETECTION AND LINKING IN WEB QUERIES.

DBPEDIA (2017).
DBPEDIA ONTOLOGY 2016-10.
https://wiki.dbpedia.org/services-resources/ontology.


Yahoo (2010).

L24 - Yahoo Search Query Log To Entities v1.0.

https://webscope.sandbox.yahoo.com/.
**Evaluation metrics**

\[
\text{prec} = \begin{cases} 
\frac{|E \cap E'|}{|E|}, & \text{if } |E| > 0 \\
1, & \text{if } |E| = 0, |E'| = 0 \\
0, & \text{if } |E| > 0, |E'| > 0 
\end{cases}
\] (1)

\[
\text{rec} = \begin{cases} 
\frac{|E \cap E'|}{|E'|}, & \text{if } |E'| > 0 \\
1, & \text{if } |E| = 0, |E'| = 0 \\
0, & \text{if } |E| > 0, |E'| = 0 
\end{cases}
\] (2)

\[
F_1 = \frac{2 \cdot \text{prec} \cdot \text{rec}}{\text{prec} + \text{rec}}
\] (3)
Evaluation metrics

\[ w = \frac{\sum_{e \in E \cap E'} rel(e)}{\sum_{e' \in E'} rel(e')} \]  

(4)

\[ rec^* = w \cdot rec \]  

(5)

\[ F_1^* = \frac{2 \cdot prec \cdot rec^*}{prec + rec^*} \]  

(6)
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>prec</th>
<th>rec</th>
<th>$F_1$</th>
<th>$rec^*$</th>
<th>$F_1^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagMe</td>
<td>.52</td>
<td>.49</td>
<td>.44</td>
<td>.42</td>
<td>.37</td>
</tr>
<tr>
<td>Smaph</td>
<td>.58</td>
<td>.48</td>
<td>.47</td>
<td>.40</td>
<td>.39</td>
</tr>
<tr>
<td>Explicit Entity Approach</td>
<td>.14</td>
<td>.47</td>
<td>.17</td>
<td>.40</td>
<td>.14</td>
</tr>
<tr>
<td>Nordlys EL</td>
<td>.64</td>
<td>.45</td>
<td>.49</td>
<td>.38</td>
<td>.41</td>
</tr>
<tr>
<td>Nordlys ER</td>
<td>.04</td>
<td>.43</td>
<td>.07</td>
<td>.37</td>
<td>.07</td>
</tr>
</tbody>
</table>