Entity Linking in Queries: Tasks and Evaluation

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg
ICTIR conference, September 2015
Entity linking

Definition from Wikipedia:

In natural language processing, entity linking, named entity disambiguation (NED), named entity recognition and disambiguation (NERD) or named entity normalization (NEN)[1] is the task of determining the identity of entities mentioned in text. It is distinct from named entity recognition (NER) in that it identifies not the occurrence of names (and a limited classification of those), but their reference.

https://en.wikipedia.org/wiki/Natural_language_processing

Why entity linking in queries?

• ~70% of queries contain entities
• To exploit semantic representation of queries

Improves:
• Ad-hoc document retrieval
• Entity retrieval
• Query understanding
• Understanding users’ task (Tasks track, TREC)

It is different …

Different from conventional entity linking:

• Limited or even no context
• A mention may be linked to more than one entity

{France, FIFA world cup}

or

{France national football team, FIFA world cup}
In this talk

How entity linking should be performed for queries?

➢ Task:
  “Semantic Mapping” or “Interpretation Finding”?

➢ Evaluation metrics

➢ Test collections

➢ Methods
In this talk

How entity linking should be performed for queries?

➤ Task:
  “Semantic Mapping” or “Interpretation Finding”?

➤ Evaluation metrics

➤ Test collections

➤ Methods
Entity linking

- Output is **set of entities**
- Each mention is linked to a single entity
- Mentions do not overlap
- Entities are explicitly mentioned

<table>
<thead>
<tr>
<th>obama mother</th>
<th>the music man</th>
<th>new york pizza manhattan</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Barack Obama}</td>
<td>{The Music Man}</td>
<td>{New York City, Manhattan}</td>
</tr>
</tbody>
</table>
Semantic mapping

- Output is **ranked list of entities**
- Mentions can overlap and be linked to multiple entities
- Entities may not be explicitly mentioned
- Entities do not need to form semantically compatible sets
- False positive are not penalized

<table>
<thead>
<tr>
<th>obama mother</th>
<th>the music man</th>
<th>new york pizza manhattan</th>
</tr>
</thead>
</table>
| Ann Dunham Barack Obama | The Music Man  
The Music Man (1962 film)  
The Music Man (2003 film)  
... | New York City  
New York-style pizza  
Manhattan  
Manhattan pizza  
... |
Interpretation finding

• Output is **set(s) of semantically related entity sets**
• Each entity set is an interpretation of the query
• Mention do not overlap within a set

<table>
<thead>
<tr>
<th>obama mother</th>
<th>the music man</th>
<th>new york pizza manhattan</th>
</tr>
</thead>
<tbody>
<tr>
<td>{{Barack Obama}}</td>
<td>{The Music Man}</td>
<td>{New York City, Manhattan}, {New York-style pizza, Manhattan}</td>
</tr>
<tr>
<td></td>
<td>{The Music Man (1962 film)}, {The Music Man (2003 film)}</td>
<td></td>
</tr>
</tbody>
</table>
## Tasks summary

<table>
<thead>
<tr>
<th>Entity Linking</th>
<th>Semantic Mapping</th>
<th>Interpretation Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities explicitly mentioned?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Mentions can overlap?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Results format</td>
<td>Set</td>
<td>Ranked list</td>
</tr>
<tr>
<td>Evaluation criteria</td>
<td>Mentioned entities found</td>
<td>Relevant entities found</td>
</tr>
<tr>
<td>Evaluation metrics</td>
<td>Set-based</td>
<td>Ranked-based</td>
</tr>
</tbody>
</table>

* Not within the same interpretation

Entity linking requirements are relaxed in semantic mapping.
In this talk

How entity linking should be performed for queries?

➢ Task:
  “Semantic Mapping” or “Interpretation Finding”?

➢ Evaluation metrics

➢ Test collections

➢ Methods
Evaluation

- Macro-averaged metrics (precision, recall, F-measure)
- Matching condition:
  - Interpretation sets should exactly match the ground truth

\[ P = \frac{|I \cap \hat{I}|}{|I|} \quad \text{System query interpretation} \]
\[ R = \frac{|I \cap \hat{I}|}{|\hat{I}|} \quad \text{Ground truth query interpretation} \]

\[ I = \{E_1, \ldots, E_n\} \]
\[ \hat{I} = \{\hat{E}_1, \ldots, \hat{E}_m\} \]

What if \( I = \emptyset \) or \( \hat{I} = \emptyset \)?
Evaluation (revisited)

Solution:

\[
P_{\text{int}} = \begin{cases} 
|I \cap \hat{I}|/|I|, & I \neq \emptyset \\
1, & I = \emptyset, \hat{I} = \emptyset \\
0, & I = \emptyset, \hat{I} \neq \emptyset 
\end{cases}
\]

\[
R_{\text{int}} = \begin{cases} 
|I \cap \hat{I}|/|\hat{I}|, & \hat{I} \neq \emptyset \\
1, & \hat{I} = \emptyset, I = \emptyset \\
0, & \hat{I} = \emptyset, I \neq \emptyset 
\end{cases}
\]

System output matches ground truth.

System output does not match ground truth.

This evaluation is methodologically correct, but strict.
Lean evaluation

- Partial matches are not rewarded in $P_{\text{int}}, R_{\text{int}}$
- E.g. $\{\{\text{New York City, Manhattan}\}\} \neq \{\{\text{New York City}\}, \{\text{Manhattan}\}\}$

Solution: Combine them with entity-based metrics.

\[
P = \frac{P_{\text{int}} + P_{\text{ent}}}{2}
\]
\[
R = \frac{R_{\text{int}} + R_{\text{ent}}}{2}
\]

\[
P_{\text{ent}} = \begin{cases} 
\frac{|E \cap \hat{E}|}{|E|}, & E \neq \emptyset \\
1, & E = \emptyset, \hat{E} = \emptyset \\
0, & E = \emptyset, \hat{E} \neq \emptyset 
\end{cases}
\]

\[
R_{\text{ent}} = \begin{cases} 
\frac{|E \cap \hat{E}|}{|\hat{E}|}, & \hat{E} \neq \emptyset \\
1, & \hat{E} = \emptyset, E = \emptyset \\
0, & \hat{E} = \emptyset, E \neq \emptyset 
\end{cases}
\]
In this talk

How entity linking should be performed for queries?

➤ Task:
   “Semantic Mapping” or “Interpretation Finding”?

➤ Evaluation metrics

➤ Test collections

➤ Methods
Test collections - ERD

The ERD challenge introduced two test collections:

- ERD-dev (91 queries)\(^1\)
- ERD-test (500 queries)
  - Unavailable for traditional offline evaluation

Annotation rules:

- The longest mention is used for entities
- Only proper noun entities are annotated (e.g., companies, locations)
- Overlapping mentions are not allowed within a single interpretation

\(^1\) http://web-ngram.research.microsoft.com/erd2014/Datasets.aspx
Test collections - YSQLE

Yahoo Search Query Log to Entities (YSQLE)

• 2398 queries, manually annotated with Wikipedia entities
• Designed for training and testing entity linking systems for queries

Issues:

• Not possible to automatically form interpretation sets
  – E.g. Query “france world cup 1998”
• Linked entities are not necessarily mentioned explicitly
  – E.g. Query “charlie sheen lohan” is annotated with Anger Management (TV series)
• Annotations are not always complete
  – E.g. Query “louisville courier journal” is not annotated with Louisville, Kentucky

YSQLE is meant for the **semantic mapping** task
Test collections - Y-ERD

Y-ERD is manually re-annotated based on:
• YSQLE annotations
• ERD rules

Additional rules:
• Site search queries are not linked
  – E.g. Query “facebook obama slur” is only linked to Barack Obama
• Clear policy about misspelled mentions
  – Two versions of Y-ERD is made available

Y-ERD is made publicly available
In this talk

How entity linking should be performed for queries?

➤ Task:
  “Semantic Mapping” or “Interpretation Finding”?

➤ Evaluation metrics

➤ Test collections

➤ Methods
Methods

Pipeline architecture for two tasks:

Semantic mapping

Mention detection → set of mentions → Candidate entity ranking → ranked list of entities → Interpretation finding → interpretations

Interpretation Finding
The goal of entity linking in queries
Mention detection

Entity name variants are gathered from:

- **KB**: A manually curated knowledge base (DBpedia)
- **WEB**: Freebase Annotations of the ClueWeb Corpora (FACC)

<table>
<thead>
<tr>
<th></th>
<th>YSQLE</th>
<th>Y-ERD</th>
<th>ERD</th>
</tr>
</thead>
<tbody>
<tr>
<td>KB</td>
<td>0.7489</td>
<td>0.7976</td>
<td>0.8556</td>
</tr>
<tr>
<td>Web</td>
<td>0.9127</td>
<td>0.9716</td>
<td>0.9956</td>
</tr>
<tr>
<td>KB+Web</td>
<td>0.9163</td>
<td>0.9724</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Recall in mention detection step

Methods

Pipeline architecture for two tasks:

Semantic mapping

Interpretation Finding
The goal of entity linking in queries
Candidate entity ranking

Ranking using language models:

\[
P(e|q) = \frac{P(q|e)P(e)}{P(q)} \propto P(e)P(q|\theta_e)
\]

Scores should be comparable across queries
- \(P(q)\) should be considered

\[
P(e|q) = P(e) \prod_{t \in q} \frac{P(t|\theta_e)P(t|q)}{\prod_{t \in q} P(t|C)P(t|q)}
\]

Mixture of Language Models (MLM)

\[
\sum_{f \in F} \mu_f P(t|\theta_{e,f})
\]

P(q) - Query length normalization

Candidate entity ranking

Combining MLM and Commonness:

\[ P(e|q) \propto \arg \max_{m \in q} P(e|m)P(q|e) \]

- **Commonness**
  - Probability of entity \( e \) being the link target of mention \( m \)

- Query length normalized MLM score
  \[ \frac{n(m, e)}{\sum_{e'} n(m, e')} \]
Candidate entity ranking

Semantic mapping results on YSQLFE:

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>S@1</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMNS</td>
<td>0.6334</td>
<td>0.5751</td>
<td>0.6442</td>
</tr>
<tr>
<td>MLM</td>
<td>0.4582</td>
<td>0.3601</td>
<td>0.4638</td>
</tr>
<tr>
<td>MLMc</td>
<td>0.6228</td>
<td>0.5413</td>
<td>0.6312</td>
</tr>
<tr>
<td>MLMcg</td>
<td>0.7078</td>
<td>0.6403</td>
<td>0.7151</td>
</tr>
<tr>
<td>TAGME</td>
<td>0.6230</td>
<td>0.6016</td>
<td>0.6385</td>
</tr>
</tbody>
</table>

TAGME is an entity linking system.
- Should not be evaluated using rank-based metrics
- Should not be compared with semantic mapping results

Methods

Pipeline architecture for two tasks:

**Semantic mapping**

- **Mention detection**
  - query → set of mentions

- **Candidate entity ranking**
  - set of mentions → ranked list of entities

**Interpretation Finding**

- The goal of entity linking in queries

- **Interpretation finding**
  - ranked list of entities → interpretations
Interpretation finding

Greedy Interpretation Finding (GIF):

Example query: “jacksonville fl riverside”

<table>
<thead>
<tr>
<th>Mention</th>
<th>Entity</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>“jacksonville fl”</td>
<td>Jacksonville Florida</td>
<td>0.9</td>
</tr>
<tr>
<td>“jacksonville”</td>
<td>Jacksonville Florida</td>
<td>0.8</td>
</tr>
<tr>
<td>“riverside”</td>
<td>Riverside Park (Jacksonville)</td>
<td>0.6</td>
</tr>
<tr>
<td>“jacksonville fl”</td>
<td>Naval Station Jacksonville</td>
<td>0.2</td>
</tr>
<tr>
<td>“riverside”</td>
<td>Riverside (band)</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Step 1: Pruning based on a score threshold (0.3)

Step 2: Pruning containment mentions

Step 3: Forming interpretation sets

\{ \{Jacksonville Florida, Riverside Park (Jacksonville)\} \}
## Interpretation finding

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TopRanked</td>
<td>0.4554</td>
<td>0.4542</td>
<td>0.4545</td>
<td>0.4771</td>
<td>0.465</td>
<td>0.4689</td>
</tr>
<tr>
<td>TAGME</td>
<td>0.6647</td>
<td>0.6642</td>
<td>0.6643</td>
<td>0.6821</td>
<td>0.6853</td>
<td>0.6815</td>
</tr>
<tr>
<td>GIF-CMNS</td>
<td>0.6927</td>
<td>0.6938</td>
<td>0.6929</td>
<td>0.7093</td>
<td>0.7072</td>
<td>0.7062</td>
</tr>
<tr>
<td>GIF-MLM</td>
<td>0.5259</td>
<td>0.5254</td>
<td>0.5255</td>
<td>0.5363</td>
<td>0.5387</td>
<td>0.5361</td>
</tr>
<tr>
<td>GIF-MLMc</td>
<td>0.6351</td>
<td>0.6354</td>
<td>0.6348</td>
<td>0.6422</td>
<td>0.642</td>
<td>0.6409</td>
</tr>
<tr>
<td><strong>GIF-MLMcg</strong></td>
<td><strong>0.7191</strong></td>
<td><strong>0.7213</strong></td>
<td><strong>0.7195</strong></td>
<td><strong>0.7305</strong></td>
<td><strong>0.7308</strong></td>
<td><strong>0.7288</strong></td>
</tr>
<tr>
<td>TopRanked</td>
<td>0.3846</td>
<td>0.3645</td>
<td>0.3700</td>
<td>0.4231</td>
<td>0.3837</td>
<td>0.3956</td>
</tr>
<tr>
<td>TAGME</td>
<td>0.7143</td>
<td>0.7015</td>
<td>0.7051</td>
<td><strong>0.7418</strong></td>
<td><strong>0.7372</strong></td>
<td><strong>0.7333</strong></td>
</tr>
<tr>
<td>GIF-CMNS</td>
<td>0.5824</td>
<td>0.5824</td>
<td>0.5824</td>
<td>0.6071</td>
<td>0.5962</td>
<td>0.5998</td>
</tr>
<tr>
<td>GIF-MLM</td>
<td>0.5824</td>
<td>0.5608</td>
<td>0.5659</td>
<td>0.5934</td>
<td>0.5718</td>
<td>0.5760</td>
</tr>
<tr>
<td>GIF-MLMc</td>
<td>0.7253</td>
<td>0.7037</td>
<td>0.7088</td>
<td>0.7445</td>
<td>0.7174</td>
<td>0.7234</td>
</tr>
<tr>
<td><strong>GIF-MLMcg</strong></td>
<td><strong>0.7143</strong></td>
<td><strong>0.7125</strong></td>
<td><strong>0.7114</strong></td>
<td><strong>0.7335</strong></td>
<td><strong>0.7262</strong></td>
<td><strong>0.7260</strong></td>
</tr>
</tbody>
</table>
Take home messages

• Entity linking in queries is different from documents

• Different flavors, different evaluation criteria:
  – Interpretation finding (yes)
  – Semantic mapping (no)

• Ultimate goal should be interpretation finding

• SM and EL should not be compared to each other

Thanks!