Thanks to ACM SIGIR for Travel Grant

Ranking Entities Through Text and Knowledge

Michael Schuhmacher, Laura Dietz, Simone Paolo Ponzetto
Current web search engines are tailed to single entity queries.
Our overall motivation is to find general query-relevant entities.
The task here is entity retrieval for general web queries

What we want:

- Retrieve entities for general web query
  “Argentine British relationships” (TREC Web Track)
The task here is entity retrieval for general web queries

What we want:

- Retrieve entities for general web query
  “Argentine British relationships” (TREC Web Track)

What we not aim for:

- Answer single entity queries
  “Jonas Salk”
- Answer entity type queries
  “European countries where I can pay with Euros” (INEX)
The task here is entity retrieval for general web queries

What we want:

- Retrieve entities for general web query
  “Argentine British relationships” (TREC Web Track)

What we not aim for:

- Answer single entity queries
  “Jonas Salk”
- Answer entity type queries
  “European countries where I can pay with Euros” (INEX)

How we do it:

1. Retrieve Documents
2. Extract and Re-rank Entities
The entities are extracted from query-relevant documents

1. **Argentina is still pained by its defeat in the Falklands conflict**
   Buenos Aires believes growing British exports and investments will help reduce the Falklanders' suspicions of all things Argentine.

2. **Survey of Argentina: Nafta option shelved - Plans for Mercosur are well advanced**
   More importantly, Argentina's new foreign policy - reflected [...] in its participation in United Nations forces in Cyprus, Bosnia and the Gulf - has been enthusiastically welcomed in Washington. [...] Mr Menem often repeats his prediction that the Falkland Islands - over which the UK and Argentina fought a brief war in 1982 - will be Argentine by the year 2000.

3. **Argentine threat to UK over S Atlantic fishing**
   Mr Guido di Tella, foreign minister, said: 'Britain will pay a very high price for this joke.'
ENTITY RANKING APPROACH
We want to understand how to rank the document entities

Given: The extracted entities

Question: Which features are helpful for ranking?
We study different ranking features in a LTR setting

Given: The extracted entities
Question: Which features are helpful for ranking?

Method Framework:
Learning-to-Rank (LTR)
- Entity Relevance Features
- Labeled Data Set
- LTR Method
We study different ranking features in a LTR setting

Learning-to-Rank (LTR)
- Entity Relevance Features
- Labeled Data Set
- LTR Method
We study 4 types of entity ranking features

Learning-to-Rank (LTR)

- Entity Relevance Features
  1. Mention Features
  2. Query-Mention Features
  3. Query-Entity Features
  4. Entity-Entity Features

- Labeled Data Set

- LTR Method
1. Mention Features

Query ➔ Ranked Documents ➔ Mentions ➔ Entities

- British
- United Nation
- UK
- United_Kingdom
- United_Nations

(1) The mention frequency includes the document retrieval infos
(2) Queries and entities are compared on the surface level

2. Query-Mention

- a) Surface Level Similarity (String/Levensthein ED)
  - **British** -> **United_Kingdom**
  - Similarity: 1.0

- b) Distribution Semantics (Glove/Jobim)
  - **UK** -> **United_Kingdom**
  - Similarity: 0.8

Argentine-British relations
(3) Query entities are compared with document entities

3. A) Query-Entity Features (Query -> Wikipedia Index)

- Query
- Ranked Documents
- Mentions
- Entities

Wiki Index: Article full text, entity types (WikiSDM)

Argentine British relations

United Kingdom

Carlos Menem

2.6
(3) Query entities are compared with document entities

3. B) Query-Entity Features (Query -> Entity Linker -> Entities)
1. Entity-Entity Features (Exploit DBpedia KB to compare entities)

- Query
- Ranked Documents
- Mentions
- Entities

- Argentina
- Argentina_national_rugby_union_team
- Falkland_Islands_sovereignty_dispute
- Falklands_War
- United_Kingdom
- Falkland_Islands
- United_Nations
- Carlos_Menem
- @UK

...
(4) Document entities are compared against each other

1. Entity-Entity Features (Exploit DBpedia KB to compare entities)

- **Argentina**
  - Carlos Menem dbo:nationality Argentina

- **Carlos Menem**

- **Falklands War**
  - Falklands_War dbo:place Falkland_Islands

- **Falkland Islands**
For evaluation, we created two new goldstandard datasets.

Learning-to-Rank (LTR)

- Entity Relevance Features

- Labeled Data

<table>
<thead>
<tr>
<th></th>
<th>REWQ Robust04</th>
<th>REWQ Clueweb12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc collection</td>
<td>TREC Disk4+5</td>
<td>ClueWeb12</td>
</tr>
<tr>
<td>Queries</td>
<td>TREC Robust Track ‘04</td>
<td>TREC Web Track ‘13/14</td>
</tr>
<tr>
<td># queries</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>Doc retrieval</td>
<td>EQFE (w/ entities)</td>
<td>SDM (text-only)</td>
</tr>
<tr>
<td>Top-k docs</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Top-k entities</td>
<td>50</td>
<td>all</td>
</tr>
<tr>
<td>Entity linker</td>
<td>KB Bridge</td>
<td>FACC1</td>
</tr>
<tr>
<td>GS annotations</td>
<td>graded, 1-5</td>
<td>binary 0/1</td>
</tr>
</tbody>
</table>

- Letor Method
For feature combination, we use established LeToR methods

Learning-to-Rank (LRT)
- Entity Relevance Features
- Labeled Data
- Letor Method

<table>
<thead>
<tr>
<th></th>
<th>Ranking SVM</th>
<th>RankLib</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR type</td>
<td>Pairwise</td>
<td>Listwise</td>
</tr>
<tr>
<td>Method/Impl.</td>
<td>Joachims Ranking SVM</td>
<td>RankLib Coordinate Asc</td>
</tr>
<tr>
<td>Optimized metric</td>
<td>Disordered pairs</td>
<td>MAP</td>
</tr>
</tbody>
</table>
EVALUATION AND FINDINGS
Our approach retrieves interesting and query-relevant entities

Retrieved Ranking for REWQ Rob04 (selected examples)

<table>
<thead>
<tr>
<th>gt</th>
<th>ndcg</th>
<th>query</th>
<th>top-1 entity</th>
<th>top-2 entity</th>
<th>top-3 entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>.895</td>
<td>schengen agreement</td>
<td>Schengen_Agreement</td>
<td>Schengen_Area</td>
<td>Schengen_Information_System</td>
</tr>
<tr>
<td></td>
<td></td>
<td>..</td>
<td>..</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>4.3</td>
<td>.879</td>
<td>poliomyelitis and post polio</td>
<td>Poliomyelitis</td>
<td>Polio_vaccine</td>
<td>Jonas_Salk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>..</td>
<td>..</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>3.3</td>
<td>.748</td>
<td>argentine british relations</td>
<td>Foreign_relations_of_Argentina</td>
<td>Argentina_national_rugby_team</td>
<td>Falklands_War</td>
</tr>
<tr>
<td></td>
<td></td>
<td>..</td>
<td>..</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>1.9</td>
<td>.966</td>
<td>agoraphobia</td>
<td>Charles_M._Schulz</td>
<td>Snoopy</td>
<td>UGM-27_Polaris</td>
</tr>
</tbody>
</table>
Combining features improves entity ranking significantly

Feature combination sign. better than best single feature (+11.3% NDCG@10)
Text & KB features both contribute significantly to the ranking

Feature ablation study (leave one-out) on REWQ Robust04

Doc Mention Frq and Wikipedia KB most important performance drivers
CONCLUSIONS
We contributed an in-depth study of doc-based entity ranking

Our contributions and major findings

a) We created two new gold standards for query-specific entity ranking

b) Finding 1: Combing document retrieval with entity retrieval is helpful (high overall NDCG/MAP)

c) Finding 2: Two simple, but complementary feature are already helpful
   a) Document-based: Mention frequency in query-specific documents (MenFrqIdf)
   b) Knowledge-based: Full-text SDM on Wikipedia (WikiSDM)
We contributed an in-depth study of doc-based entity ranking

Our contributions and major findings

a) We created two new gold standards for query-specific entity ranking

b) Finding 1: Combing document retrieval with entity retrieval is helpful (high overall NDCG/MAP)

c) Finding 2: Two simple, but complementary feature are already helpful
   a) Document-based: Mention frequency in query-specific documents (MenFrqIdf)
   b) Knowledge-based: Full-text SDM on Wikipedia (WikiSDM)

REWQ Ground truth datasets available
http://rewq.dwslab.de