Entity Linking: Test Collections Revisited
Laurent Mertens, Thomas Demeester, Johannes Deleu, Matthias Feys, Chris Develder
Overview

- Introduction
- Query Selection Strategies
- Influence of Query Set Size
- Conclusion
Introduction
Introduction: Task Description

- What is Entity Linking?

President Obama promised a sustained effort to build trust between police and the communities they serve during a meeting at the White House on Monday.
Introduction: System Evaluation

- How to evaluate the performance of an EL system?
  - Need “Golden Truth”: an annotated query set
- Query?
  - A particular mention in a particular document
- Annotated?
  - You know the correct answer

“Mention Washington in Document X refers to Wikipedia article Washington, D.C.”
Query Selection Strategies
Generating Random Queries

- 1: Random Document
  - QueryID: EL0192
  - DocID: doc_xxx
  - Mention: “Obama”

- 2: Random Mention
  - QueryID: EL0192
  - DocID: doc_xxx
  - Mention: “Obama”
Random Query Results

<table>
<thead>
<tr>
<th></th>
<th>RD</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td>753 (54%)</td>
<td>169 (18%)</td>
</tr>
<tr>
<td>NIL</td>
<td>654 (46%)</td>
<td>777 (82%)</td>
</tr>
<tr>
<td>Total</td>
<td>1407</td>
<td>946</td>
</tr>
</tbody>
</table>

Number of queries per selection strategy

<table>
<thead>
<tr>
<th></th>
<th>RD</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Link</td>
<td>89.1%</td>
<td>80.5%</td>
</tr>
<tr>
<td>NIL</td>
<td>83.2%</td>
<td>92.5%</td>
</tr>
<tr>
<td>Total</td>
<td>87.0%</td>
<td>86.1%</td>
</tr>
</tbody>
</table>

Precision and Recall on RD and RM queries
Influence of Query Set Size
Motivation

F1 for increasing query set size for top scoring TAC2013 systems

Chaos!  Order?
Theoretical Model

- **Question:** when removing $x$ queries from a given set, what are the maximum, minimum and expected performance on the reduced set?

- **Answer:** model probability of removing a query with specific properties
  - Is it a NIL or a link?
  - Was it answered correctly by the system or not?

  !When a system predicts NIL for NIL, it is always right, but when it predicts a link for a link, it might predict a wrong link!
Theoretical Model: Verification

System vs. Theory: F1 distribution over 10,000 iterations after removal of x% queries
## Theoretical Model: Application

<table>
<thead>
<tr>
<th>X = 10</th>
<th>Min. F1</th>
<th>Max. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>0.875</td>
<td>0.877</td>
</tr>
<tr>
<td>Set 2</td>
<td>0.874</td>
<td>0.882</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X = 50</th>
<th>Min. F1</th>
<th>Max. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>0.873</td>
<td>0.879</td>
</tr>
<tr>
<td>Set 2</td>
<td>0.869</td>
<td>0.888</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X = 100</th>
<th>Min. F1</th>
<th>Max. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>0.872</td>
<td>0.881</td>
</tr>
<tr>
<td>Set 2</td>
<td>0.865</td>
<td>0.893</td>
</tr>
</tbody>
</table>

Min and max observed F1 after removal of X queries for query set sizes 2853 (Set 1) and 1000 (Set 2)
System Comparison: Concept

- **Question**: does changing the query set size influence the ordering of the systems?
- **Proposed answer**: perform following experiment

\[ \text{SysT} = \text{Top scoring system for TAC2013} \]
\[ \text{For increasing Query Set size:} \]
\[ \quad \text{For all best-submissions from other teams, check if SysT is significantly better according to F1 value (} p = 0.05 \text{ over 1000 bootstraps)} \]
System Comparison: Result

Number of systems not discernible from top scoring system
Conclusion
Conclusion

- Introduced 2 ways of automatically selecting query mentions with different focus:
  - Random Document: popular, easy to resolve
  - Random Mention: “long-tail”, difficult to resolve
- Demonstrated the size of a query set has a severe impact on system assessment
- Investigated effect of query set size on its discriminating power: more queries is better!
...questions...