Evaluation Initiatives for Entity-oriented Search

Krisztian Balog
University of Stavanger

Evaluation Initiatives for Entity-oriented Search

Krisztian Balog
University of Stavanger

KEYSTONE Spring WG Meeting 2015 | Kosice, Slovakia, 2015

Evaluation Initiatives for Entity-oriented Search

Krisztian Balog
University of Stavanger

In this talk

- Keyword-based search in knowledge bases
- Query understanding with the help of knowledge bases
- Test collections, metrics, evaluation campaigns

Share tasks at world-wide evaluation campaigns (TREC, CLEF)

Provide ingredients for evaluating a given task
- Data collection
- Gold-standard annotations
- Evaluation metrics

Distribution of web search queries (Pound et al., 2010)

Distribution of web search queries (Lin et al., 2011)

Distribution of web search queries


Entity Retrieval

Ad-hoc entity retrieval
- Input: keyword query
  - "telegraphic" queries (neither well-formed nor grammatically correct sentences or questions)
  - E.g., semantic web curl, "American embassy in Nairobi"
- Output: ranked list of entities
- Collection: semi-structured documents

Evaluation initiatives
- INEX Entity Ranking track (2007-09)
- TREC Entity track (2009-11)
- Semantic Search Challenge (2010-11)
- INEX Linked Data track (2012-13)

Baseline models
- Standard document retrieval methods applied on entity description documents
- E.g., language modeling (MLM)

Baseline models
- Fielded extensions of document retrieval methods
- E.g., mixture of language models (MLM)
### Setting field weights

- **Heuristically**
  - Proportional to the length of text content in that field, to the field’s individual performance, etc.
- **Empirically** (using training queries)
- **Problems**
  - Number of possible fields is huge
  - It is not possible to optimise their weights directly
  - Entities are sparse w.r.t. different fields
  - Most entities have only a handful of predicates

### Predicate folding

- **Idea**: reduce the number of fields by grouping them together based on, e.g., type

### Probabilistic Retrieval Model for Semistructured data (PRMS)

- **Extension to the Mixture of Language Models**
- **Find which document field each query term may be associated with**

\[
P(t|\theta_d) = \sum_{j=1}^{m} P(d_j|t)P(t|\theta_{d_j})
\]

- Term probability
- Mapping probability estimated for each query term

### Estimating the mapping probability

\[
P(t|C_j) = \frac{\sum_{d_j} P(t|d_j)P(d_j)}{\sum_{d_k} P(t|d_k)P(d_k)}
\]

- Prior field probability
- Term likelihood
- Probability of a query term occurring in a given field type

### Example (IMDB)

- **meg ryan war**

### Baseline results

- **MAP**
- **Location** statistics

### Related entity finding

- **TREC Entity track (2009-2011)**
- **Input**: natural language query
- **Output**: ranked list of entities
- **Collection**: combination of unstructured (web crawl) and structured (DBpedia)
- **Queries** contain an entity (E), target type (T), and a required relation (R)
- Entity and target type are annotated in the query

### Examples

- **airlines that currently use Boeing 747 planes**
- **Members of The Beaux Arts Trio**
- **What countries does Eurail operate in?**
Modeling related entity finding

- Exploiting the available annotations in a three-component model

\[ p(e|E,T,R) \propto p(e|E) \cdot p(T|e) \cdot p(R,e) \]

** Wikification **

- Recognizing concepts in text and linking them to the corresponding entries in Wikipedia

2. Understanding Queries

Entity linking

- Typically only named entities are annotated
- Reference KB can be different from Wikipedia
- Usages
  - Improved retrieval, enabling semantic search
  - Advanced UI/UX, help users to explore
  - Knowledge base population/acceleration

Entity linking methods

- Mention detection
  - Identifying entity mentions in text
- Candidate entity ranking
  - Generating a set of candidate entries from the KB for each mention
- Disambiguation
  - Selecting the best entity for a mention
  - Machine learning; features: commonness, relatedness, context, …

Entity linking evaluation

<table>
<thead>
<tr>
<th>ground truth ( \hat{A} )</th>
<th>system annotation ( \hat{A} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</td>
<td>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</td>
</tr>
</tbody>
</table>

Examples

- Can we obtain such annotations automatically?

airlines that currently use Boeing 747 planes

Members of The Beaux Arts Trio

What countries does Eurail operate in?

Entity linking evaluation

<table>
<thead>
<tr>
<th>ground truth ( \hat{A} )</th>
<th>system annotation ( \hat{A} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</td>
<td>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</td>
</tr>
</tbody>
</table>
**Entity linking evaluation**

**ground truth** $\hat{A}$

Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.

**system annotation** $A$

Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.

$$P = \frac{|A \cap \hat{A}|}{|A|} \quad \quad R = \frac{|\hat{A} \cap A|}{|\hat{A}|} \quad \quad F = \frac{2 \cdot P \cdot R}{P + R}$$

**Entity linking for queries**

- **Challenges**
  - search queries are short
  - limited context
  - lack of proper grammar, spelling
  - multiple interpretations
  - needs to be fast

**Example**

the governor

**Example**

the governor movie

**Example**

new york pizza manhattan

**Example**

new york pizza manhattan

**ERD’14 challenge**

- **Task**: finding query interpretations
- **Input**: keyword query
- **Output**: sets of sets of entities
- **Reference KB**: Freebase
- Annotations are to be performed by a web service within a given time limit

**Evaluation**

**ground truth** $\hat{I}$

New York City, Manhattan

**system annotation** $I$

New York City, Manhattan

New York-style pizza, Manhattan

New York-style pizza

$$P = \frac{|I \cap \hat{I}|}{|I|} \quad \quad R = \frac{|\hat{I} \cap I|}{|\hat{I}|} \quad \quad F = \frac{2 \cdot P \cdot R}{P + R}$$
ERD’14 results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>F1</th>
<th>latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SMAPH Team</td>
<td>0.7076</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>NTUNLP</td>
<td>0.6797</td>
<td>1.04</td>
</tr>
<tr>
<td>3</td>
<td>Seznam Research</td>
<td>0.6693</td>
<td>3.91</td>
</tr>
</tbody>
</table>

Single interpretation is returned

3

Living Labs for IR Evaluation

Goals & focus

- **Overall goal**: make information retrieval evaluation more realistic
  - Evaluate retrieval methods in a live setting with real users in their natural task environments
- **Focus**: medium-sized organizations with fair amount of search volume
  - Typically lack their own R&D department, but would gain much from improved approaches

What is in it for participants?

- Access to privileged commercial data
  - (Search and click-through data)
- Opportunity to test IR systems with real, unsuspecting users in a live setting
  - (Not the same as crowdsourcing!)

"Give us your ranking, we’ll have it clicked!"

Research aims

- Understanding of online evaluation and of the generalization of retrieval techniques across different use cases
- Specific research questions
  - Are system rankings different when using historical clicks from those using online experiments?
  - Are system rankings different when using manual relevance assessments (“expert judgments”) from those using online experiments?

Key idea

- Focus on frequent (head) queries
  - Enough traffic on them (both real-time and historical)
  - Ranked result lists can be generated offline
- An API orchestrates all data exchange between live sites and experimental systems

Methodology

- Queries, candidate documents, historical search and click data made available
- Rankings are generated for each query and uploaded through an API
- When any of the test queries is fired, the live site request rankings from the API and interleaves them with that of the production system
- Participants get detailed feedback on user interactions (clicks)
- Ultimate measure is the number of “wins” against the production system

Interleaving

System A (production system)
- Result A1
- Result A2
- Result A3

System B (experimental system)
- Result B1
- Result B2
- Result B3

Partners & Use-cases

<table>
<thead>
<tr>
<th>Provider</th>
<th>Product search</th>
<th>Local domain search</th>
<th>Web search</th>
</tr>
</thead>
<tbody>
<tr>
<td>regiojatek.hu</td>
<td>raw queries and documents</td>
<td>raw queries and documents</td>
<td>pre-computed document-query features</td>
</tr>
<tr>
<td>uva.nl</td>
<td>raw queries and (highly structured) documents</td>
<td>raw queries and (generally textual) documents</td>
<td></td>
</tr>
<tr>
<td>seznam.cz</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data

- raw queries and (highly structured) documents
- raw queries and (generally textual) documents
- pre-computed document-query features

Site traffic

- relatively low (~4K sessions/day)
- relatively low
- high

Info needs

- (mostly) transactional
- (mostly) navigational
- vary

Logs

- timestamp: 1402052038;7902b47fbbd45360a672670367cccf8d;IDX;28178,28188,33533,36450,32394,34188,76395,85649,33991,27747,30649,34561,00841,33039,34543;event: 1;event attributes: 1,76770,07921,80075,36149,85739,09726,50463,06679,77285,56687,80878,00101,03984,77712
- timestamp: 1402052040;fd593d671402dd5ea21e39d00aa30e26;PRD;16402;event: 14420
- timestamp: 1402052040;fd593d671402dd5ea21e39d00aa30e26;PRC;16402;1;event attributes: 80875,36419,65736,09726,85683,09679,13867,25718,13866,25714,36152,92991
- timestamp: 1402052041;fd593d671402dd5ea21e39d00aa30e26;IDX;34188,34561,34543,30649,27747,28188,28178,76395,85649,00841,33039,33991,33533,32394,36450;event: 14595
- timestamp: 1402052043;44ee8beae78426b0eb5662d25191984c;PRD;71983;event: 14420
- timestamp: 1402052043;44ee8beae78426b0eb5662d25191984c;PRC;71983;1;event attributes: 88986,73681,49141,02215,85632,85633,40628,40497,71970,63293,79479,40942
- timestamp: 1402052043;897deef88cfd77d59cf2eebe6c9cab43;CAT;12;9;event attributes: 97975,07976,76715,99293,88844,64632,64528,64631,64708,76800,64530,29804,14946,21163
- timestamp: 1402052044;dc1db4bef11f027a0a8d161b8354a318;IDX;00841,34543,76395,28188,34188,33039,33533,33991,30649,36450,27747,32394,85649,34561,33039;event: 14420
- timestamp: 1402052045;bb3a12f50e5c1a53765885a90a8d8992;IDX;00841,28188,27747,76395,30649,28178,34188,36450,33991,34561,33533,32394,85649,34543,33039;event: 14420

Relational database

- products
  - product_id
  - product_name
  - price
  - manufacturer_id
  - description
  - short_description
  - age_min
  - age_max
- product_characters
  - product_id
  - character_id
  - name
- product_categories
  - product_id
  - category_id
  - name
- categories
  - category_id
  - name
  - parent
  - priority
  - description
- manufacturers
  - manufacturer_id
  - name

Guide for CLEF participants

1. Guide for CLEF Participants

1.1 Schedule

Product search

- Ad-hoc retrieval over a product catalog
- Several thousand products
- Limited amount of text, lots of structure
- Categories, characters, brands, etc.
Test queries

Product data

Results

Summary

Questions?

Join us!

- Evaluation will continue to run even after the CLEF deadline
- Additional use-cases are welcome

<table>
<thead>
<tr>
<th>Method</th>
<th>Impr. total</th>
<th>Impr. per query</th>
<th>#clicks</th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1884</td>
<td>18.84</td>
<td>581</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>1407</td>
<td>14.07</td>
<td>630</td>
<td>0.45</td>
</tr>
<tr>
<td>3</td>
<td>1334</td>
<td>13.34</td>
<td>483</td>
<td>0.36</td>
</tr>
</tbody>
</table>

- Entities as the unit of retrieval
- Understanding queries with the help of entities
- From traditional test-collection based evaluations to evaluation as a service

Contact  | @krisztianbalog | krisztianbalog.com