Track aims

- Evaluate entity-oriented search tasks on the Web
- Address information needs that are better answered by returning objects (instead of just any type of documents)

What is an entity?

An entity has...

- A unique identifier
- Name(s)
- Type
- Attributes
- Relationships to other entities

Related Entity Finding

- Return a ranked list of entities (of a specified type) that engage in a given relationship with a given source entity

Example entity-oriented information need

airlines that currently use Boeing 747 planes
Why REF?

- Grounded in practical applications
- More than 50% of Web search queries target entities [Pound et al., WWW’10]
- Challenging
  - These information needs are difficult to answer
  - Existing approaches are not directly applicable
- Meets pragmatic constraints
  - Results sets are of reasonable size

The Related Entity Finding (REF) task

- Given
  - Input entity, defined by its name and homepage
  - Target type (person, organization, or product)
  - Narrative (describing the nature of the relation in free text)
- Return
  - Ranked list of homepages of related entities

Example

Information need:

airlines that currently use Boeing-747 planes

TREC topic definition:

<narrative>airlines that currently use Boeing-747 planes</narrative>
Example

Information need:
airlines that currently use Boeing-747 planes

TREC topic definition:
<narrative>airlines that currently use Boeing-747 planes</narrative>
<entity_name>Boeing 747</entity_name>
<entity_URL>clueweb09-en0005-75-02292</entity_URL>

Example

Information need:
airlines that currently use Boeing-747 planes

TREC topic definition:
<narrative>airlines ... planes</narrative>
<entity_name>Boeing 747</entity_name>
<entity_URL>clueweb09-en0005-75-02292</entity_URL>

Example

Topics, data collection
• 20 topics
• ClueWeb09 “Category B”
• 50 million English webpages
• Topic development is difficult
• Not many entities have homepage in CatB
• What is a homepage, anyway?
• Trivial topics

Example topics
• Organizations that award Nobel prizes
• CDs released by the King’s Singers
• Companies that John Hennessey serves on the board of
• Universities that are members of the SEC conference for football
• Scotch whisky distilleries on the island of Islay

Relevance assessments
Run 1: 2 2
Run 2: 1 0
Run N: 0 1

Required output
• Homepages of the entity (up to 3)
• Wikipedia page of the entity
• String answer that represents the entity
• Supporting documents (up to 10)
Relevance assessments

Evaluation metrics

- **NDCG@R** Normalized discounted cumulative gain at rank R
  - R is the number of primaries and relevant pages for the topic
- **P@10** Fraction of primary pages in the top 10 results
- **#rel** Number of relevant pages
- **#pri** Number of primary pages

Variations on computing graded relevance

<table>
<thead>
<tr>
<th>Homepages</th>
<th>Wikipedia page</th>
<th>Name</th>
<th>Max gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP-only</td>
<td>+</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>HP+Name</td>
<td>+</td>
<td>+</td>
<td>3</td>
</tr>
<tr>
<td>WP-only</td>
<td>+</td>
<td>+</td>
<td>2</td>
</tr>
<tr>
<td>HP+WP</td>
<td>+</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>HP+WP+Name</td>
<td>+</td>
<td>+</td>
<td>3</td>
</tr>
</tbody>
</table>

Observations

- ClueWeb B dataset doesn’t contain enough entity pages
- Heavy use of Wikipedia/DBpedia
- Need more precise definitions of what constitutes an entity homepage
- Primary homepages are not rewarded enough
- Simplify entity record - traditional single-record TREC format is preferred

Changes in 2010

- ClueWeb English (500 million pages)
- New entity type: location
- Single record submission format
- No supporting documents
- Wikipedia pages are not accepted
- Primary homepages are rewarded more

Entity homepage

<table>
<thead>
<tr>
<th>Homepage</th>
<th>Primary:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>devoted to and in control of the entity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not a homepage</th>
<th>Relevant:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>devoted to the entity, but not in control of the entity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Wikipedia page of the entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>News articles and blog posts, even if exclusively about the entity</td>
<td></td>
</tr>
<tr>
<td>Topics, assessment</td>
<td>Metrics</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------</td>
</tr>
<tr>
<td>• 20 topics from 2009 reassessed</td>
<td></td>
</tr>
<tr>
<td>• 50 new topics</td>
<td></td>
</tr>
<tr>
<td>• 47 of which are judged (pooled to depth 20)</td>
<td></td>
</tr>
<tr>
<td>• Main metric: <strong>NDCG@R</strong></td>
<td></td>
</tr>
<tr>
<td>• Gain=1 for relevant</td>
<td></td>
</tr>
<tr>
<td>• Gain=3 for primary</td>
<td></td>
</tr>
<tr>
<td>• RPrec and MAP for both relevance levels</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entity 2010</th>
<th>Entity 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A variation: Entity List Completion</th>
<th>Billion Triple Collection (BTC-2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Find entities that are engaged in a specific relation with an input entity</td>
<td></td>
</tr>
<tr>
<td>• Entities are represented by a unique URI in the Linked Open Data (LOD) cloud</td>
<td></td>
</tr>
<tr>
<td>• Topic definition includes example entities</td>
<td></td>
</tr>
<tr>
<td>• Sample of the Linked Open Data cloud</td>
<td></td>
</tr>
<tr>
<td>• Semantic Web crawl</td>
<td></td>
</tr>
<tr>
<td>• ~1.14 billion statements</td>
<td></td>
</tr>
<tr>
<td>• 17GB gzipped</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topics</th>
<th>Example topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 13 out of the 20 REF 2009 topics</td>
<td></td>
</tr>
<tr>
<td>• Excluded topics with complete result sets and topics without example entities</td>
<td></td>
</tr>
<tr>
<td>• Mapped type to the most specific class within the DBpedia ontology</td>
<td></td>
</tr>
</tbody>
</table>
| • Mapped known relevant entity HPs to their URI in BTC-2009 | <!-- On top of the REF topic definition -->
| <entity_URIs> |
| <URI>http://dbpedia.org/resource/Boeing_747</URI> |
| <target_type_dbpedia>dbpedia-owl:Airline</target_type_dbpedia> |
| <examples> |
| <entity> |
| <URI>http://dbpedia.org/resource/Northwest_Airlines</URI> |
| <entity> |
| <URI>http://dbpedia.org/resource/British_Airways</URI> |
| <URI>http://twitter.com/British_Airways</URI> |
| ... |
| </examples> |
Observations

- Found entities not covered by ClueWeb
- Still many "missing" entities
  - Perhaps the collection is too small?

Entity 2011

Task, data set, topics

- Given an input entity, target type, and narrative (nature of the relation in free text), return a ranked list of homepages of related entities.
- 50 new REF topics
- ClueWeb09 English

Key changes

- Only primary homepages are accepted, relevance is binary
- For each answer a single supporting document is required
- Target type is not limited
- Groups that use Web Search Engines are required to submit an obligatory run
  - Using the Lemur ClueWeb online query service

Example topic

<narrative>airlines that fly the Airbus 320 plane</narrative>

<entity_name>Airbus</entity_name>
<entity_URL>clueweb09-en0004-69-31337</entity_URL>
<target_entity>airline</target_entity>

Results: best run per group

<table>
<thead>
<tr>
<th>runID</th>
<th>type</th>
<th>MAP</th>
<th>R-Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRISREF1</td>
<td>M</td>
<td>0.2509</td>
<td>0.2908</td>
</tr>
<tr>
<td>TongKeyEN2</td>
<td>M</td>
<td>0.1266</td>
<td>0.1984</td>
</tr>
<tr>
<td>WhuRun1</td>
<td>M</td>
<td>0.0063</td>
<td>0.0176</td>
</tr>
<tr>
<td>ICSTmaxSni</td>
<td>A</td>
<td>0.0004</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Results: all runs

- MAP
- R-Prec
**REF 2011 Assessments**

- Runs pooled down to depth 30
- \(<docid, namestring>\) pairs (instead of \(<docid, namestring, supportdoc>\) triples) were judged
- Two judgment files

**REF 2011 Standard trec_eval qrels file**

<table>
<thead>
<tr>
<th>topic</th>
<th>docid</th>
<th>judgment</th>
</tr>
</thead>
</table>

Two sources
- Pooled results
- Answers found by assessors during topic development
- Supporting documents are not taken into account

**REF 2011 Judgment file participants-only**

<table>
<thead>
<tr>
<th>topic</th>
<th>docid</th>
<th>supportdoc</th>
<th>judgment</th>
<th>namestring</th>
<th>nameclass</th>
<th>namecorrectness</th>
</tr>
</thead>
</table>

Judgment \(<topic, docid, supportdoc>\)
- 0: incorrect
- 1: correct, supported
- -1: correct, not supported

Support document is arbitrarily chosen among all submitted for a \(<docid, namestring>\) pair

**Entity 2011**

Task Examples
- \(<docid, namestring>\)
  - 0: incorrect
  - 1: inexact
  - 2: correct

Support document is arbitrarily chosen among all submitted for a \(<docid, namestring>\) pair

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
<th>Entity ID</th>
<th>ClueWeb09</th>
<th>Sindice'11</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF 2011</td>
<td>N</td>
<td>URL</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>REF-LOD 2011</td>
<td>N</td>
<td>URI</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>ELC 2011</td>
<td>Y</td>
<td>URI (URL)</td>
<td>Opt.</td>
<td>Y</td>
</tr>
</tbody>
</table>

**REF-LOD 2011**

- Same as the REF task, BUT
- Entities are identified by their URIs in the LOD crawl (not by their homepages)
- Supporting documents are optional

**Task Examples**

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
<th>Entity ID</th>
<th>ClueWeb09</th>
<th>Sindice'11</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF 2009</td>
<td>N</td>
<td>URL</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>REF 2010</td>
<td>N</td>
<td>URL</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>ELC 2010</td>
<td>Y</td>
<td>URI (URL)</td>
<td>Opt.</td>
<td>Y</td>
</tr>
</tbody>
</table>
REF-LOD 2011 Results

Entity 2011

ELC 2011

Task

• In addition to source entity, type, and narrative, examples are also given
• Return URIs of relevant entities

Key changes

• Larger and more representative LOD crawl
• Examples are not mapped manually to LOD
  • Examples may also be returned, but are worth less credits

Data set and topics

• 50 REF topics from 2010 turned into ELC topics
• Target type mapped to the DBpedia ontology
• Known relevant entities are provided as examples
• Sindice-2011 dataset

Sindice-2011 dataset

http://data.sindice.com/trec2011/

• Dump of Sindice search engine
• 11 billion RDF statements from 230 million documents
• 1.7 billion entities
• All standardised data formats (RDFa, Microformats, etc.) are covered
## Sindice-2011 dataset

http://data.sindice.com/trec2011/

- Structured in two ways
  - Organized around documents
  - Organized around entities
- Tool support based on SIREn (Lucene extension) to index and search the collection

## ELC 2011 Results and assessments

- 21 runs by 7 groups
- Community judgments
- Underway...

## Summary

- Task(s) generated interest
- 120 REF topics, 70 ELC topics
  - Need to consolidate qrels, evaluation scripts, etc.
- Strong end-to-end focus seems to be an issue
  - Difficult to untangle the performance contributions of individual components

## The future

- Join mailing list
  http://groups.google.com/group/trec-entity
- LinkedIn discussion group
  "Entity-Oriented Search"
- Entity plenary session on Friday

## Acknowledgments

- NIST
  - Ellen Voorhees
  - Ian Soboroff
  - Assessors
- Sindice team, DERI
- Participants

## Questions?