Towards Summarizing the Web of Entities

August 15, 2012
Thomas Hofmann
Director of Engineering
Search Ads Quality
Zurich, Google Switzerland
thofmann@google.com

contributors:
Enrique Alfonseca
Yasemin Altun
Massimiliano Ciaramitria
Jean-Yves Delort
Katja Filippova
Ioannis Tsochantaridis
1. The Big Picture
2. World of Entities
3. Web of Entities
4. Event Detection
5. Event Summarization
6. Timeline View
1

The Big Picture
How to Summarize "the Web"?

- Web = content creation "in the wild".
  - Multi-authored, multi-lingual, absence of editorial oversight, huge diversity (quality, style, intent)

- How can we summarize content on the Web? And what for?

Separate **what is talked about** (referred to) from **what is said** (propositional content)
Annotate all Web content by references to a known "set of things" = **entities** (across languages, styles)

Create an **entity index of the Web**

Cluster entity mentions around common **events** or based on their **propositional content**

Entity-centric **summarization** around memorable events and **aggregation** of "what is said"
This Talk!

- Not primarily about advances in academic research
- Present an engineering-centric view: long-term vision plus simplified end-to-end prototype system that operates on Web-scale

- Need to cut many corners...
- Leave a lot of work half-finished or undone (UGC, unsupervised parsing, relation extraction, ....)
- Identify & contribute to complex problems on the way. Shift away from standard multi-document summarization...

Enrique Alfonseca
2

World of Entities
Curated knowledge base with (~05/2011):

- **22 M nodes** (concepts, named entities, locs, ...)
- **400 M connections** (typed relations)
- Community-driven effort; larger internal base
MQL

MQL: query language to query entity graph

Schema part of graph representation
cf. Google I/O talk
Entity-centric knowledge base is only one side of the effort to establish a "semantic Web"

Manual annotation (by authors, Web masters, publishers; see schema.org) requires large collaborative effort (slow, conflicting interests, lack of comprehensiveness,...)

Enable automatic understanding of text by identifying mentions of entities & concepts.

Grounding of mentions/references in curated 'ontology' of things (vs. open domain extraction)
Entities in Web Search

query 2 entity mapping
Web of Entities
Entity Grounding

Step 1
Identify entity candidates (names)

Freebase mids
Entity Grounding

Step 2
Load known entity relations and create entity graph for document
Entity Grounding

Step 3

Joint inference over entity graph to disambiguate references

Inferred entities
Reference Disambiguation Task

1. Observe a set of mentions
2. Produce list of scored candidate entities using a 'local' model
3. Rerank candidates using topic model
4. Combine the scores, and search the space for the best set of entities

<table>
<thead>
<tr>
<th>CoNLL Test</th>
<th>Hoffart et al 2011</th>
<th>LogisReg</th>
<th>Reranked LogisReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>81.82</td>
<td>87.67</td>
<td>89.44</td>
</tr>
<tr>
<td>Recall</td>
<td>~70.00</td>
<td>64.83</td>
<td>66.41</td>
</tr>
<tr>
<td>F-1</td>
<td>75.45</td>
<td>74.54</td>
<td>76.22</td>
</tr>
</tbody>
</table>

Generative prior model for sets of entities

CoNLL data annotated by Weikum group @MPI
4 News Event Detection
Towards Entity Timelines

Events influence what the Web says about entities. Temporality matters, so use it as an organizing principle!

Given an entity and a time period, provide a summary of the most memorable events involving this entity.

Current "strawman" system:
- Identify relevant news stories/clusters about the entity.
- Filter and rank news clusters in order of importance.
- Concisely summarize each news cluster
- Visualize summaries in temporal order
Web Search is not enough ...

Have to use left navigation restrict to see useful recent results
Web News is not enough ...

Confusion around entities in Google News
Even with a little help, no good results, since only today's news.

Web News is not enough ...
Entity-Centric View on News

**Entity Timeline**

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 25, 2011</td>
<td>REM's farewell marks the end of an era</td>
</tr>
<tr>
<td>September 21, 2011</td>
<td>REM breaks up after 31 years, band says 'our deepest thanks for listening'</td>
</tr>
<tr>
<td>March 6, 2011</td>
<td>REM's New Album Collapse Into Now Released Today on Warner Bros. Records</td>
</tr>
<tr>
<td>March 2, 2011</td>
<td>REM's New Album, &quot;Collapse Into Now,&quot; Leaks Online</td>
</tr>
<tr>
<td>March 18, 2010</td>
<td>Alex Chilton's Big Star bandmates confirm SXSW tribute show featuring REM</td>
</tr>
<tr>
<td>January 21, 2010</td>
<td>REM Song To Be Recorded As Simon Cowell's Charity Single</td>
</tr>
</tbody>
</table>

R.E.M. was an American rock band from Athens, Georgia, formed in 1980 by singer Michael Stipe, guitarist Peter Buck, bassist Mike Mills and drummer Bill Berry. One of the first popular alternative rock bands, R.E.M.
Event Clustering & Cluster Selection

- Restrict to Google news corpus + entities
- Piggyback on clustering of stories (hack)
- Given an entity, score news clusters
- [ A lot more work on event detection required... ]

1. Event prominence. The size of the news cluster (in how many articles the news was published?).

2. Entity topicality. The average topicality of the entity in the articles in the cluster (derived from mentions)

3. Entity prominence. The percentage of titles in the cluster mentioning the entity.
Event Summarization
Given a set of news articles, generate an extractive summary by:

- "Extractive" multi-document summarization, plus:
- Use entity as focus (vs. generic summary)
- Headline selection for timeline view
- Sentence compression

State-of-the-art results on academic corpora (TAC 2011)
Multi-document Summarization

TopicSum
Haghighi and Vanderwende

Jean-Yves Delort
Enrique Alfonseca

- Augmented for update summarization
- Integrate paraphrase scoring from MT
- Entity-specific sentence weighting
- Special attention to titles

Ioannis Tsochanta
ridis
Sentence Compression

Beyond extractive summarization: compression

Learn from title / first sentence pairs.
Valerie Amos said on Thursday that the U.N. humanitarian chief was devastated by the destruction and wants to know what happened in district Baba Amr of the city of Homs as result of an assault by government forces.
Time Line View
Entity Timeline Prototype

Google

Entity Timeline

Any time
Past 3 months
Past 6 months
Past year

Chronology
Annotated timeline

Top 10 events
Top 20 events
Top 50 events
Top 100 events

REM

Select an item from the list:

R.E.M. Film music contributor
Michael Stipe Celebrity
Rem Koolhaas Person
Rem City/Town/Village
Rem Musical Recording
Rem Mythological Figure
Rem Person
Rem Topic
REM rebound
R.E.M. Musical Album
R.E.M. Live Musical Album

R.E.M.
Musical Group: Alternative rock, College rock, Jangle pop
Place Musical Career Began: Athens
Active as Musical Artist (start): 1980

R.E.M. was an American rock band from Athens, Georgia, formed in 1980 by singer Michael Stipe, guitarist Peter Buck, bassist Mike Mills and drummer Bill Berry. One of the first popular alternative rock bands, R.E.M. gained early attention due to Buck...

March 2, 2011
REM's New Album, "Collapse Into Now," Leaks Online

March 18, 2010
Alex Chilton's Big Star bandmates confirm SXSW tribute show featuring REM

January 21, 2010
REM Song To Be Recorded As Simon Cowell's Charity Single
Some screenshots ....
Fernando Alonso roars from 11th to win phenomenal race

He now takes his 20-point lead in the championship to the British Grand Prix at Silverstone next month, hoping to emulate his victory in the 2011 race.

Alonso called the win in front of his home fans one of the best of his 29-win career.

Vettel's exit left Alonso being chased by Lotus' Romain Grosjean, but remarkably another alternator issue forced the Frenchman to retire on lap 41.
Johnny Depp splits from Vanessa Paradis after 14 years together

May 8, 2012

Johnny Depp plays Barnabas Collins in "Dark Shadows"

Johnny Depp has described his performance as Barnabas in Dark Shadows as an homage to original star Jonathan Frid.

Depp plays vampire Barnabas Collins, who is released in 1972 after being confined in a coffin for nearly two centuries by a vengeful witch.

"Dark Shadows," based on the TV soap opera of the late 1960s and early 1970s, is Johnny Depp's eighth collaboration with director Tim Burton.

Johnny Depp channels departed gonzo journalist Hunter S. Thompson again for 'Rum Diary'

May 14, 2011

Depp, Piratas 4 now Estimated $88 Million weekend with Top Opening Night of Year
François Hollande

July 1, 2012
France's budget needs more tightening, audit says, despite Hollande's anti-austerity message

June 26, 2012
Hollande, markets welcome eurozone crisis deal but French debt rises

June 20, 2012
François Hollande's industry goal meets Peugeot town reality

June 14, 2012
Hollande's Socialists triumph in French Parliamentary election giving president free rein to implement anti-austerity measures

June 14, 2012
Hollande allies take reins of power

June 7, 2012
French first lady takes virtual swipe at Hollande's ex

June 6, 2012
Eurozone crisis tests Hollande's credentials

June 4, 2012
Hollande set to revoke 2010 pension reform
Where from Here?
Future Directions

- More serious event detection & extraction. Tying events to entities and their relations.
- Include UGC, e.g. by mapping/linking to new stories
- Improve summarizer by better understanding of propositional content
- Developing a full-blown time-line view
THANK YOU!

I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions.