Contextualizing Trending Entities in News Stories

Marco Ponza, Diego Ceccarelli, Edgar Meij, Paolo Ferragina, Sambhav Kothari

The 14th ACM International Conference on Web Search and Data Mining (WSDM 2021)
8-12 March, 2021

Marco Ponza
AI Research Engineer

TechAtBloomberg.com
Introduction
Introduction

▷ Every day an overwhelming amount of data is produced
  ○ Millions of news stories, web pages, and social media posts

▷ Important to provide automatic tools for:
  ○ Identify the most relevant information
  ○ Understand why a particular piece of information was selected

▷ Trends distill units (e.g., keywords, phrases, or names) meaningful for characterizing the news stories content
  ○ Units allow readers to discover and stay focused on relevant information
Introduction

▷ Bloomberg Terminal functionality:

<table>
<thead>
<tr>
<th>Security</th>
<th>Pub.</th>
<th>↓GN</th>
<th>Δ Price</th>
<th>Δ AVAT</th>
<th>News Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba Group Holdings</td>
<td>190</td>
<td></td>
<td>0.00%</td>
<td>+9.83%</td>
<td>Alibaba Group Founder Jack Ma First Appearance</td>
</tr>
<tr>
<td>FactSet Research Systems</td>
<td>112</td>
<td></td>
<td>0.00%</td>
<td>+18.50%</td>
<td>FactSet: 144 01/14/2021</td>
</tr>
<tr>
<td>LG Electronics Inc</td>
<td>56</td>
<td></td>
<td>0.00%</td>
<td>+153.32%</td>
<td>17:12 LG Electronics May Exit Smartphone Business</td>
</tr>
<tr>
<td>MediaTek Inc</td>
<td>56</td>
<td></td>
<td>+0.23%</td>
<td>+40.94%</td>
<td>Like Qualcomm and Apple, MediaTek Designs Chips</td>
</tr>
<tr>
<td>BASF SE</td>
<td>55</td>
<td></td>
<td>+1.24%</td>
<td>+52.13%</td>
<td>BASF Sees FY Adj. Ebit EU3.56 Billion</td>
</tr>
<tr>
<td>Tesla Inc</td>
<td>54</td>
<td></td>
<td>0.00%</td>
<td>-34.58%</td>
<td>Tesla Breaking Into India After China Success</td>
</tr>
<tr>
<td>Burberry Group PLC</td>
<td>53</td>
<td></td>
<td>+4.43%</td>
<td>+887.53%</td>
<td>Burberry Sales Fall</td>
</tr>
<tr>
<td>Taiwan Semiconductor Industry Corp</td>
<td>53</td>
<td></td>
<td>+0.47%</td>
<td>+67.62%</td>
<td>ASML Beats Grapples With Chip Supply Shortage</td>
</tr>
<tr>
<td>Alphabet Inc</td>
<td>50</td>
<td></td>
<td>0.00%</td>
<td>+32.49%</td>
<td>Trump Pardons Former Google Car Engineer</td>
</tr>
<tr>
<td>Microsoft Corp</td>
<td>48</td>
<td></td>
<td>0.00%</td>
<td>-3.99%</td>
<td>Microsoft GM Cars</td>
</tr>
</tbody>
</table>

▷ Can we do one step further and **contextualize** the **trending entities**?
Introduction

▷ Can we do one step further and contextualize the trending entities?

*What do we mean for contextualization?*

- Given a trending entity, we want to retrieve and rank contextual entities
- Contextual entity is an entity that helps explaining why the trending entity is actually trending

▷ Result of such contextualization can have different applications

- Entity-centric summary driven by trending and contextual entities
- Recommend related entities or news stories
- Query expansion
- *And many others!*
Our Contributions

New research problem!

Unsupervised and supervised solutions based on:
1. Personalized PageRank and entity embeddings
2. Feature engineering and learning to rank with improvements ranging from 7% to 12% in terms of Precision@1

Creation and release of a test collection built with crowdsourcing. Available at 
🔗 https://doi.org/10.5281/zenodo.4422044
Problem Formulation

```python
>>> np.random.shuffle(X)
>>> X
```
The CEO of Tesco Dave Lewis resigned and Ken Murphy is supposed to become the new CEO.
Solutions
Two Main Solutions
Unsupervised vs. Supervised

Unsupervised Solution based on Personalized PageRank, Salience, and Embeddings

Supervised Solution based on Feature Engineering and Learn to Rank
Unsupervised Solution

Daily News Stories

Trending Entity
Tesco PLC

Contextual Entities
Market
Ken Murphy
Dave Lewis
ASDA
Supermarket
Hertfordshire
Unsupervised Solution

Schema: All contextual entities linked to the trending entity

Intuition: Graph pivoted around the trending entity

Tesco PLC

- Market
- Ken Murphy
- Dave Lewis
- ASDA
- Supermarket
- Hertfordshire
Unsupervised Solution

Schema: Edges connections draw by stories co-occurrences

Intuition: Co-occurring entities will “vote” each other

- Tesco PLC
- Market
- Ken Murphy
- Dave Lewis
- ASDA
- Supermarket
- Hertfordshire
Unsupervised Solution

Schema: Edges weighted via cosine embeddings generated with Wikipedia2Vec (Yamada, ACL 2016)

Intuition: Using background knowledge for strengthen/weaken the entities edges

Tesco PLC

Market
Ken Murphy
Dave Lewis
ASDA
Supermarket
Hertfordshire

TechAtBloomberg.com
© 2021 Bloomberg Finance L.P. All rights reserved.
Unsupervised Solution

Node Ranking Problem! We can use *Personalized PageRank*!

**Teleport Vector:** Initial relevance score for each node

**Intuition:** Inject in PageRank the information we lost by building the graph (e.g., position, salience, etc.)
Unsupervised Solution

Node Ranking Problem! We can use Personalized PageRank!

Teleport Vector based on Entity Salience

Tesco PLC 0.8
Market 0.05
Ken Murphy 0.6
Dave Lewis 0.7
ASDA 0.3
Supermarket 0.1
Hertfordshire 0.2

TechAtBloomberg.com
© 2021 Bloomberg Finance L.P. All rights reserved.
Unsupervised Solution

Node Ranking Problem! We can use *Personalized PageRank*!

Teleport Vector based on Entity Salience

![Diagram showing node ranking with entities like Tesco PLC, Market, Ken Murphy, Dave Lewis, ASDA, Supermarket, and Hertfordshire, with associated teleport vectors.]
Experimental Results

- Dataset built via crowdsourcing
- 149 trends ~120K contextual entities (~800 per trend)

Publicly available at https://doi.org/10.5281/zenodo.4422044
## Experimental Results

### “Relevant” & “Somewhat Relevant” as Gold Labels

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P@1</th>
<th>P@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.098</td>
<td>0.262</td>
<td>0.224</td>
<td>0.168</td>
<td>0.233</td>
<td>0.448</td>
</tr>
<tr>
<td>Position</td>
<td>0.237</td>
<td>0.195</td>
<td>0.152</td>
<td>0.237</td>
<td>0.319</td>
<td>0.354</td>
</tr>
<tr>
<td>Co-Occurrence</td>
<td>0.359</td>
<td>0.477</td>
<td>0.295</td>
<td>0.441</td>
<td>0.479</td>
<td>0.604</td>
</tr>
<tr>
<td>PMI</td>
<td>0.147</td>
<td>0.161</td>
<td>0.15</td>
<td>0.186</td>
<td>0.209</td>
<td>0.324</td>
</tr>
<tr>
<td>Milne&amp;Witten</td>
<td>0.177</td>
<td>0.141</td>
<td>0.136</td>
<td>0.179</td>
<td>0.242</td>
<td>0.311</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.214</td>
<td>0.248</td>
<td>0.183</td>
<td>0.234</td>
<td>0.276</td>
<td>0.394</td>
</tr>
<tr>
<td>Stories’ Embeddings</td>
<td>0.210</td>
<td>0.208</td>
<td>0.161</td>
<td>0.238</td>
<td>0.287</td>
<td>0.373</td>
</tr>
<tr>
<td>Wikipedia Embeddings</td>
<td>0.206</td>
<td>0.221</td>
<td>0.154</td>
<td>0.214</td>
<td>0.274</td>
<td>0.372</td>
</tr>
<tr>
<td>Reciprocal Rank</td>
<td>0.418</td>
<td>0.523</td>
<td>0.291</td>
<td>0.460</td>
<td>0.508</td>
<td>0.630</td>
</tr>
<tr>
<td>Salience</td>
<td>0.497</td>
<td>0.570</td>
<td>0.394</td>
<td>0.556</td>
<td>0.612</td>
<td>0.727</td>
</tr>
<tr>
<td>PPR</td>
<td>0.519</td>
<td>0.644</td>
<td>0.391</td>
<td>0.586</td>
<td>0.637</td>
<td>0.773</td>
</tr>
</tbody>
</table>

### “Relevant” as Gold Label

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P@1</th>
<th>P@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.097</td>
<td>0.208</td>
<td>0.177</td>
<td>0.179</td>
<td>0.242</td>
<td>0.382</td>
</tr>
<tr>
<td>Position</td>
<td>0.247</td>
<td>0.114</td>
<td>0.105</td>
<td>0.249</td>
<td>0.331</td>
<td>0.274</td>
</tr>
<tr>
<td>Co-Occurrence</td>
<td>0.441</td>
<td>0.416</td>
<td>0.221</td>
<td>0.486</td>
<td>0.515</td>
<td>0.528</td>
</tr>
<tr>
<td>PMI</td>
<td>0.173</td>
<td>0.107</td>
<td>0.105</td>
<td>0.195</td>
<td>0.219</td>
<td>0.252</td>
</tr>
<tr>
<td>Milne&amp;Witten</td>
<td>0.177</td>
<td>0.094</td>
<td>0.087</td>
<td>0.174</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.229</td>
<td>0.174</td>
<td>0.116</td>
<td>0.240</td>
<td>0.282</td>
<td>0.299</td>
</tr>
<tr>
<td>Stories’ Embeddings</td>
<td>0.237</td>
<td>0.148</td>
<td>0.110</td>
<td>0.253</td>
<td>0.299</td>
<td>0.295</td>
</tr>
<tr>
<td>Wikipedia Embeddings</td>
<td>0.213</td>
<td>0.154</td>
<td>0.096</td>
<td>0.210</td>
<td>0.276</td>
<td>0.276</td>
</tr>
<tr>
<td>Reciprocal Rank</td>
<td>0.488</td>
<td>0.430</td>
<td>0.219</td>
<td>0.501</td>
<td>0.542</td>
<td>0.541</td>
</tr>
<tr>
<td>Salience</td>
<td>0.555</td>
<td>0.456</td>
<td>0.286</td>
<td>0.593</td>
<td>0.640</td>
<td>0.622</td>
</tr>
<tr>
<td>PPR</td>
<td>0.605*</td>
<td>0.564*</td>
<td>0.282</td>
<td>0.639*</td>
<td>0.678*</td>
<td>0.686*</td>
</tr>
</tbody>
</table>

Average improvements around +3/10% wrt Salience

- PPR better for detection of **Highly Relevant Entities**
- Qualitative better ranking for top ranked entities
Two Main Solutions
Unsupervised vs. Supervised

**Unsupervised** Solution based on Personalized PageRank, Salience, and Embeddings

**Supervised** Solution based on Feature Engineering and Learn to Rank
Supervised Solution

Trend in Question

Input

Supervised Solution

Feature Engineering

Learn to Rank

Output

Ranked Contextual Entities

Resources & Tools

Past Stories

Wikipedia Resources

- Wikipedia Statistics
- Wikipedia Graph
- Wikipedia2Vec Embeddings

Tools

- Entity Salience
Supervised Solution
Feature Engineering

FREQUENCY

Provide a sense on how often an entity is mentioned across the news stories.
Supervised Solution
Feature Engineering

Provide a sense on *where* an entity is *mentioned* across the news stories.
Supervised Solution
Feature Engineering

**POPULARITY**

Provide a sense on *general popularity* of an entity from background knowledge

Ken Murphy vs. Wikipedia
Supervised Solution
Feature Engineering

PO Popularity

Provide a sense on general popularity of an entity from background knowledge

Ken Murphy

VS.

Market

en.wikipedia.org/wiki/Ken_Murphy

en.wikipedia.org/wiki/SimonRoberts

en.wikipedia.org/wiki/Dave_Lewis

en.wikipedia.org/wiki/Financial_Market

en.wikipedia.org/wiki/Economy_of_the_UK

en.wikipedia.org/wiki/Wall_Street

en.wikipedia.org/wiki/NASDAQ
Supervised Solution
Feature Engineering

**POPULARITY**

Provide a sense on **general popularity** of an entity from background knowledge

- Ken Murphy: Wikipedia In-Degree = 12
- Market: Wikipedia In-Degree = 648
Supervised Solution
Feature Engineering

TEXT COHERENCE

Provide a sense of **coherence** in the texts of the **news stories**

**Murphy** new **Tesco** CEO

Will **Ken Murphy** succeed after **Lewis**?

\[
P(\text{Ken Murphy} | \text{Murphy}) = 0.4 \quad P(\text{Tesco} | \text{Tesco}) = 0.95
\]

\[
P(\text{Ken Murphy} | \text{Ken Murphy}) = 0.9 \quad P(\text{Dave Lewis} | \text{Lewis}) = 0.7
\]

TechAtBloomberg.com
Supervised Solution
Feature Engineering

NEURAL COHERENCE

Provide a sense of coherence of an entity with the other entities.
Supervised Solution
Feature Engineering

NEURAL COHERENCE

Provide a sense of coherence of an entity with the other entities

Ken Murphy
Market
ASDA
Tesco PLC
Hertfordshire
Dave Lewis
Supermarket
Supervised Solution
Feature Engineering

NEURAL COHERENCE

Provide a sense of coherence of an entity with the other entities

Ken Murphy

AVG Embeddings’ Cosine
0.57

Hertfordshire

AVG Embeddings’ Cosine
0.12
Supervised Solution
Feature Engineering

SALIENCE

Provide a sense on stories’ importance on an entity
Supervised Solution
Feature Engineering

SALIENCE

Provide a sense on stories’ importance on an entity

Cosine

\[
\begin{bmatrix}
0.7 & 0.4 & \ldots & 0.5 \\
0.9 & 0.5 & \ldots & 0.6
\end{bmatrix} = 0.73
\]
Supervised Solution
Feature Engineering

**SALIENCE**

Provide a sense on *stories’ importance* on an entity

Past Stories

Story Frequency where *Tesco PLC* and *Ken Murphy* have salience score ≥ τ (= 0.3)

Trend in Question

Ken Murphy

TechAtBloomberg.com

© 2021 Bloomberg Finance L.P. All rights reserved.
Experimental Results

- Dataset split into training / validation / test sets
- Model tuned on training / validation sets
- Results of the methods reported on the test set
### Experimental Results

#### “Relevant” & “Somewhat Relevant” as Gold Labels

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P@1</th>
<th>P@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salience</td>
<td>0.474</td>
<td>0.569</td>
<td>0.364</td>
<td>0.526</td>
<td>0.584</td>
<td>0.714</td>
</tr>
<tr>
<td>PPR</td>
<td>0.495</td>
<td>0.646</td>
<td>0.364</td>
<td>0.565</td>
<td>0.617</td>
<td>0.767</td>
</tr>
<tr>
<td>LTR</td>
<td><strong>0.574</strong></td>
<td><strong>0.708</strong></td>
<td><strong>0.472</strong></td>
<td><strong>0.629</strong></td>
<td><strong>0.682</strong></td>
<td><strong>0.815</strong></td>
</tr>
</tbody>
</table>

#### “Relevant” as Gold Label

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>P@1</th>
<th>P@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salience</td>
<td>0.534</td>
<td>0.462</td>
<td>0.251</td>
<td>0.566</td>
<td>0.616</td>
<td>0.604</td>
</tr>
<tr>
<td>PPR</td>
<td>0.591</td>
<td>0.554</td>
<td>0.256</td>
<td>0.622</td>
<td>0.659</td>
<td>0.665</td>
</tr>
<tr>
<td>LTR</td>
<td><strong>0.609</strong></td>
<td><strong>0.569</strong></td>
<td><strong>0.308</strong></td>
<td><strong>0.654</strong></td>
<td><strong>0.696</strong></td>
<td><strong>0.710</strong></td>
</tr>
</tbody>
</table>

Average improvement around +6/9% wrt PPR

- Qualitative better ranking in the head+tail
- Able to detect partially relevant entities
Conclusion
Take-Home Messages

▷ We introduced the problem of ranking contextual entities for a trend

▷ We experimented with two methods:
  ○ Unsupervised, based on Personalized PageRank.
  ○ Supervised, based on feature engineering and learning to rank.

▷ The Supervised method generally works better than unsupervised ones
  ○ With improvements between +10% in terms of P@3
  ○ Feature analysis show most important features are based on salience and text coherence
Thank you!

We are hiring: bloomberg.com/engineering