Quotebank: A Corpus of Quotations from a Decade of News

Timoté Vaucher (EPFL), Andreas Spitz (EPFL), Michele Catasta (Stanford), Robert West (EPFL)

WSDM 2021 – Israel
Motivation

- Ease of access to news
- Prevalence of misinformation and Fake News
- Necessity of fact checking

Who said what?

Did they really say that?

By Joseph D’Urso
BBC Political Research Unit

© 16 February 2017
Contributions

**Quobert**
- End-to-end framework for extracting and attributing quotations
- Scalability for large corpora
- Distantly and minimally supervised
- Single seed pattern (e.g. “Q”, said S)

**Quotebank**
- Corpus of 178 mio unique attributed quotations from 163 mio news article in English
- Spans over a decade (2008 – 2020)
Quotation Attribution (in a nutshell)

Example 1

“Quotebank contains millions of quotations,” Tim told the audience.
Quotation Attribution
(in a nutshell)

Example 1
“Quotebank contains millions of quotations,” Tim told the audience.

Example 2
Tim has made strong claims to Jane, a fellow student, such as “Quotebank contains millions of quotations.”

- Multiple entities to choose from
- No syntactic clues
- Requires semantic understanding
### Historical Approaches

<table>
<thead>
<tr>
<th>Unsupervised</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ High precision rules</td>
<td>Can make a prediction for every quotation based on context</td>
</tr>
<tr>
<td>Bootstrap and discovery of rules (Quootstrap)</td>
<td>Capable of handling implicit mentions</td>
</tr>
<tr>
<td>− Requires many rules to account for the complexity of a language</td>
<td>Requires a large, fully annotated corpus (does not scale)</td>
</tr>
<tr>
<td>Low recall</td>
<td></td>
</tr>
</tbody>
</table>

Attribution framework

Our approach

- Combine the best of the supervised and unsupervised world
- Use pattern-based bootstrapping to generate enough training data for a supervised transformer architecture

"Quotebank contains millions of quotations"

“Quotebank [...],” said Tim.

“Quotebank [...],” Tim told the audience.

Tim has made strong claims to Jane, a fellow student, such as “Quotebank [...].”

“Quotebank [...],” the student said.
Unsupervised pattern matching using bootstrapping to generate new rules.

Start with a single seed: “Q”, said S

“Quotebank [...],” said Tim.

“Quotebank [...],” Tim told the audience.

Tim has made strong claims to Jane, a fellow student, such as “Quotebank [...].”

“Quotebank [...],” the student said.
2: Data Expansion

Finding all the remaining occurrences of quotation–entity pairs that were not covered by rules in Quootstrap

“Quotebank [...]” said Tim.

“Quotebank [...]” Tim told the audience.

Tim has made strong claims to Jane, a fellow student, such as “Quotebank [...].”

“Quotebank [...],” the student said.
3: Quobert

Fine tune a transformer architecture with a classification head to predict a probability distribution over the candidate entities.
Quobert – Evaluation

Crowd annotated 1500 quotations and their context

Categories

- Implicit and complex contexts not recognized by Quootstrap
- Contexts with many entities to choose from
- Representative randomly sampled contexts

Baselines

- Nearest Entity
- Sieve based quotation attribution model available in Stanford CoreNLP

## Quobert – Results

<table>
<thead>
<tr>
<th></th>
<th>Quootstrap</th>
<th>Baseline</th>
<th>CoreNLP</th>
<th>Quobert(_{BAL,sum})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implicit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>0</td>
<td>0</td>
<td>0.763</td>
<td>0.635</td>
</tr>
<tr>
<td>−</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.494</td>
</tr>
<tr>
<td>All</td>
<td>0</td>
<td>0.471</td>
<td>0.581</td>
<td></td>
</tr>
<tr>
<td><strong>Many-choice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>0.316</td>
<td>0.606</td>
<td>0.606</td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.194</td>
</tr>
<tr>
<td>All</td>
<td>0.253</td>
<td>0.485</td>
<td>0.592</td>
<td></td>
</tr>
<tr>
<td><strong>Representative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>0.331</td>
<td>0.803</td>
<td>0.773</td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>0</td>
<td>0</td>
<td>0.466</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>0.253</td>
<td>0.614</td>
<td>0.701</td>
<td></td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>0.167</td>
<td>0.528</td>
<td>0.629</td>
<td></td>
</tr>
</tbody>
</table>
Quotebank in Numbers

- 178 mio unique attributed quotations
- Over 900,000 distinct speakers
- 377,000 unique web domains
- During 13 years (Sep 2008 – Apr 2020)
Using Quotebank

WOW
such quotations
many speakers
much sources
so excited

imgflip.com
Using Quotebank

4 axes for you to explore

Which sources reported “who said what” when?

- Data-driven analyses at scale of the public statements of hundreds of thousands of speakers
- Generation of claim provenance graphs for fact checking and combating fake news
- Research into the propagation and coverage of events in the news

NLP, Computational Social Science, Journalism
Example Application
Shift in Political Attitude

How Lindsey Graham Went From Trump Skeptic to Trump Sidekick

- Sources: all
- Who: Lindsey Graham
- What: Quotations about Donald Trump
- When: From June 2015, when Trump announced his candidacy for President of the United States onwards
Conclusion

Along **Quotebank**, the code for **Quobert**, our distantly and minimally supervised end-to-end, language-agnostic framework for extracting and attributing quotations is available at

https://github.com/epfl-dlab/Quotebank

We hope you will make ample use of Quotebank and Quobert in your research!

“Thank you for your attention!” said Tim.