



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?

BBC Sign in Home News Sport

NEWS

Home Coronavirus Video World UK Business Tech Science Sto

Politics Parliaments Brexit Election 2019

Did they really say that?

By Joseph D'Urso
BBC Political Research Unit

© 16 February 2017

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

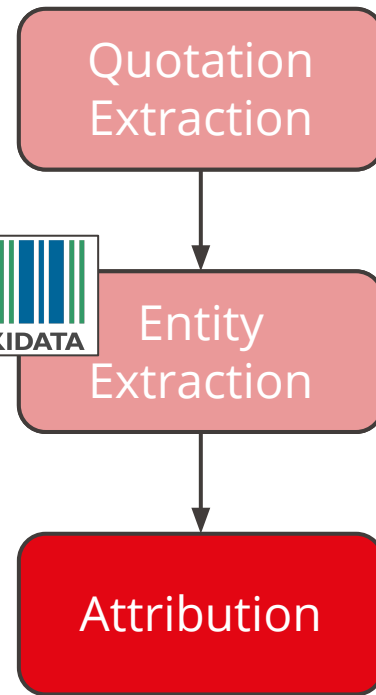
- Corpus of 178 mio unique attributed quotations from 163 mio news article in English
- Spans over a decade (2008 – 2020)



Quotation Attribution (in a nutshell)

Example 1


"Quotebank contains millions of quotations," Tim told the audience.



Quotation Attribution (in a nutshell)

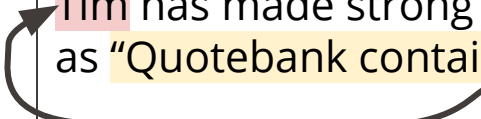
Example 1

"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

Quotation
Extraction

Entity
Extraction

Attribution

Historical Approaches

Unsupervised

- + High precision rules
Bootstrapping and discovery of rules
(Quootstrap)
- Requires many rules to account for
the complexity of a language
Low recall

Supervised

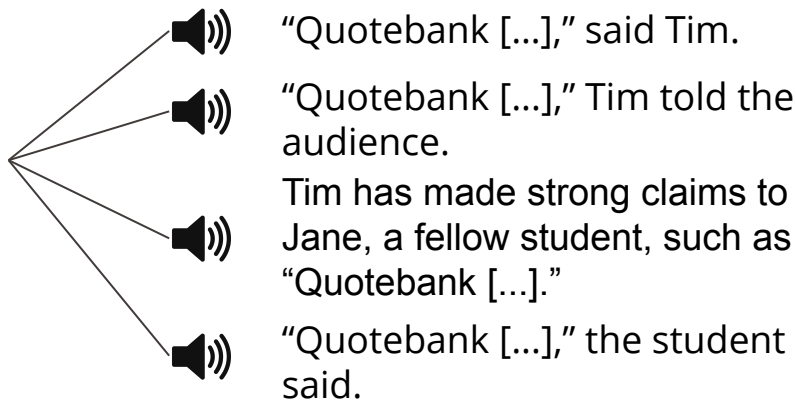
- Can make a prediction for every
quotation based on context
Capable of handling implicit mentions
- Requires a large, fully annotated corpus
(does not scale)

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"

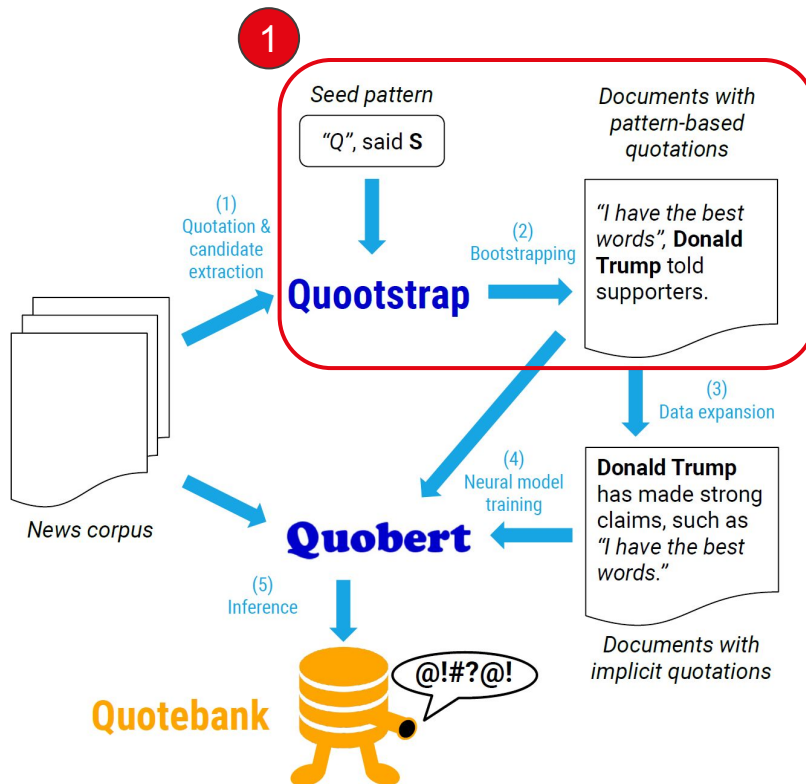


1: Bootstrapping with Quootstrap

Unsupervised pattern matching using bootstrapping to generate new rules.

Start with a single seed: “Q”, said S

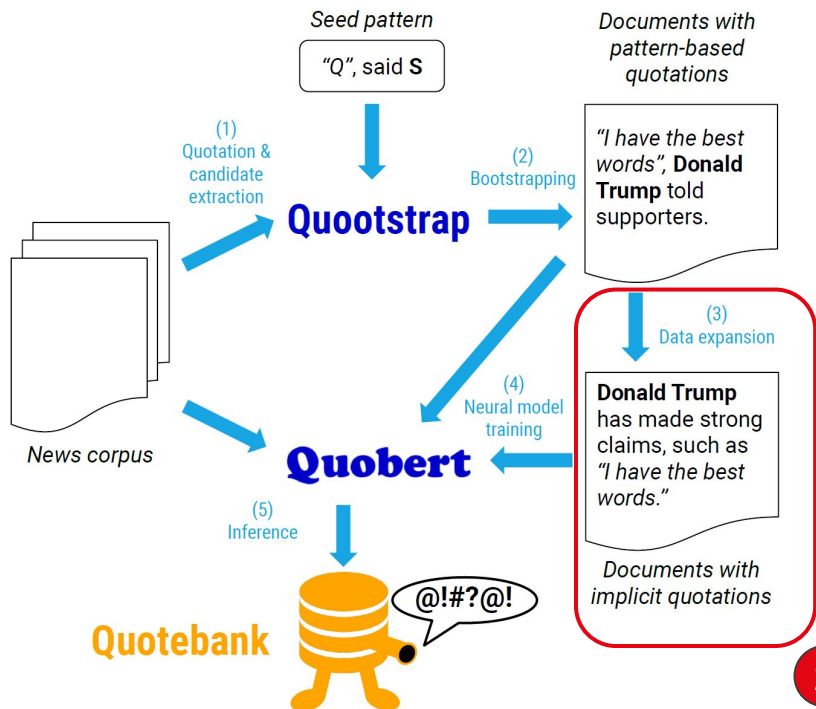
- “Quotebank [...],” said Tim.
- “Quotebank [...],” Tim told the audience.
- Tim has made strong claims to Jane, a fellow student, such as “Quotebank [...].”
- “Quotebank [...],” the student said.



2: Data Expansion

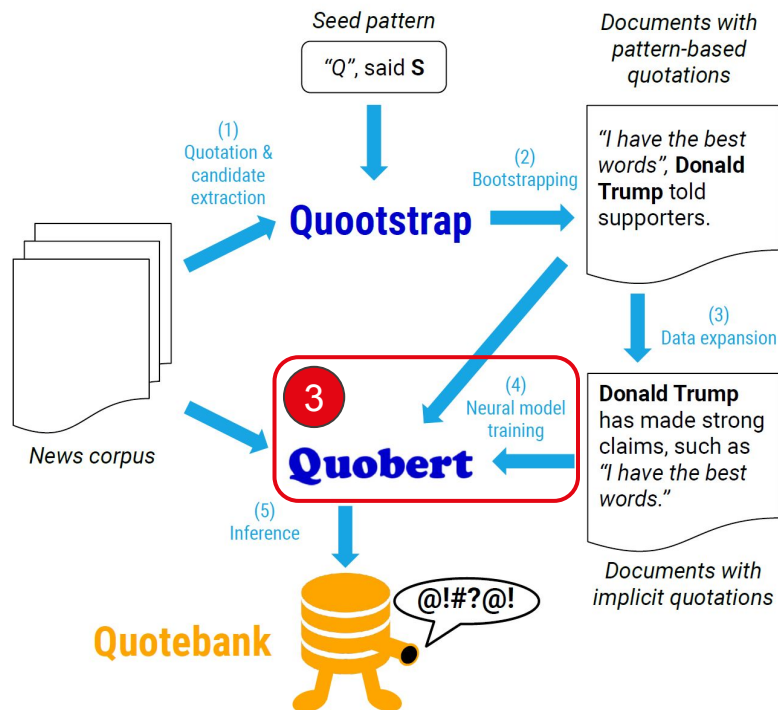
Finding all the remaining occurrences of quotation–entity pairs that were not covered by rules in Quootstrap

- “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

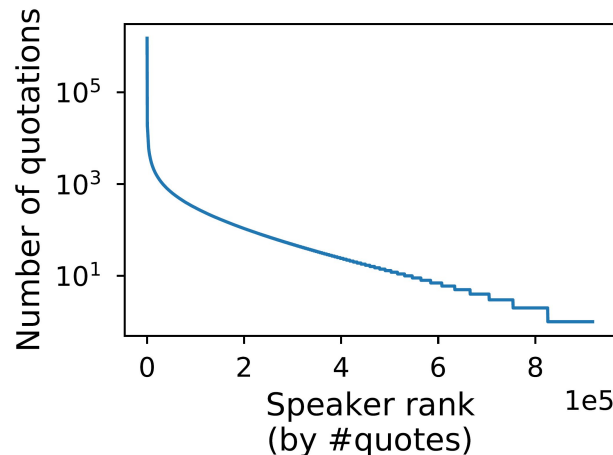
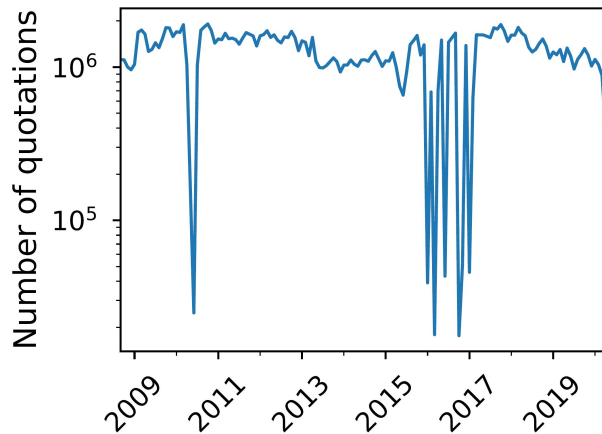
- Nearest Entity
- Sieve based quotation attribution model available in Stanford CoreNLP

Quobert – Results

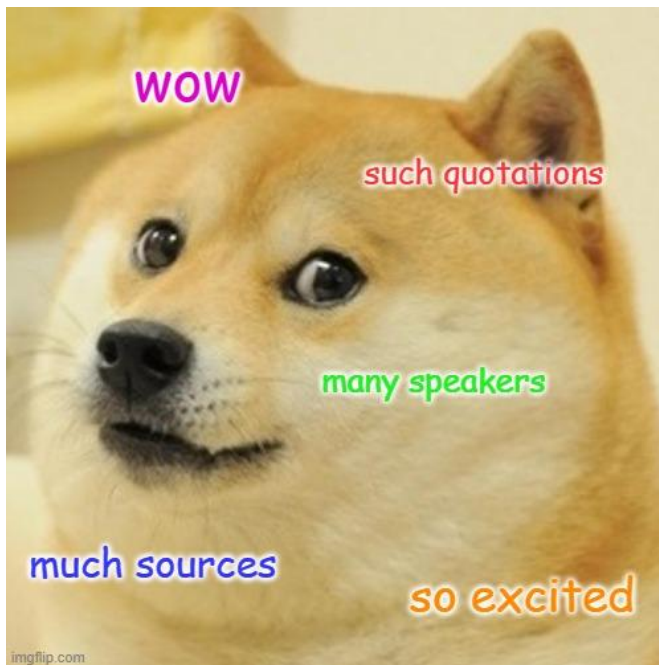
		Quootstrap	Baseline	CoreNLP	Quobert _(BAL, sum)
Implicit	+	0	0.763	0.635	0.855
	–	0	0	0.494	0.857
	All	0	0.471	0.581	0.856
Many-choice	+	0.316	0.606	0.606	0.911
	–	0	0	0.194	0.731
	All	0.253	0.485	0.592	0.875
Representative	+	0.331	0.803	0.773	0.899
	–	0	0	0.466	0.806
	All	0.253	0.614	0.701	0.877
Overall		0.167	0.528	0.629	0.869

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



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

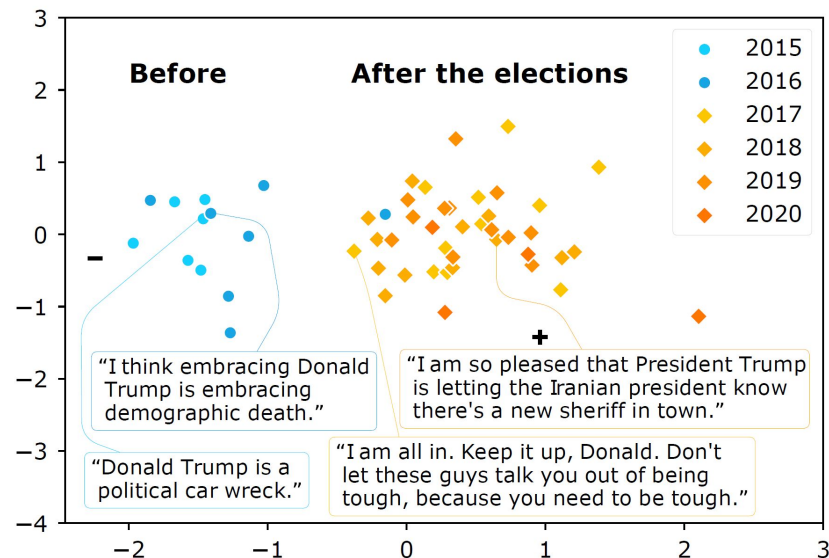


Example Application

Shift in Political Attitude



- 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.

