Confronted with the influx of daily news, journalists and social scientists struggle with answering the question “who said what?” at Web-scale. The process of extracting quotations and determining the speaker is known as quotation attribution.

To address this problem, we introduce Quobert, a minimally supervised framework for extracting and attributing quotations in massive corpora. We use it to create Quotebank, a Web-scale corpus of annotated quotations.

To generate Quotebank, we apply Quobert to 196M English news articles that were extracted from 377k websites between 2008 and 2020.

The results are available to the community and contain 178M unique quotations attributed to more than 900k speakers.

How Quobert Works

We combine the advantages of supervised and unsupervised learning by using pattern-based bootstrapping to generate training data for a supervised transformer architecture:

1. Extract all quotations and candidate speakers from the articles.
2. Leverage Quootstrap [1], an unsupervised pattern matcher that uses bootstrapping to generate new rules. The process starts with a single language-agnostic rule.
3. Find all remaining occurrences of the extracted quotations that were not covered by rules in Quootstrap to generate rich training contexts.
4. Fine tune a transformer architecture with a classification head to predict a probability distribution over the candidate entities.
5. Apply the model on the news corpus to generate Quotebank.

Evaluation of Quobert

To verify how well the correct speaker is attributed, we evaluate Quobert on 1500 crowd-annotated quotations of three different types:

- Implicit and complex contexts
- Contexts with many entities to choose from
- A representative set, sampled at random

We compare to three baselines:

- Heuristic: select the nearest candidate entity
- Rule-based attribution by Quootstrap [1]
- The CoreNLP sieve approach [2]

Exploring Quotebank

To highlight the potential of Quotebank to enable data-driven analyses of the public statements of hundreds of thousands of speakers, consider this PCA projection of the embeddings of quotations from US Senator Lindsey Graham (aggregated by month), who is famous for reversing his stance on Donald Trump after the 2016 U.S. elections.

“We think embracing Donald Trump is embarrassing,” demographic death.

“I am so pleased that President Trump is letting the Iranian president know there’s a new sheriff in town.”

“I am all in. Keep it up, Donald. Don’t let these guys talk you out of being tough, because you need to be tough.”

References

[1] Dario Pavlov, Tiziano Piccardi and Robert West. Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. 2018, ICWSM’18

This poster was presented at the 14th International Conference on Web Search and Data Mining (WSDM ’21), March 8-12, 2021, Jerusalem, Israel.