

The provided paper illustrates the creation of Quotebank, a corpus that associates quotations taken from English-speaking news articles to their respective speakers. Here are a couple ideas of projects that can be done based on Quotebank.

Idea 1:

We know that Quotebank essentially extracts quotations from provided news articles and attempts to associate them to the person that uttered those words. However, in addition to that, Quotebank also indicates the source of the quotation, essential in providing context to the quotation. Therefore, a possible use for this would be to use Quotebank as a way to factcheck possible fake news or wrongly attributed quotations. Basing ourselves off the fact that Quotebank's association of speaker to quotation is mainly correct, an idea could be to take a dataset composed of a number of quotations used on social media for example and crosscheck these quotations with Quotebank to verify if the attributed speaker is correct or not. Related to this, an additional function would be to use the source referencing on Quotebank to determine if the given quotation is used in the appropriate context or not. Even though the latter concept seems complicated for a model to discriminate, we could define different contexts with a list of keywords for each, allowing us to determine if the quotation is in a misplaced context or not.

Idea 2:

Another possible use for Quotebank would be to compile quotations from a particular given date. Indeed, Quotebank also indicates the date of the specified quotation, based on the publication date of the news article. Therefore, we could crossreference a dataset containing a list of dates pertaining to specific events for example and construct a list containing keywords, speakers per date. Performing this experiment the other way around could be interesting as well, meaning that creating such a list based on events could result in an interesting compilation of keywords and themes. Although only events from 2019 are considered for this report, we could easily imagine this over a larger time period.

Idea 3:

Finally, another potentially interesting project would be to analyse the nature of the quotations extracted from various news articles. Indeed, as mentioned in the paper, Quotebank depends on a supervised framework called Quobert to sort and correctly associate quotation and speaker. However, Quobert was "taught" to recognise various patterns of quotations commonly used in news articles of different genres. Indeed, there are direct quotations, indicated by quotation marks, as there are indirect quotations, used in anaphora, in multi-entity sentences, ... An initial analysis providing statistics on the use of these different quotation types based on the genre of news article could give an interesting insight in the writing style of each genre and in the English language in general. Furthermore, one could imagine a similar database as Quotebank in a different language (French for example!) and go on to compare such statistics concerning quotation type based on the news article genre between different languages.

Feedback: **Needs improvements**

Overall, we like the way you explained the ideas.

Below the comments for each idea:

Idea #1

- Not sure about your idea. You basically want to verify if the dataset we gave you is accurate, by looking for the same quotations on other sources (Twitter?).

Idea #2

- Can you provide an example of the dataset you would use?

Idea #3

- Okay, the idea can be interesting. You have to think on how you want to proceed, but we like the effort in thinking about this original idea.

Best,