Contextualizing AI-Generated Transcripts for Researchers

University of California San Francisco

Lubov McKone, Pratt Institute
Geoffrey Boushey, University of California, San Francisco

Pratt Graduate Student Engagement Fund

Background

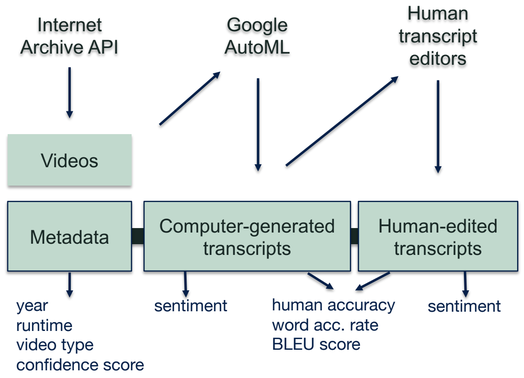
At the UCSF Library, technical staff often provide text transcripts generated from image, audio, or video collections for researchers. Our collections often involve a wide variety of media types, some of which provide more accurate transcriptions than others. For example, videos from community meetings, when audio is poor and multiple people speak at once, may translate less accurately through speech-to-text than a video of a nightly news program.

Unless we communicate the variance in transcription accuracy that results from these AI tools to researchers, we risk introducing bias into the data we provide. As a result, we decided to evaluate the variance and accuracy for transcripts generated from on a small collection of videos from the UCSF Tobacco Archives

Research Questions

- Taking into account factors such as year and runtime, how does computer transcription accuracy differ between television commercials and court proceedings?
- How might transcription accuracy impact the conclusions drawn from the data?
- What guidance can we give researchers to prevent uninformed conclusions?

Data collection



Methods

We compared transcripts using the following metrics:

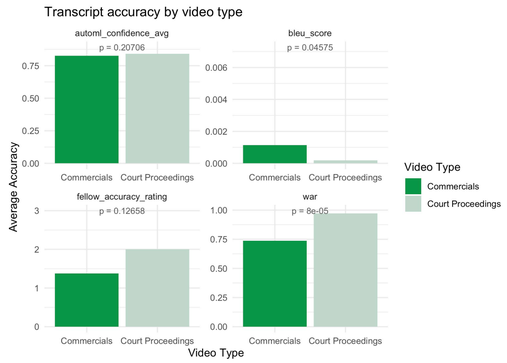
Accuracy

- **Word Accuracy Rate** – how many changes convert the test transcript into the reference transcript (inverted)
- **BLEU Score** – a more advanced algorithm measuring n-gram matches normalized for n-gram frequency.
- **Human-evaluated accuracy** – Good, Fair, or Poor
- **AutoML confidence score** – a score indicating how accurate Google believes its transcription to be.

Meaning

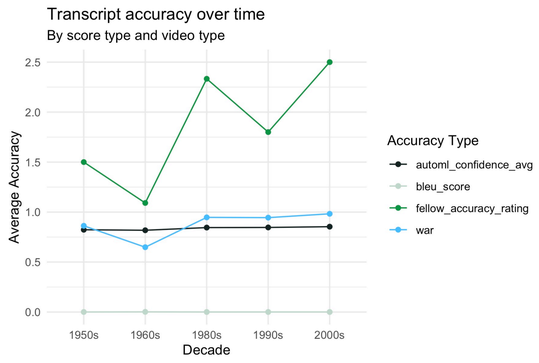
- **Sentiment score**
- **Topic modeling**

Findings



Court proceeding transcripts are more accurate than commercials

Across most metrics, the court proceeding transcripts were more accurate than the commercials. One potential reason for this is that commercials typically include some form of singing or more stylized speaking, which Google AutoML had trouble transcribing.

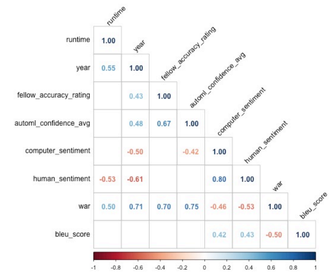


The more recent the video, the more accurate the transcript

There's a general improvement in transcription quality after the 1960s, but not a dramatic one. Interestingly, this trend disappears when looking at each video type separately.

Metrics of transcript accuracy

Significant correlations



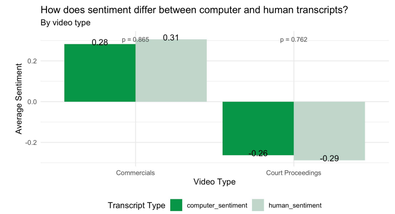
AutoML's confidence score is significantly positively correlated with human-evaluated accuracy and word accuracy rate

Recommendations

- Include the Google AutoML confidence scores when providing AutoML-generated transcripts to researchers
- Inform researchers that newer videos may generate more accurate computer transcriptions.
- Inform researchers that video transcripts from media that contain singing or stylized speaking may be less accurate.

Future directions

Some aspects of the data that we did not investigate as much as we had hoped were sentiment and topic modeling, although we did find that the sentiment of the computer transcripts was a slight underestimation in both categories.



We are also curious to investigate the same questions for other transcription services, such as Whisper

Recommendations for other text analysis researchers

- Our research culminated in a reusable dataset that lives in our **github repository**: **https://github.com/geoffswc/Internet-Archives-Transcripts**
- BLEU score is supposed to range from 0-1, but in our study its range was 0.0001 – 0.007 because the transcripts generated by Google AutoML did not contain any punctuation. We don't recommend the use of the BLEU score metric on transcripts generated by Google AutoML, or on other computer-generated transcripts that lack punctuation.