RadioTalk: a large-scale corpus of talk radio transcripts

Doug Beeferman, William Brannon, Deb Roy

{dougb5, wbrannon, dkroy}@media.mit.edu

Lab for Social Machines, MIT Media Lab

Introduction

RadioTalk is a corpus of automatic transcripts from US talk radio broadcasts in 2018 and 2019.

- 2.8 billion words of speech from 250+ radio stations from Oct 2018 to Mar 2019.
- 284,000 hours of broadcasts, sampled to every other 10 minute period.
- Much additional metadata: partial diarization, speaker gender, radio program schedule info, studio/telephone origin of audio (to flag call-ins) etc.

Why does radio matter?

- Last naturally local part of broadcast media (cf TV) and provides local opinions not easily available elsewhere.
- Call-ins are the original tweets: short expressions of local opinion, still going strong.
- Radio is still a major source of news and info in rural America.

Stations Included

Two groups of included radio stations:

- A nationally representative, stratified random sample of 50 US radio stations.
- A convenience sample of another ~200 stations, esp. in MA, WI and NE.

Figure 1: Initial (top) and current (bottom) panels of ingested stations.

Ingestion System

Processing occurs in two asynchronous phases:

- Audio ingest from online station streams. High redundancy for high availability.
- Processing captured audio: transcription and metadata creation. These outputs are turned into bulk datasets.

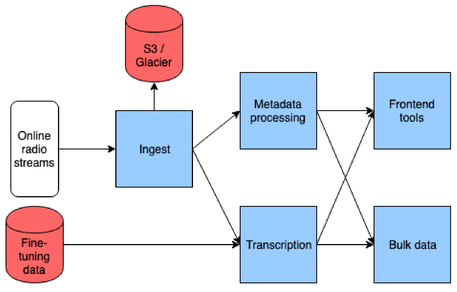

Figure 2: A high-level diagram of the radio processing pipeline.

Application: Word Embeddings

Word embeddings trained on RadioTalk can be used in conversational speech applications, but also perform well in conventional analogy evaluation:

	RadioTalk Skipgram	Original
Google (sem)	41.1%	55.0%
Google (syn)	40.3%	59.0%
MSR (syn)	38.7%	39.6%

Table 1: Note: these embeddings are a proof of concept and need not have state of the art performance. See Mikolov et al 2013 (arXiv:1301.3781) for original values.

Application: Syndication's Influence

Syndicated radio may influence local radio. That is, national media may set the agenda for local media. Many ways to test whether this happens, including:

- Coverage of events: does news break earlier / is the news cycle faster on syndicated radio?
- Distinctive phrases: do phrase variants starting on syndicated radio diffuse to local radio?
- Modeling: do lags of syndicated coverage predict local coverage?

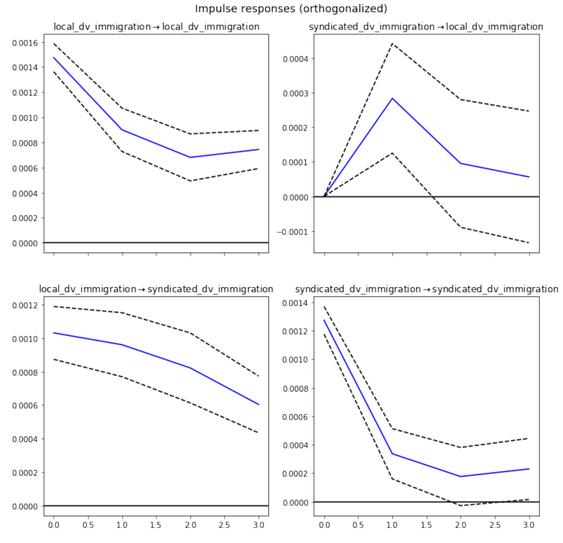

Figure 3: Impulse response functions of VAR model on local and syndicated coverage of immigration.

Dataset Available at:

https://github.com/social-machines/RadioTalk

Transcription

Custom speech to text model, not focus of research:

- Based on an open-source model from Johns Hopkins, entered in a US government competition.
- Neural acoustic model, n-gram language model.
- Language model fine-tuned on several thousand hours of talk radio (Rush Limbaugh and NPR) from shows with human-produced transcripts.
- Word error rate averages about 13%. More accurate models have not been cost-effective at scale.

Next Steps

- Applications: understanding news cycle dynamics, in radio and vs social media.
- Better automatic transcription. Current word error rates of about 13% are adequate but can be reduced.
- Better methods for identifying syndicated content.

Conclusions

Why is this novel?

- One of the largest and most diverse corpuses of conversational speech in English.
- No previous comprehensive corpus of radio transcripts.
- Radio callers are disproportionately from groups underrepresented on social media.

Who should care about this?

- Social sciences and media studies: research on the dynamics and political role of talk radio.
- Natural language processing: a large and diverse variety of conversational speech for training NLP models.
- Linguistics / conversation analysis: a window into dialect and spontaneous speech for a large variety of US English speakers.

Acknowledgements

The authors would like to thank the Ethics and Governance of AI Fund for supporting this work.