Auto-annotation for Voice-enabled Entertainment Systems

Wenyan Li, Ferhan Ture

Comcast Applied AI Research Lab

Motivation

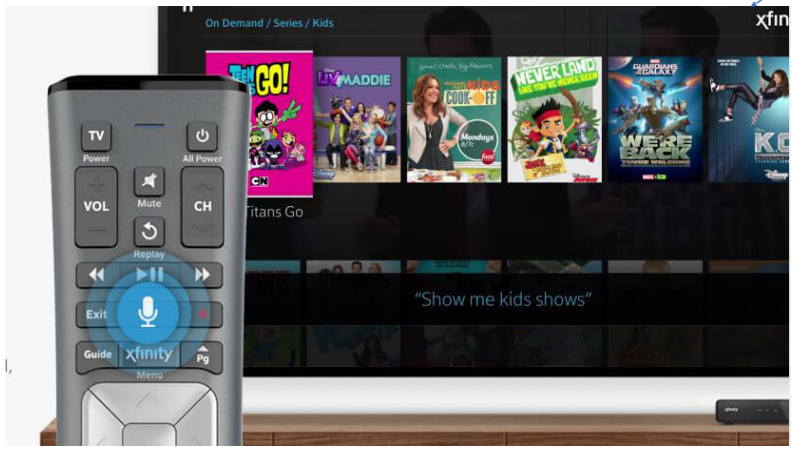
- Voice-activated intelligent entertainment systems are prevalent in our daily lives

in specialized domains
shifting customer trends

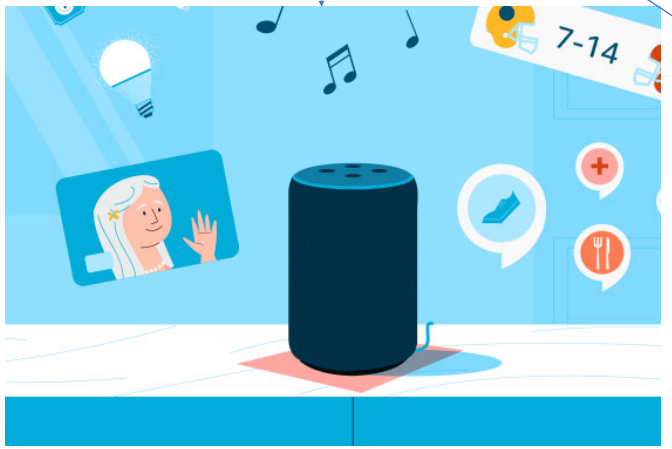
Labeled audio data

time consuming
labor intensive

Automatic Speech Recognition



Comcast X1

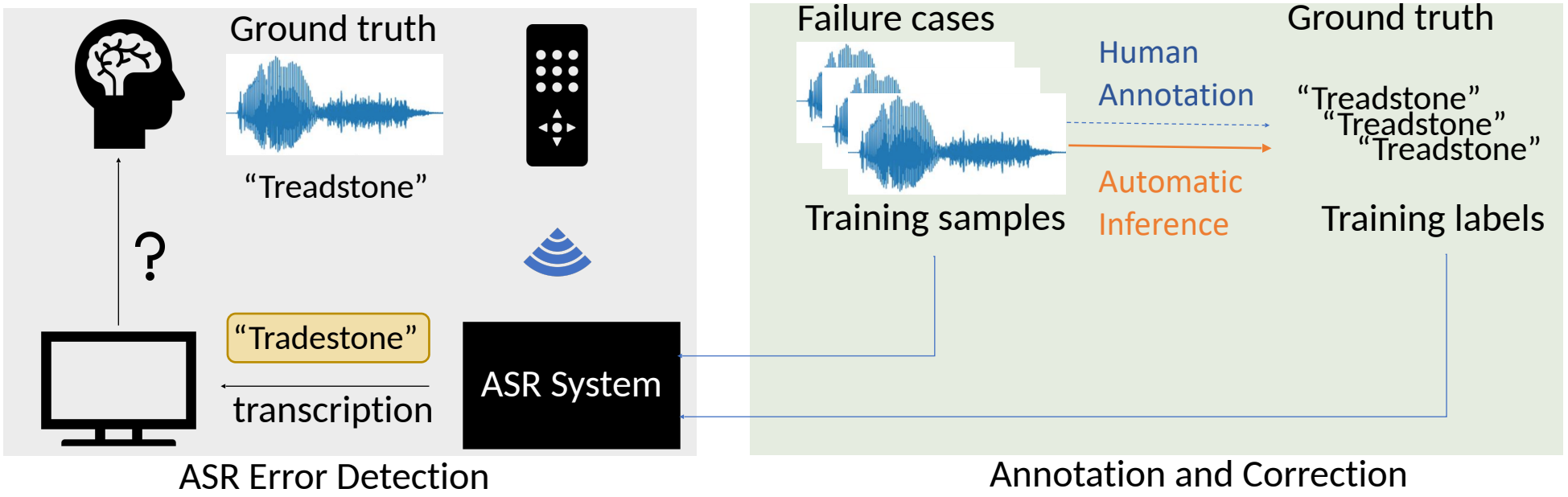


Amazon Alexa

Google Nest Hub

Proposal

- Identify errors and provide reasonable corrections in ASR systems automatically
- Obtain training data for production-level ASR systems in an **unsupervised** manner



RoadMap

- Motivation
- Proposal
- Prior work
- Auto-annotation Methods:
  - Utterance-based
  - Interaction-based
- Evaluation
- Conclusion

Prior Work

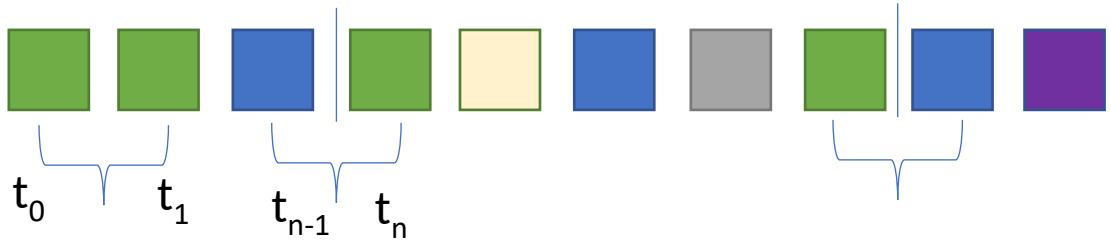
- Fully Supervised
  - Lexical and acoustic features of query reformulation [Tang, et al. 2019; Errattahi, et al. 2018; Errattahi, et al. 2016]
  - Audio-only interactions [Hong and Findlater. 2018]
- Weakly supervised
  - acoustic audio event detection [Kumar and Raj. 2016, Kumar et al. 2017]
- Production-level
  - human-in-the-loop [Amazon SageMaker]

RoadMap

- Prior work
- Auto-annotation Methods:
  - Utterance-based :
    - transcriptions only
  - Interaction-based:
    - transcriptions + interaction events (app, tune, keypress)
- Evaluation
- Conclusion

Data: utterance sessions

- Each query is issued by the same device.
- Each non-first query occurs within 45 seconds of the last. (Tang et al., 2018)



transcriptions are grouped by device and sorted by time

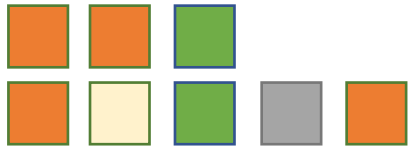
If t_n-t_{n-1} > 45s, have a session = events[t_0 : t_{n-1}]

Sessions: [, ,]

[[{‘time': 'Mon Oct 7 00:11:19 2019', 'transcription': 'NBC America'}
{‘time': 'Mon Oct 7 00:11:23 2019', 'transcription': 'NBC America'},
{‘time': 'Mon Oct 7 00:11:35 2019', 'transcription': 'NBC'}]

Utterance-based Detection and Annotation

- Step 1: Identifying erroneous transcriptions
  - Session



☑ have more than one utterances

☑ contain repeats

  - Query  → Erroneous

☑ Number of session > T_s

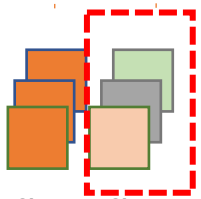
☑ Repeat likelihood > T_{rep}

☑ Median of time interval between repeats < T_t

- Step 2: Automatic annotation with query reformulation patterns

query = Log house

sessions



q_{err} $q_{i'}$

☑ more than one utterances

☑ no repeated transcriptions

☑ second last query is erroneous

☑ Median of time interval < T_t

- Extract possible candidates from all sessions
- Select correction by confidence

$$P(q_i' \mid q_{err}) = \frac{count(q_i', q_{err})}{count(q_{err})}$$

('Log House', 'Loud House'), 0.44

('Log House', 'The Loud House'), 0.125

…

RoadMap

- Prior work
- Auto-annotation Methods:
  - Utterance-based :
    - transcriptions only
  - Interaction-based:
    - transcriptions + interaction events (app, tune, keypress)
- Evaluation
- Conclusion

Interaction-based Detection and Annotation

- Three scenarios where subsequent user interactions lead to a confirmed transcription

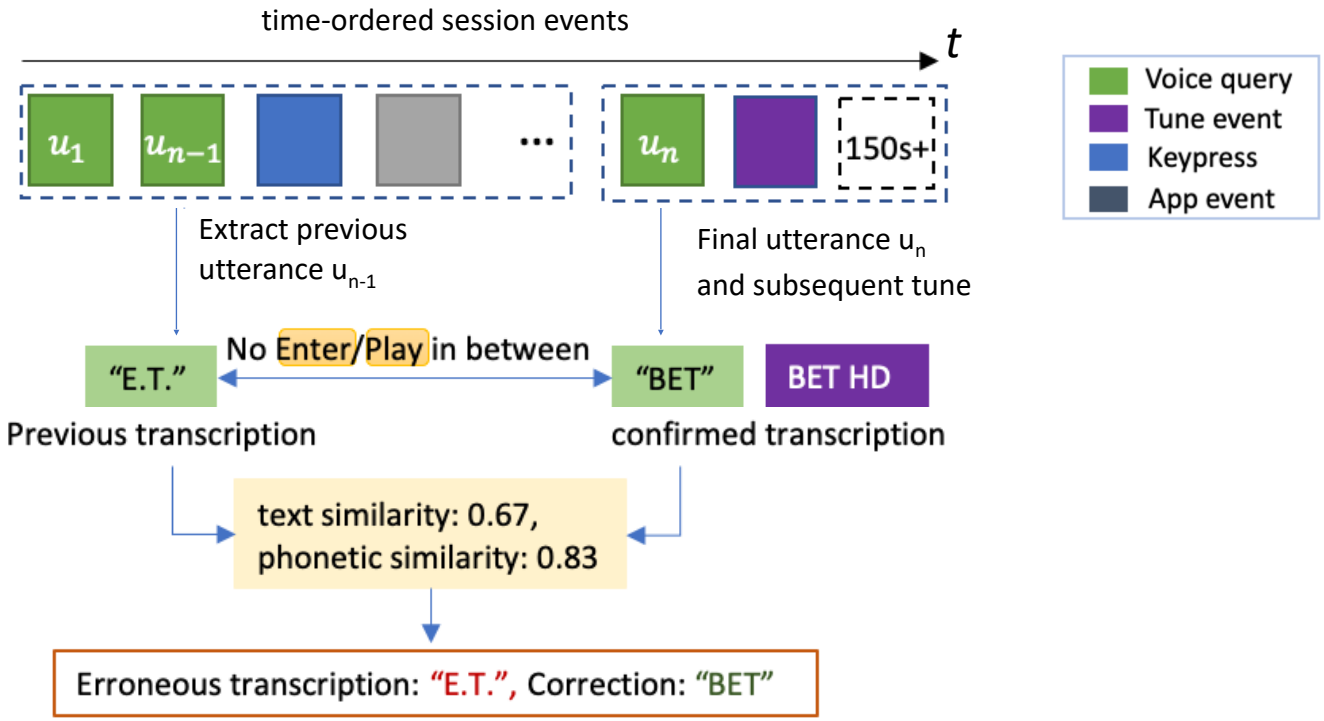
Scenario 1: User clicks a button, tunes to a program, then keeps watching > 30 seconds.

Scenario 2: User launches an app and stays for at least 30 seconds

Scenario 3: User tunes to a program and stays for at least 150 seconds.

Interaction-based Detection and Annotation

- After we confirm the transcription of the last voice query:
  - Look at the previous query from the same user to see if it points to a potential system error



Evaluation

- Utterance-based Detection and Annotation

Data	474 sessions with 925 utterances on 68 detected erroneous queries
Error detection	Precision: 77.49% Recall: 76.11%
Suggest correction	Accuracy: 69.76% WER would have reduced by 0.907

- Interaction-based Detection and Annotation

Data	225 queries with 15 utterance samples for 15 unique queries
Confirmed transcriptions	Accuracy: 99.60% WER of 0.002
Identify suspicious transcriptions	Accuracy: 87.1.%
Suggest correction	Accuracy: 80.0%

Methods	Suspicious transcriptions	Suggested annotation
Utterance-based Annotation	Academy	Pup Academy
	CNET Latino	Cinelatino
	Entouch	In Touch
	Murder in the First	Murder in the Thrist
	Joe Josie Wah	JoJo Siwa
	Antena*	Antenna
Interaction-based Annotation	Play Select	Player Select
	Just Go With It	Just Roll With It
	Find	Friends/Blind Date
	Stephen	Steven Universe
	60 Days N	60 Days In
	The Singer	The Masked Singer/Masked Singer

Conclusion

- We present an automated annotation system that provides an unsupervised approach to identify erroneous transcriptions and to suggest possible fixes for detected errors.

- Our auto-annotated training data reaches an overall word error rate (WER) of 0.002 and we obtained a reduction of 0.907 in WER after applying the auto-suggested fixes.

- Our system can be directly applied to improve annotation efficiency and robustness of ASR systems