SHORT UTTERANCE DIALOGUE ACT CLASSIFICATION USING A TRANSFORMER ENSEMBLE

University of East London
Pioneering Futures Since 1898

Harry Maltby (Harrym906@gmail.com), Julie Wall, Tracy Goodluck Constance, Mansour Moniri, University of East London
Cornelius Glackin, Marvin Rajwadi, Nigel Cannings, Intelligent Voice Ltd., London, UK

INTELLIGENT VOICE

WHAT IS A DIALOGUE ACT?

Essentially, a dialogue act is a type of sentence classification that denotes its linguistic purpose, e.g. statement, question. It's a label to explain the intent and purpose of what was said.

WHY IS THIS IMPORTANT?

- "really(?)", "yeah(?)", "this is your car(?)", can you tell what is meant by the text alone?
- Reliable dialogue act classification is crucial for digital assistants to be able to understand requests and respond appropriately, transcription tools need to know the dialogue act to add punctuation.
- AI agents need dialogue act classification to understand user requests.
- Most dialogue act classification methods in the literature are text-based and lack context for short utterances.

WHAT DID WE DO?

- We explored the use of both text and audio to classify dialogue acts, improving on previous attempts to combine the two.
- Using state-of-the-art models we combined what was said with how it was said by combining text and audio to classify dialogue acts.

THE PROBLEM

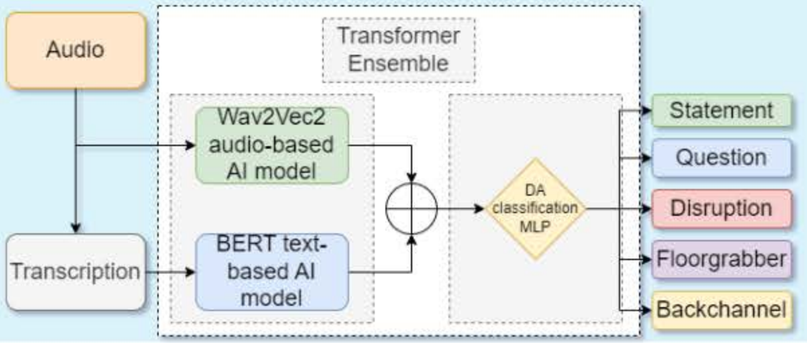
- Taking in what is said is not always good enough, what we need to know is how it's said.
- Declarative questions and short utterances cause ambiguity for both people and AI.

- Sometimes the same phrase can mean different things depending on how it's said.
- Text-based AI models can only look at what was said and not how it was said.

OUR SOLUTION

- Taking in what is said is not always good enough, what we need to know is how it's said.
- Declarative questions and short utterances cause ambiguity, to reduce this ambiguity we combined what was said with how it was said.
- Using state-of-the-art models we combined what was said (the text), with how it was said (the audio).
- To combine what was said with how it was said we used an artificial neural network (ANN).
- We used Facebook's pre-trained XLRS Large 960 self Wav2Vec2 model fine-tuned for the acoustic model and BERT-base-uncased for the lexical model, both gave their probabilities to the artificial neural network.
- Using an artificial neural network was crucial for getting performance gains as just simply combining the model outputs was not enough to improve matters.

Technique	F (F1)	S (F1)	D (F1)	B (F1)	Q (F1)	Macro F1	Accuracy %
BERT	0.82	0.93	0.91	0.75	0.96	0.8740	89.44
Wav2Vec2	0.61	0.80	0.34	0.54	0.39	0.5354	67.88
Ensemble	0.82	0.94	0.91	0.76	0.96	0.8765	90.18
Baseline	0.57	0.92	0.82	0.69	0.71	0.6904	80.05



RESULTS

- The incorporation of an acoustic model allows for more reliable classification of short utterances, increasing overall performance.
- This approach improves upon previous attempts of utilising both the lexical and acoustic information for the MRDA dataset.
- We used ANN classifier in an ensemble to combine the outputs from the text and audio models.
- We compared our ANN classifier with a baseline weak classification method that showed an ANN was necessary.
- We also investigated the performance on a word count-basis and found that the two models complement each other for different utterance lengths.
- Combining both models allowed for the more accurate classification of dialogue acts that would have not been classified reliably by text-based alone.
- Text-based models perform very well but to improve further innovations such as using audio will need to be used.
- Our numerical results show that the accuracy and other metrics improved from combining the two models using an artificial neural network.

ACCURACY AND F-SCORE BY WORD COUNT