Short Utterance Dialogue Act Classification Using a Transformer Ensemble

Harry Maltby, Julie Wall, Tracy Gooodluck Constance, Mansour Moniri
University of East London, London, UK
harrym906@gmail.com

Cornelius Glackin, Marvin Rajwadi, Nigel Cannings
Intelligent Voice Ltd., London, UK

Abstract

An influx of digital assistant adoption and reliance is demonstrating the significance of reliable and robust dialogue act classification techniques. In the literature, there is an over-representation of purely lexical-based dialogue act classification methods. A weakness of this approach is the lack of context when classifying short utterances. We improve upon a purely lexical approach by incorporating a state-of-the-art acoustic model in a lexical-acoustic transformer ensemble, with improved results when classifying dialogue acts in the MRDA corpus. Additionally, we further investigate the performance on an utterance word-count basis, showing classification accuracy increases with utterance word count. Furthermore, the performance of the lexical model increases with utterance word length and the acoustic model performance decreases with utterance word count, showing the models complement each other for different utterance lengths.

Index Terms: Dialogue act classification, ensemble, NLP, transformers, lexico-acoustic models

1 Introduction

There is a surge in the use of digital assistants in our homes. As device hardware and capabilities have improved in recent years, it has resulted in people, especially the elderly and disabled, becoming more reliant on them for everyday purposes. An essential requirement for an AI agent is to reliably understand a user's intended request and distinguish between a command and a question. This task is known as Dialogue Act (DA) classification and is essential to comprehending communication as it exposes the intent of the requests [1]. Detecting questions from speech is difficult for machines as automatic speaker recognition (ASR) transcripts lack punctuation, relying solely on lexical information, leading to ambiguity. Acoustic information can aid in resolving ambiguity, as demonstrated by Ortega [2]. For example, in the case of a Declarative Question in the text without question marks, such as **'this is your car(?)'**, determining whether it is a statement or a question requires additional contextual information. Our study shows that combining lexical and acoustic information can enhance DA classification. We use predictions from both a lexical and an acoustic model to improve overall performance. Our proposed transformer ensemble, which consists of BERT, Wav2Vec2, and a multi-layer perception (MLP), fuses lexical and acoustic features in a cascaded process for DA classification. By combining the strengths of both models and addressing their weaknesses, our ensemble approach achieves better results than either single model alone, as demonstrated on the MRDA corpus.

2 Related Work

DA classification is the task of classifying an utterance to capture the intent of the communication, i.e. whether it is a question, statement, etc [3]. The most used publicly available benchmark data sets for this task are Switchboard Dialogue Act (SwDA) and the Meeting Recorder Dialogue Act (MRDA) corpus. The MRDA corpus was first described by Shriberg in 2004 [4]. The corpus was later revised and relabeled into 5 classes, as described by Ang in 2005 [5], where the data set was separated into the labels: *backchannel*, *disruption*, *floorgrabber*, *question* and *statement*. Ang et al. note that previous work of the time by other authors encountered many challenges due to the data set containing multiple speakers, frequent speaker overlap and high rates of speaker interruption. These challenges hindered the effectiveness of traditional techniques of the time to perform tasks such as DA classification and speaker diarisation. Overall, there have been several approaches to Dialogue Act (DA) classification, Li et al. and Colombo et al. [6] both utilize neural network models to classify DAs, with Colombo et al. specifically using a novel guided attention mechanism and hierarchical encoder to capture context-dependent correlations without handcrafted features and Li et al using a recurrent neural network. Chen et al. [7] introduce a CRF attentive structured network, while Raheja et al.

[8] propose a context-aware self-attention mechanism with a hierarchical recurrent neural network. The work of Ortega et al. is among the few that have explored the effectiveness of acoustic features and found that they can be effective when lexical information is insufficient or strong lexical indicators are absent. However, their analysis also revealed the heavy reliance on question marks for question classification, demonstrating the need for acoustic information in the classification process. Recent works propose neural and attention-based models for DA classification in spoken dialogues. However, only a few have explored the effectiveness of acoustic information in DAs. Research for this task appears to be converging and maturing into the use of hierarchical techniques to capture context correlation data within the lexical data of the data set. We build upon this and further explore the efficacy of using the acoustic portion of the data set to augment the performance of mature lexical techniques for DA classification.

3 Methodology

3.1 Data Set

The experiments were performed on the MRDA corpus [5], and the data set split was performed on the 5-class version of the MRDA corpus. The 5 classes are Statement (S), Question (Q), Floorgrabber (F), Backchannel (B), and Disruption (D). The MRDA corpus is an audio data set consisting of 75 meetings (each roughly an hour in duration) with 53 unique speakers (an average of about 6 speakers per meeting), mainly male speakers (40 male and 13 female speakers). We used the processing utilities from [9] and followed the guidelines outlined by the authors in [4]. The train, validation and test splits are split meeting-wise according to these guidelines. Of the 75 meetings, the guidelines suggest the omission of the last 2 meetings which has been followed.

3.2 Model and Rationale

The rationale behind the inclusion of both lexical and acoustic information is that often words can lexically look the same but belong to different linguistic classes when punctuation is absent, an example is **'Yeah(?)'** which could belong to any of the 5 aforementioned classes. The ensemble combines both a BERT lexical and Wav2Vec2 acoustic model to improve DA classification, shown in Figure 1. The models' output probabilities are concatenated and fed into an MLP classifier. The approach can utilize both lexical and acoustic information to enhance performance. The lexical model bert-base-uncased [10] was fine-tuned on the MRDA corpus, the acoustic model, Facebook's XLRS_Large_960_self Wav2Vec2 [11] was pre-trained on Libri-Light/Librispeech. The audio was segmented into speaker utterances from the 75 meetings using timestamped data.

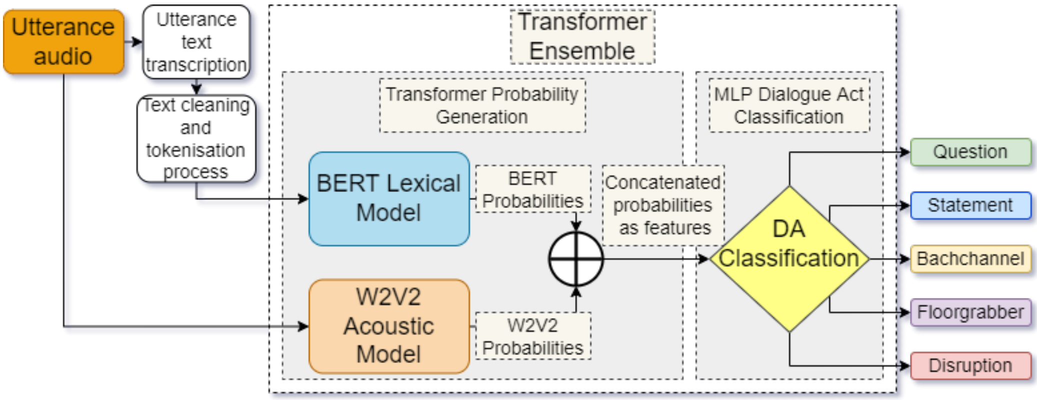

Figure 1: Transformer ensemble classification architecture.

4 Experiments and Results

Our BERT model achieved 89.44% accuracy and 0.87 F1 score, while the Wav2Vec2 model provided 67.88% accuracy and 0.54 F1 score shown in Table 1. Our MLP ensemble outperformed both models, increasing accuracy by 0.74% and F-score by 0.25. The MLP results are comparable to Ortega et al. while providing a baseline for the efficacy of the MLP classifier within the ensemble as shown in Table 2. When samples are aggregated by utterance word count and plotted as a grouped bar chart we can see a positive trend in the F1 score for the BERT classifier as the word count increases suggesting a positive correlation. Likewise, we can also see that the F1 score for the Wav2Vev2 model decreases as word count increases suggesting a negative trend both shown in Figure 2. This information falls in line with our hypothesis of lexical models performing better for longer utterances and acoustic models performing better for shorter utterances. It can also be observed that the MLP ensemble gained the most from combining lexical and acoustic when classifying short utterances as the greatest difference between bars between the BERT and the MLP ensemble was in the 1-2 word utterance category.

Table 1: Model performance metrics

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

Table 2: Performance comparison table

Author	Accuracy	Features	Technique
Chapuis et al (2021) [12]	92.4%	Lexical	Bespoke hierachical transformer encoder
Li et al (2019) [1]	92.2%	Lexical	CRF
Chen et al (2017) [7]	91.7%	Lexical	CRF
Colombo et al (2020) [6]	91.6%	Lexical	Hierarchical BiGRU Seq2Seq
Raheja et al (2019) [8]	91.1%	Lexical	Hierarchical Dual-Attention RNN CRF
Ortega et al (2018) [2]	84.7%	Lexico-Acoustic	CNN LSTM + CNN
Ortega et al (2018) [2]	84.1%	Lexical	CNN LSTM
Ortega et al (2018) [2]	67.8%	Acoustic	CNN
Our system	90.18%	Lexico-Acoustic	Transformer Ensemble
Our system	89.44%	Lexical	BERT
Our system	67.88%	Acoustic	Wav2Vec2

5 Conclusions

To summarise, the analysis of the performance of BERT, Wav2Vec2, and an MLP ensemble for the task of DA classification using both lexical and acoustic features showed that there is potential for performance improvement by exploiting both features. The results showed that BERT performs better on longer utterances, while Wav2Vec2 performs better on shorter ones. The MLP ensemble performed better than any single model, especially for short utterances. The negative correlation between the output of BERT and Wav2Vec2 suggests that they complement each other as an ensemble. Overall, the results suggest that there is room for further research to fully exploit both lexical and acoustic features in this task.

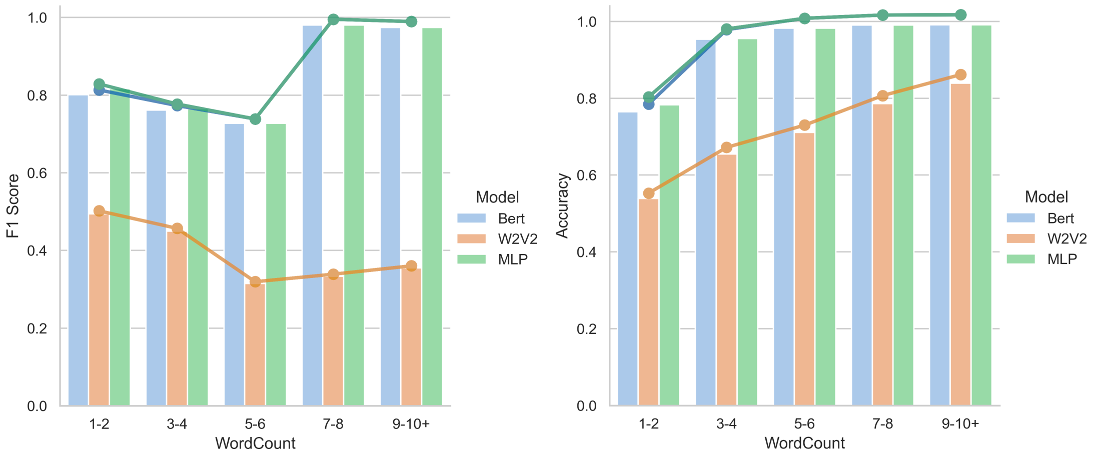

Figure 2: Per-model accuracy and F-score on sets of utterances grouped by word-count.

References

[1] R. Li et al. “A dual-attention hierarchical recurrent neural network for dialogue act classification”. In: *arXiv preprint arXiv:1810.09154* (2018).

[2] D. Ortega and Ngoc Thang Vu. “Lexico-acoustic neural-based models for dialog act classification”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 6194–6198.

[3] J.L. Austin. *How to do things with words*. Oxford University Press, 1975.

[4] E. Shriberg et al. *The ICSI meeting recorder dialog act (MRDA) corpus*. Tech. rep. International Computer Science Inst. Berkeley CA, 2004.

[5] J. Ang, Yang Liu, and Elizabeth Shriberg. “Automatic dialog act segmentation and classification in multiparty meetings”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. 2005, pp. I–1061.

[6] P. Colombo et al. “Guiding attention in sequence-to-sequence models for dialogue act prediction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7594–7601.

[7] Z. Chen et al. “Dialogue act recognition via CRF-attentive structured network”. In: *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018, pp. 225–234.

[8] V. Raheja and Joel Tetreault. “Dialogue act classification with context-aware self-attention”. In: *arXiv preprint arXiv:1904.02594* (2019).

[9] N. Duran. *Utilities for Processing the Meeting Recorder Dialogue Act Corpus*. 2018. URL: `https://github.com/NathanDuran/MRDA-Corpus`.

[10] T. Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[11] A. Baevski et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12449–12460.

[12] E. Chapuis et al. “Hierarchical pre-training for sequence labelling in spoken dialog”. In: *arXiv preprint arXiv:2009.11152* (2020).