Dual-Task Dialogue Understanding

Sibgha Anwar, Nirmalie Wiratunga and Mark Snaith

School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen AB10 7GJ, Scotland, UK

Abstract

In dialogue systems, utterances do not occur in isolation. One conversation might involve interactions between several speakers. It's crucial to determine the intentions behind utterances in multi-party conversations when more than two interlocutors are interacting. Beyond directly capturing the speaker's intention, our proposed model first focuses on identifying speakers from utterances, and based on this knowledge, it classifies the corresponding dialogue acts. For the speaker identification process, the study extracted linguistic features related to speakers from conversations and incorporated them during the fine-tuning process, which is particularly beneficial in dealing with multiple speakers. After that our model aims to improve dialogue act recognition baselines on shorter utterances by implementing a pipe-lining approach based on speaker model predictions. The effectiveness of our approach is demonstrated using two benchmark datasets, MRDA and SwDA, which are based on multiparty and twofold conversations, respectively.

Keywords

Speaker identification, dialogue act recognition, dual-task learning, conversational structure learning

Conversations, both written and verbal, are crucial for human communication. Speaker identification (SI) and dialogue act recognition (DAR) are essential tasks for understanding spoken language, identifying speakers, and facilitating human-computer interaction applications. SI and DAR have historically been considered separate tasks in natural language processing (NLP) and speech processing [1, 2, 3]. Some contributions to the field of SI have been made by [3, 4] while [5, 6] conducted significant research on dialogue act recognition. The goal is to identify each speech in a conversation based on its communicative function, such as question, statement, command, or request, which aids in understanding conversational flow and anticipating future exchanges.

The literature identifies several major challenges with dialogue act recognition that require improvement. Firstly, dialogue act models hugely rely on statistical patterns rather than speaker-specific traits, predicting actions based on similar sequences but not considering unique speaking styles [7, 8]. Secondly, as dialogue act recognition algorithms generalise across all speakers without taking details into account, they frequently provide inaccurate classifications and fail to correctly detect specific speakers' speech patterns or personal styles [9, 10]. Finally, most current methods often see conversation actions as discrete categories, which may overlook the subtle variations in how various speakers convey identical intents and the pragmatic implications [11, 12].

Research shows that speaker identification is a crucial aspect of dialogue act recognition systems, enabling personalised recognition and distinguishing between orders, enquiries, and assertions based on speaker-specific patterns [13]. It is especially helpful for unclear utterances such as "Really?" and aids in identifying and adjusting to unusual or non-standard dialogue behaviours. Inaccurate categorisation and disruption of discourse can result from mis-identification. SI tracks conversational roles and interactions, which preserves dialogue flow, improves turn-taking modelling, and increases DAR accuracy [10, 14]. Table 1 displays conversation snippets tagged with speaker IDs and dialogue acts from the MRDA corpus.

The literature research suggests that combining SI and DAR can enhance conversational flow, particularly in multiparty interactions. However, the challenge of simultaneously identifying the speaker and their intent has not been adequately addressed in literature. Therefore our research combines SI models with DAR to tackle dialogue act recognition challenges, forming the basis for addressing following research questions.

SICSA REALLM Workshop 2024

s.anwar3@rgu.ac.uk (S. Anwar)
0000-0001-5839-5314 (S. Anwar); 0000-0003-4040-249 (N. Wiratunga); 0000-0001-9979-9374 (M. Snaith)

Table 1
Example conversation snippet annotated by dialogue acts from the MRDA corpus.

Speaker	Utterance	DA
me012	Who would be the subject of this trial run?	Question
mn015	Pardon me?	Request
me012	Is one of you going to be the subject?	Question
mn015	Liz volunteered to be the first subject, which might be even better than us.	Statement
fe004	Good.	Agreement
me003	One of us.	Acknowl-edgement

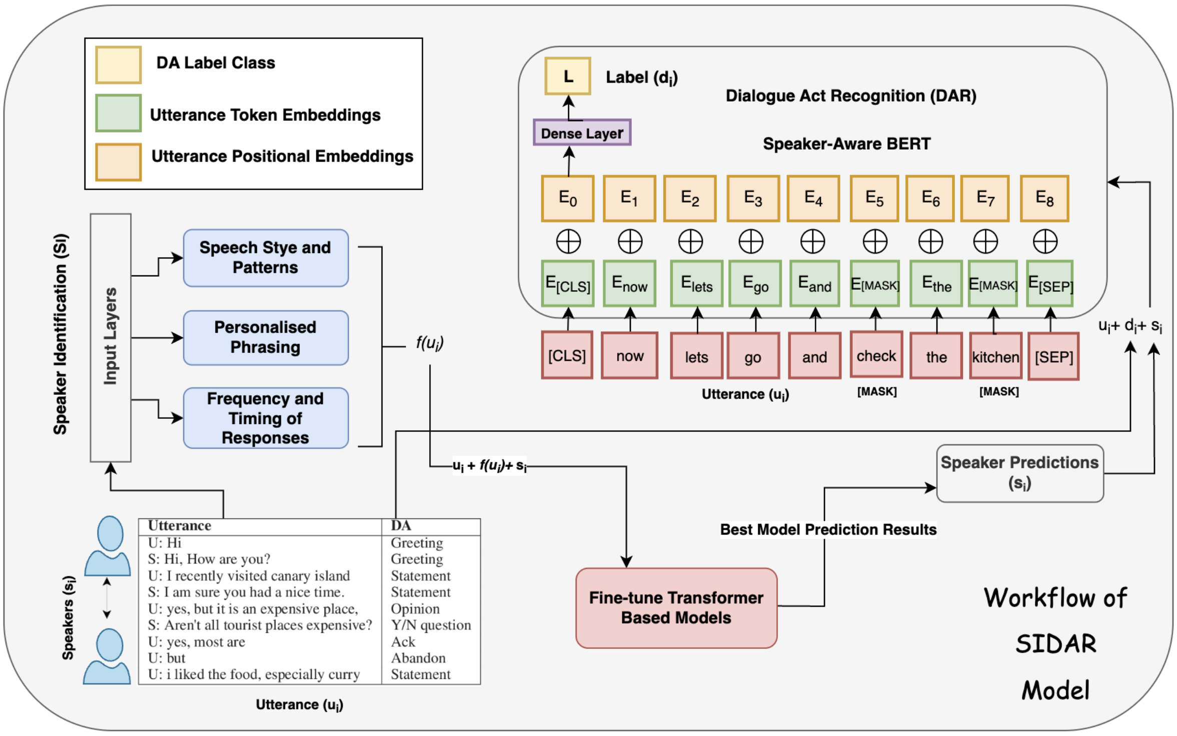
- What is the most effective method to properly combine speaker identification and dialogue act recognition into a single framework to improve our overall understanding of multiparty conversations?
- When speaker-specific data such as speech patterns and styles, frequency, response time and personalised phrasing are included, how does the accuracy of dialogue recognition systems improve?
- How does the proposed system handle imprecise speaker transitions such as overlapping speech and shifting topics unexpectedly in real-world multi-party interactions?
- What strategies may be used to guarantee this dual-task system's performance and scalability in intricate multiparty interactions across a range of conversational contexts?

1. Related Work

In natural language processing (NLP), speaker identification (SI) and dialogue act recognition (DAR) are important research fields [5]. Early speaker identification techniques primarily focused on linguistic information derived from speech transcriptions. Discourse patterns, grammatical structures, and semantic content are key indicators that provide valuable insights into the linguistic preferences of specific speakers. The study [15] suggests classifying film dialogue speakers based on discrete stylistic features using the K Nearest Neighbour Algorithm, Naive Bayes Classifier, and Conditional Random Field [3, 16, 17]. The approaches were difficult to handle a variety of language styles and complex transcribing conditions, even if they worked well in controlled settings. Due to their thorough contextual representations, pre-trained language models such as BERT and RoBERTa have shown success in speech processing and conversational tasks [18, 19].

Traditionally, DAR relied on statistical models and rule-based systems, including Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), to categorise dialogue acts according to lexical and syntactic aspects [5, 20]. Deep learning algorithms Improve DAR Accuracy by using LSTM networks to enhance the representation of ambiguity in real-world conversations. when multiple participants are involved [21]. The recent advancements such as transformer-based models such as BERT and RoBERTA enhance conversational relationships by adjusting dialogue act recognition and enhancing performance when paired with dialogue-specific data such as dialogue history and speaker information [1, 22]. But these models have drawbacks, especially when it comes to shorter utterances, when poor performance is caused by inadequate context [1]. Furthermore, models that just use pre-trained embeddings to identify dialogue actions frequently ignore the entirety of the conversational context and speaker-specific information. The model's inability to differentiate between dialogue acts is hindered by this lack of contextual richness and customisation, which may make it more difficult to comprehend and interpret the conversation's intended meaning. Adding speaker-specific data can help to improve dialogue act recognition by giving the required context.

The integration of SI with DAR has not received much attention since it relies heavily on auditory features and is not immediately relevant to text transcriptions [23]. For text-based, multiparty conversation contexts, additional research is required [24]. Recent works have investigated the use of speaker



Utterance	DA
U: Hi	Greeting
S: Hi, How are you?	Greeting
U: I recently visited canary island	Statement
S: I am sure you had a nice time.	Statement
U: yes, but it is an expensive place,	Opinion
S: Aren't all tourist places expensive?	Y/N question
U: yes, most are	Ack
U: but	Abandon
U: i liked the food, especially curry	Statement

Figure 1: Workflow of the proposed dual-task system for dialogue understanding, integrating Speaker Identification and Dialogue Act Recognition (SIDAR). The method improves both tasks' accuracy by utilising conversational context and speaker-specific features, which makes it possible to understand multiparty conversations more effectively.

embeddings in dialogue act models, offering a basis for increasing DAR through speaker identification. Study conducted in [25] indicates that discourse structure has an important role in understanding utterance purpose, enhancing model performance, and recognising dialogue acts. Recent studies have also explored techniques involving discourse structure analysis and speaker identification in dialogue act recognition [26]. Therefore, our study intends to create complex, context-aware chat systems by utilising discourse structure and speaker identity to increase dialogue act recognition accuracy and coherence.

2. Datasets

The dual-task learning method will be tested on two publicly available datasets to demonstrate its reliability in accurately identifying speakers regardless of their complexity or speaking style. This study uses two datasets: the ICSI Meeting Recorder Dialogue Act (MRDA) [27] and the Switchboard Dialogue Act (SwDA) [28], which contain over 180,000 real-world meeting conversation utterances and 223605 utterances from phone talks between two speakers on a predetermined topic, respectively, to analyse academic and professional meetings.

3. Proposed Approach

The proposed method aims to enhance dialogue act recognition by integrating speaker identification into the conversational analysis pipeline. This dual-task learning approach addresses the drawbacks of existing models in handling intricate conversational contexts by integrating speaker identification with dialogue act recognition to better capture the link between discourse purpose and speaker identity. Following are the major elements of the our proposed methodology. The Figure 1 illustrates the overall workflow and how different components are linked in the proposed SIDAR model.

3.1. Speaker Identification (SI) Model

The Speaker Identification (SI) model plays a major role in our dual-task method by enhancing contextual awareness in multiparty conversations. It enhances the identification process by capturing distinct speech and behaviour patterns by adding speaker-specific information.

3.1.1. Features of SI Model

The SI model improves interactions between multiple speakers in text-based translations by incorporating speaker-specific features that influence communication styles. We aim to utilise the following features.

- **Speech Style and Patterns:** Each speaker uses different syntactic patterns, repeats particular phrases, and builds utterances in different ways. Through the integration and application of these patterns, the model enhances its ability to identify dialogue acts, hence improving its understanding of the speaker's communication style and context [20, 29].

- **Personalised Phrasing:** "Would you mind" and "Can you" are examples of spoken words and phrases that speakers frequently employ. The model is able to accurately predict the meaning of an utterance by identifying personalised phrases since some words, such as requests, directions, and queries, are suggestive of certain conversation activities [3, 4].

- **Frequency and Timing of Responses:** The SI model analyses the frequency of a speaker's replies as well as their timing, including start and end timings, in order to comprehend their response in the conversation. Faster answers demonstrate interaction, but slower responses imply more complex contributions. People who take longer to reply, for instance, could be writing more intricate or in-depth remarks, such long suggestions. Overall the SI model enhances the DAR model's performance by understanding the speaker's identity and the correlation between response timing and the dialogue act [4, 30].

3.1.2. SI Model Architecture

Advanced pre-trained language models, such as DeBERTa, RoBERTa, BART and Llama, are used in this work to identify the speaker in text-based transcriptions; these models are perfect for conversations that vary widely in context. To simulate long-range interactions in conversation text, these models, according to [22, 24], specifically account for effective speaker recognition and gather extensive contextual information. The research aims to increase speaker identification models' performance by utilising contextual factors from these models and the extracted linguistic elements. Byte-pair encoding (BPE) adds context by breaking utterances into subword units and assigning each token to an embedding vector. Speaker embeddings recognise distinct speech patterns, whereas position embeddings preserve word order in utterance [23, 24].

Furthermore, position embeddings identify speaker transitions while preserving the utterances' sentence structure by tracking the conversation's flow. Large numbers of speakers are a challenge for traditional speaker identification techniques such as tf-idf vectors, speaker tokens, and word2vec [2, 3, 29, 31, 32]. But even in comparable utterances, our work makes use of speaker embeddings to improve the model's ability to distinguish between speakers in intricate multiparty conversations [24, 29]. Therefore, as stated in Section 2, conversational datasets that highlight multiple speakers in various contexts are used to train the DAR model. By learning both speaker identification and dialogue act recognition simultaneously, it maximises their effectiveness. The dual-task technique leverages speaker-specific traits for enhanced recognition in scenarios when the speaker identification is crucial to the discourse [25, 30]. By focusing on pertinent segments of the input dialogue utterance, attention mechanisms are also intended to be used to dynamically balance the importance of conversational segments [26, 33].

3.2. Dialogue Act Recognition (DAR) Model

DAR models categorise speech in conversations based on communicative goals such as inquiry, statement, or order [5]. The most advanced models capture linguistic and contextual features by using pre-trained language models such as BERT or RoBERTa. Transformer ensembles using lexical-based techniques (BERT) have been developed recently as a result of advancements in spoken language processing; however, these models frequently perform inefficiently on shorter utterances. For instance, the utterance "Sure" in a customer support chat system can signify agreement, acknowledgement, or confirmation [1]. Input representations, such as word embeddings, positional embeddings, and speaker-specific data, are used to analyse data. Moreover, transformer encoders decode the input data, while multi-head self-attention mechanisms understand conversation progression and connections between dialogue turns.

Traditional models often face challenges due to short utterances lacking context. We often include speaker-specific embeddings from speaker identification models, such as DeBERTa, RoBERTa, BART and Llama, to get over the limitation and enhance the model's ability to handle brief or unclear utterances, particularly in multiparty interactions [34]. With the aid of these personalised embeddings, the model is better able to understand speaker behaviour, including patterns of reaction from customers. By clarifying the goal and making speech acts simpler to identify, this enhances interpretation and interaction patterns. The model's last layer improves overall accuracy in real-world conversational scenarios by predicting dialogue actions based on the speaker's location and the substance of the utterance. In order to improve dialogue act recognition for shorter utterances, our SIDAR models will identify the speakers first, perhaps providing additional information to understand unique speech patterns and conversational styles.

4. Conclusion and Future Work

In this work, we have presented a proposed methodology that integrates SI and DAR in a dual-task learning approach to improve conversational flow precision. We suggest using state-of-the-art language models such as DeBERTa, RoBERTa, BART and Llama in light of the shortcomings of traditional speaker identification techniques and the limitations with BERT models have when recognising dialogue acts, especially with shorter utterances. By including speaker-specific information and conversational history into the dialogue act recognition process, our methodology aims to improve upon the inadequacies of current techniques. This research provides a conceptual framework; however, further work will need to be done to put our suggested approaches into practice and validate them through experimentation. The study findings are expected to have a substantial impact on the domains of dialogue act recognition and speaker identification, which will eventually improve the efficacy of conversational AI systems.

5. Acknowledgements

The authors would like to thank Robert Gordon University which supported this work through a funded PhD studentship.

References

[1] H. Maltby, J. Wall, T. Goodluck Constance, M. Moniri, C. Glackin, M. Rajwadi, N. Cannings, Short utterance dialogue act classification using a transformer ensemble, UA-DIGITAL 2023: UA Digital Theme Research Twinning (2023).

[2] D. Holmer, L. Ahrenberg, J. Monsen, A. Jönsson, M. Apel, M. B. Grimaldi, Who said what? speaker identification from anonymous minutes of meetings, in: The 24rd Nordic Conference on Computational Linguistics, 2023.

[3] K. Ma, C. Xiao, J. D. Choi, Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks, in: Proceedings of ACL 2017, Student Research Workshop, 2017, pp. 49–55.

[4] S. Salim, S. Shahnawazuddin, W. Ahmad, Automatic speaker verification system for dysarthric speakers using prosodic features and out-of-domain data augmentation, Applied Acoustics 210 (2023) 109412.

[5] V. Raheja, J. Tetreault, Dialogue act classification with context-aware self-attention, arXiv preprint arXiv:1904.02594 (2019).

[6] Y. Si, L. Wang, J. Dang, M. Wu, A. Li, A hierarchical model for dialogue act recognition considering acoustic and lexical context information, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7994–7998.

[7] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, M. Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech, Computational linguistics 26 (2000) 339–373.

[8] T. Saha, S. Srivastava, M. Firdaus, S. Saha, A. Ekbal, P. Bhattacharyya, Exploring machine learning and deep learning frameworks for task-oriented dialogue act classification, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[9] M. Kim, H. Kim, Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances, Pattern recognition letters 101 (2018) 1–5.

[10] A. Qamar, A. Pyarelal, R. Huang, Who is speaking? speaker-aware multiparty dialogue act classification, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 10122–10135.

[11] C. Sun, L.-P. Morency, Dialogue act recognition using reweighted speaker adaptation, in: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2012, pp. 118–125.

[12] A. Enayet, G. Sukthankar, An analysis of dialogue act sequence similarity across multiple domains, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3122–3130.

[13] Z. He, L. Tavabi, K. Lerman, M. Soleymani, Speaker turn modeling for dialogue act classification, arXiv preprint arXiv:2109.05056 (2021).

[14] P. Żelasko, R. Pappagari, N. Dehak, What helps transformers recognize conversational structure? importance of context, punctuation, and labels in dialog act recognition, Transactions of the Association for Computational Linguistics 9 (2021) 1163–1179.

[15] A. Kundu, D. Das, S. Bandyopadhyay, Speaker identification from film dialogues, in: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), IEEE, 2012, pp. 1–4.

[16] R. Lowe, N. Pow, I. Serban, J. Pineau, The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, arXiv preprint arXiv:1506.08909 (2015).

[17] M. K. Singh, S. Manusha, K. Balaramakrishna, S. Gamini, Speaker identification analysis based on long-term acoustic characteristics with minimal performance, International Journal of Electrical and Electronics Research 10 (2022) 848–852.

[18] C. S. Xia, Y. Wei, L. Zhang, Practical program repair in the era of large pre-trained language models, arXiv preprint arXiv:2210.14179 (2022).

[19] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2023) 1–40.

[20] H. Kumar, A. Agarwal, R. Dasgupta, S. Joshi, Dialogue act sequence labeling using hierarchical encoder with crf, in: Proceedings of the aaai conference on artificial intelligence, volume 32, 2018.

[21] C. Bothe, C. Weber, S. Magg, S. Wermter, A context-based approach for dialogue act recognition using simple recurrent neural networks, arXiv preprint arXiv:1805.06280 (2018).

[22] G. Guaquiere, P. ENSAE, A. N. T. SON, Roberta vs bert for intent classification (2021).

[23] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech communication 52 (2010) 12–40.

[24] Z. Jia, Y. Shi, W. Liu, Z. Huang, X. Sun, Speaker-aware interactive graph attention network for emotion recognition in conversation, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2023) 1–18.

[25] Z. Shi, M. Huang, A deep sequential model for discourse parsing on multi-party dialogues, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 7007–7014.

[26] C.-J. Peng, Y.-J. Chan, C. Yu, S.-S. Wang, Y. Tsao, T.-S. Chi, Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2021, pp. 1–5.

[27] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, H. Carvey, The icsi meeting recorder dialogue act (mrda) corpus, in: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004, 2004, pp. 97–100.

[28] D. Jurafsky, E. Shriberg, Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13 daniel jurafsky*, elizabeth shriberg+, and debra biasca** university of colorado at boulder &+ sri international (1997).

[29] M.-Q. Nghiem, N. Roberts, D. Sityaev, Speaker role identification in call centre dialogues: Leveraging opening sentences and large language models, in: Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, 2023, pp. 388–392.

[30] R. Le, W. Hu, M. Shang, Z. You, L. Bing, D. Zhao, R. Yan, Who is speaking to whom? learning to identify utterance addressee in multi-party conversations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 1909–1919.

[31] D. Baum, Recognising speakers from the topics they talk about, Speech Communication 54 (2012) 1132–1142.

[32] E. Ekstedt, G. Skantze, Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog, arXiv preprint arXiv:2010.10874 (2020).

[33] D. Bahdanau, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.