Leveraging BERT, Context, and Syntactic Features for Enhanced Speaker Identification

Meng-Chiao (Joe) Wang
wangjoe@berkeley.edu

Abstract

This project tackles the challenge of speaker identification in the TV series Friends by integrating a fine-tuned BERT model with contextual utterance windows and syntactic features derived from Part of Speech (POS) tags. By merging syntactic information with BERT's robust contextual embeddings, the model effectively captures long-range dependencies within dialogues. This approach surpasses the existing state-of-the-art multi-document CNN-based model by 9 percentage points in accuracy (39.74% vs. 31.06%), demonstrating its superior performance in accurately identifying speakers in multiparty conversations.

1 Introduction

Speaker identification in multiparty dialogues is a fundamental task in Natural Language Processing (NLP) with diverse applications, including relationship analysis, power dynamics assessment, dialogue summarization, and enhancing accessibility through automatic subtitle tagging in videos. Additionally, accurate speaker identification can be instrumental in legal contexts, such as prosecutorial investigations or detective work.

This project focuses on identifying speakers in the TV series Friends by leveraging a novel approach that integrates a fine-tuned BERT model with syntactic parsing information, including POS tags from spaCy. Furthermore, the model incorporates contextual information by considering prior and subsequent utterances surrounding each target utterance. This combination aims to enhance the model's ability to capture both the semantic nuances and grammatical structures inherent in conversational dialogues.

This approach achieved a 39% F1-score and 40% accuracy, outperforming the baseline multi-document CNN-based model by 9 percentage points in each metric. These results underscore the effectiveness of integrating contextual and syntactic features with advanced transformer-based models for improved speaker identification in multiparty conversations.

2 Related Work

Speaker identification in dialogues has been explored using various machine learning models. Prior work (Ma et al., 2017) introduced a multi-document CNN-based model for text-based speaker identification in multiparty conversations. While effective, their approach does not incorporate syntactic information, potentially limiting its ability to capture grammatical nuances and handle longer-range dependencies within dialogues. Additionally, such models may struggle with uncommon words, colloquialisms, or intentional misspellings present in transcripts.

Transformer-based architectures have shown significant promise in NLP tasks due to their ability to model contextual relationships through self-attention mechanisms. Other prior work (Li and Choi, 2020) applied SpanBERT, a variant of BERT tailored for span-based tasks, to the Friends dataset for a question answering task. However, their focus was not on speaker identification, leaving a gap in leveraging transformer models for this specific application.

By combining BERT's subword tokenization capabilities (Devlin et al., 2019) with contextual utterance windows and syntactic parsing information, this model aims to enhance the accuracy and robustness of speaker identification in multiparty dialogues, setting it apart from previous works that either lack syntactic integration or do not utilize transformer-based models for this task.

3 Methods

3.1 Dataset and Preprocessing

The dataset utilized in this project was sourced from publicly available transcripts of the TV series

Speaker	Training	Validation	Test
Other	**6294**	894	979
Ross Geller	5604	822	984
Rachel Green	5470	**1093**	**1086**
Chandler Bing	5431	849	677
Monica Geller	5183	907	812
Joey Tribbiani	4774	939	909
Phoebe Buffay	4451	810	773

Table 1: Utterance distribution by speaker across data sets.

Friends, provided by the Emory Character Mining Project on GitHub. The initial dataset comprised 67,373 utterances. To streamline the data for speaker identification, all speakers except the six main characters—Ross Geller, Rachel Green, Chandler Bing, Monica Geller, Joey Tribbiani, and Phoebe Buffay—were categorized under an "Other" label. This grouping was essential to focus the model on distinguishing among the primary characters while managing the diversity of additional speakers.

Subsequent preprocessing involved the removal of utterances with NULL entries, reducing the dataset by 6,063 utterances and resulting in a refined total of 61,310 utterances. To ensure consistency and comparability with prior work, the dataset was partitioned based on episode numbers. Specifically, seasons 1 through 6 were allocated to the training set, encompassing 37,207 utterances. Season 7 was designated for validation, containing 6,314 utterances, while season 8 was reserved for testing, comprising 6,220 utterances. Utterances from seasons 9 and 10, totaling 11,569, were excluded as they were not utilized in Ma et al.'s experiment (Ma et al., 2017). The distribution of utterances across speakers in each dataset split is detailed in Table 1.

3.2 Baseline Models

To establish benchmarks for evaluating the effectiveness of the proposed approach, two baseline models were implemented: a Majority Class Classifier and a Fine-Tuned BERT Model. The Majority Class Classifier serves as a simple baseline by always predicting the most frequent class, which is "Other" in this dataset. This baseline helps in understanding the minimum performance threshold and the impact of class imbalance. The Fine-Tuned BERT Model represents a more sophisticated baseline, utilizing only BERT embeddings without incorporating syntactic features or contextual utterances. This model consists of a single hidden Dense layer, a Dropout layer, and a final Softmax layer, trained over five epochs with early stopping to prevent overfitting. By comparing our enhanced model against these baselines, we can quantitatively assess the contributions of contextual and syntactic integrations to speaker identification performance.

3.3 Feature Extraction

3.3.1 Contextual Utterances

To enrich the model's understanding of conversational dynamics, contextual utterances were incorporated by considering three preceding and three following utterances surrounding each target utterance. These contextual utterances were concatenated and separated by special tokens that denote their relative positions, such as [PREV_1] for the first preceding utterance and [NEXT_1] for the first following utterance. This structured input ensures that the model captures the flow and nuances of conversation without exceeding BERT's maximum token limit of 512. Consequently, the input sequence length was set to 500 tokens to accommodate these special tokens and provide a buffer for essential dialogue information. Empirical testing revealed that a window size of three utterances on either side provided the optimal balance between contextual richness and computational efficiency, as larger windows did not yield significant accuracy improvements and occasionally led to performance degradation.

3.3.2 Syntactic Features

In addition to contextual information, syntactic features were integrated to enhance the model's ability to distinguish between speakers based on grammatical styles. Syntactic information was extracted using spaCy's POS tags. These syntactic features were then transformed into numerical representations by assigning each POS tag a unique numerical identifier, which were subsequently embedded into dense vector spaces using an embedding layer, capturing the structural relationships and grammatical nuances within the dialogue. By encoding syntactic relationships, the model gains insights into each speaker's unique grammatical patterns, facilitating more accurate identification. As seen in Figure 2, these numerical syntactic features were subsequently fed into the classification

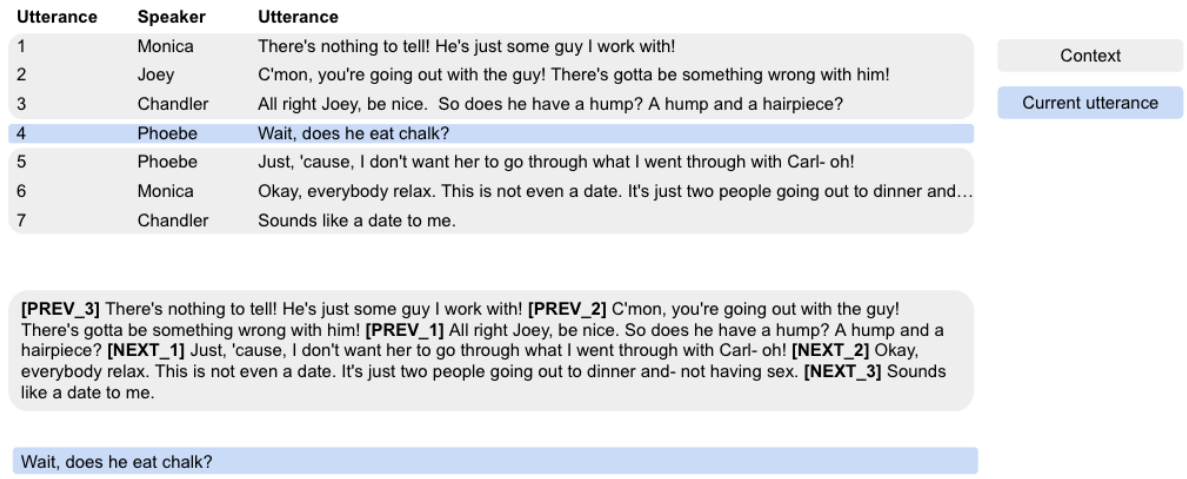

Figure 1: Sample "context" utterances with their corresponding "current" utterance.

model alongside BERT's contextual embeddings, allowing the model to leverage both semantic and syntactic information in making speaker predictions.

3.3.3 Model Architecture

The core architecture of the proposed speaker identification model leverages a fine-tuned BERT base model, enhanced with syntactic features and contextual utterances. The model begins with four input layers: input_ids, attention_mask, token_type_ids, and pos_tag_ids. The input_ids, attention_mask, and token_type_ids are fed into the BERT encoder (TFBertModel), which generates contextual embeddings capturing the semantic nuances of the dialogue. Concurrently, the pos_tag_ids are passed through an embedding layer to convert part-of-speech tags into dense vectors. These syntactic embeddings are then concatenated with BERT's contextual embeddings along the feature dimension, resulting in an 800-dimensional vector for each token. To distill this rich information, a Global Average Pooling layer averages the token vectors across the sequence length, producing a single fixed-size vector that summarizes the entire input. This pooled representation is then passed through a Dense layer with ReLU activation to learn complex patterns, followed by a Dropout layer with a rate of 0.3 to mitigate overfitting. Finally, a Softmax-activated Dense layer outputs probability distributions across seven speaker categories, enabling the model to classify each utterance accurately.

3.3.4 Training Procedure

All models were trained using consistent configurations to ensure fair comparisons. The training process utilized the Adam optimizer with an initial learning rate of 2e-5, which was reduced to 1e-5 in the final epoch to fine-tune the learning process. A batch size of 16 was selected to balance computational efficiency and model performance. Training was conducted over five epochs, with early stopping triggered if the validation loss did not improve for four consecutive epochs, thereby preventing overfitting. The loss function employed was Sparse Categorical Cross-Entropy, suitable for multi-class classification tasks. Additionally, class weights were applied to address the inherent class imbalance, particularly the predominance of the "Other" category. However, this adjustment yielded limited improvements, enhancing the F1 score for Rachel from 0.37 to 0.40 while decreasing the F1 score for "Other" from 0.47 to 0.42. This outcome suggests that while class weighting can mitigate some imbalance effects, further strategies may be necessary to achieve optimal performance across all speaker categories.

4 Results and Discussion

Our approach achieved 39% F1-score and 40% accuracy, outperforming the baseline by 9ppt for each metric. Starting with a fine-tuned BERT model, incorporating the three preceding utterances as contextual input provided a five percentage point boost to the model's accuracy, increasing it from 26.98% to 32.07%. Adding Part-of-Speech (POS) tags further enhanced performance by two percent-

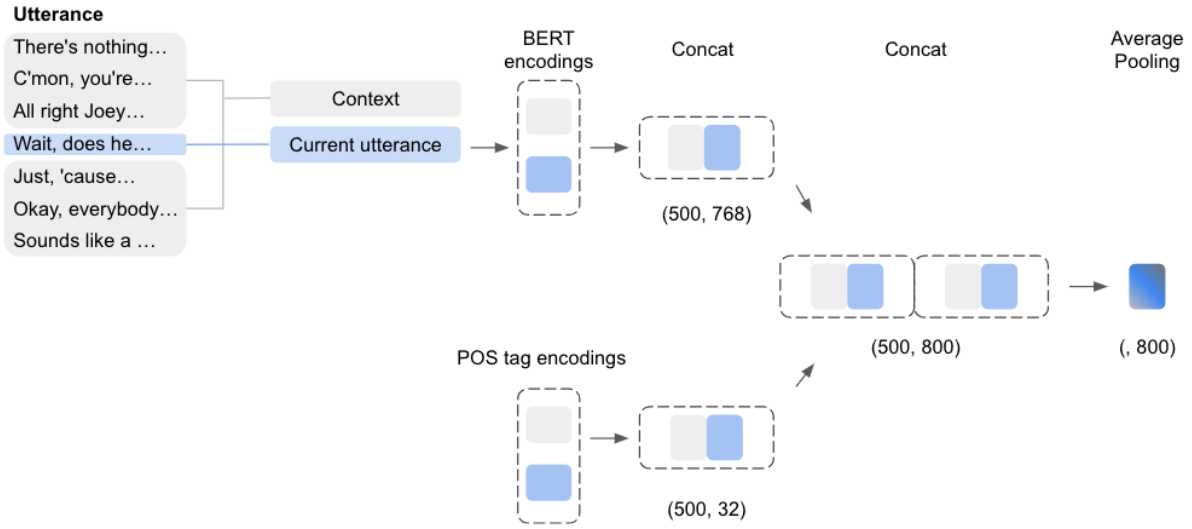

Figure 2: Model input architecture.

Model	Precision	Recall	F1	Accuracy
Multi-Document-CNN ((Ma et al., 2017)			29.72	31.06
Baseline (Majority Class)	2.25	14.29	3.89	15.74
Baseline (BERT)	28	27	27	26.98
BERT + context (-3)	34	32	31	32.07
BERT + context (-3) + syntactic	35	34	34	34.18
BERT + context (-3, EP) + syntactic	36	34	34	34.32
BERT + context (±3, EP) + syntactic	**41**	**39**	**39**	**39.74**

Table 2: Model performance

age points, raising accuracy to 34.18%. To reduce noise, contextual utterances were subsequently restricted to those within the same episode, resulting in a modest improvement of 0.14 percentage points, bringing accuracy to 34.32%. Analysis revealed that utterances at the beginning of episodes were frequently misclassified, likely due to insufficient prior context. Addressing this issue, three following utterances were included as additional context, which significantly boosted the model's accuracy to 39.74%.

Subgroup analysis reveals that the model excels in precision when predicting Ross as the speaker, achieving a 0.62 precision score. However, it struggles with recall for Ross, capturing only 0.34 of his actual utterances. Joey stands out with the highest F1 score of 0.48, correctly identifying nearly half of his utterances, which may indicate that Joey's speech patterns are more distinctive in content and grammatical structure. Notably, the model demonstrates a strong recall for the "Other" category (0.63), likely due to the larger number of "Other" utterances in the training dataset. The confusion matrices show that the model frequently misclassifies utterances from Rachel and Phoebe as "Other," suggesting a tendency to default to "Other" when uncertain about the speaker identity. Specifically, Rachel is often confused as "Other" in approximately 24% of cases, while Phoebe experiences similar confusion rates. This pattern indicates that interactions involving Rachel and Phoebe may be less distinctive, making them more susceptible to misclassification.

Monica seems to be the character being confused the most for non-"Other" speakers, with Chandler (18%) and Rahcel (17%) being the top two candidates for confusion. This could potentially reflect the high frequency of interaction between Monica, Chandler (Monica's romantic interest), and Rachel (Monica's long-time best friend), although it is interesting that the same is not true for Ross and Rachel. Overall, the model achieves an accuracy of 40%, with varying performance across different speakers, highlighting areas for potential improvement in distinguishing speakers with overlapping or less distinctive speech characteristics.

Speaker	precision	recall	f1-score	support
Chandler Bing	0.27	0.38	0.31	677
Joey Tribbiani	0.49	0.48	**0.48**	909
Monica Geller	0.33	0.27	0.3	812
Other	0.37	**0.63**	0.47	979
Phoebe Buffay	0.37	0.34	0.35	773
Rachel Green	0.44	0.33	0.37	**1086**
Ross Geller	**0.62**	0.34	0.44	984
accuracy				0.4
macro avg	0.41	0.39	0.39	6220
weighted avg	0.42	0.4	0.4	6220

Table 3: Model performance by speaker

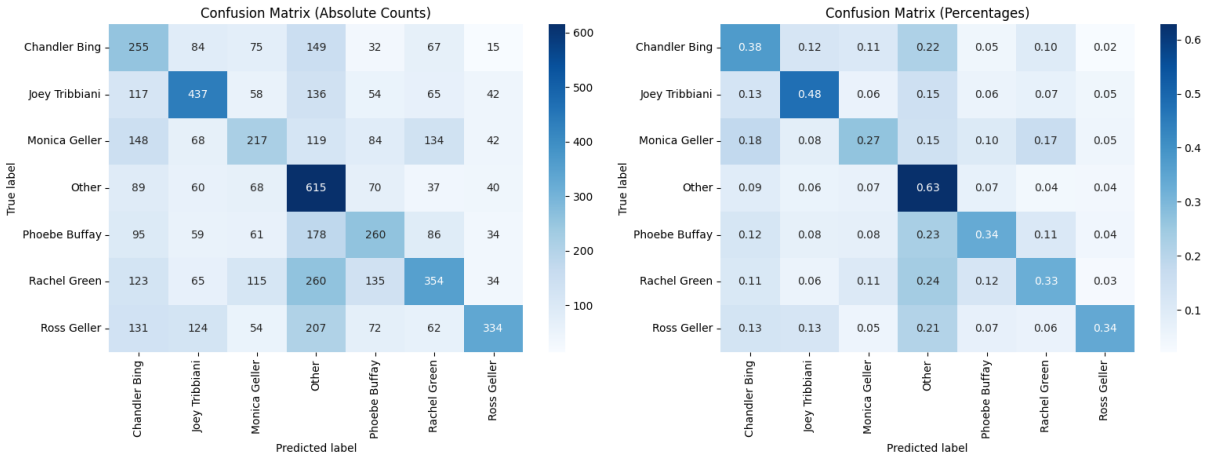

Figure 3: Confusion matrix between True vs. Predicted speakers from the best BERT results.

5 Conclusion

This project presents an innovative approach to speaker identification in multiparty dialogues by integrating a fine-tuned BERT model with contextual utterance windows and syntactic features derived from POS tagging. Our method achieved a notable improvement over baseline models, attaining a 39% F1-score and 40% accuracy, which surpasses the existing CNN-based model by 9 percentage points in each metric. This enhancement underscores the effectiveness of combining contextual information and grammatical structures with advanced transformer-based embeddings for more accurate speaker classification.

Future work can explore the adoption of more advanced transformer architectures such as RoBERTa to potentially achieve higher contextual understanding and efficiency. Furthermore, incorporating named entity linking (Chen and Choi, 2016) could add deeper contextual insights, aiding in distinguishing speakers with similar linguistic styles. Addressing class imbalance through advanced techniques and augmenting syntactic feature representations through dependency or constituency parsing are also promising avenues for enhancing model performance. In conclusion, this project advances the field of speaker identification by effectively merging contextual and syntactic information with transformer-based models. The insights gained pave the way for developing more refined and accurate models capable of handling complex, multiparty conversational settings.

References

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Changmao Li and Jinho D. Choi. 2020. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5709–5714, Online. Association for Computational Linguistics.

Kaixin Ma, Catherine Xiao, and Jinho D. Choi. 2017. Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks. In *Proceedings of ACL 2017, Student Research Workshop*, pages 49–55, Vancouver, Canada. Association for Computational Linguistics.