Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks

Kaixin Ma
Math & Computer Science
Emory University
Atlanta, GA 30322, USA

Catherine Xiao
Math & Computer Science
Emory University
Atlanta, GA 30322, USA

Jinho D. Choi
Math & Computer Science
Emory University
Atlanta, GA 30322, USA

{kaixin.ma, catherine.xiao, jinho.choi}@emory.edu

Abstract

We propose a convolutional neural network model for text-based speaker identification on multiparty dialogues extracted from the TV show, *Friends*. While most previous works on this task rely heavily on acoustic features, our approach attempts to identify speakers in dialogues using their speech patterns as captured by transcriptions to the TV show. It has been shown that different individual speakers exhibit distinct idiolectal styles. Several convolutional neural network models are developed to discriminate between differing speech patterns. Our results confirm the promise of text-based approaches, with the best performing model showing an accuracy improvement of over 6% upon the baseline CNN model.

1 Introduction

Speakers verbalize their thoughts in different ways through dialogues. The differences in their expressions, be they striking or subtle, can serve as clues to the speakers' identities when they are withheld. This paper investigates the possibility of identifying speakers in anonymous multiparty dialogues.

Impressive advancements have been achieved in the field of speech recognition prior to this paper (Sadjadi et al., 2016; Fine et al., 2001; Campbell et al., 2006). Research on dialogue systems has also involved considerable efforts on speaker identification, as it constitutes an important step in building a more natural and human-like system (Raux et al., 2006; Hazen et al., 2003). Research in this area, however, has mostly been focused on acoustic features, which are absent in many situations (e.g., online chats, discussion forums, text messages). In addition, it is commonly acknowledged that natural language texts themselves reflect the personalities of speakers, in addition to their semantic content (Mairesse et al., 2007). Various experiments have demonstrated significant differences in the linguistic patterns generated by different participants, suggesting the possibility to perform speaker identification with text-based data. An increasing number of large unstructured dialogue datasets are becoming available, although they comprise only the dialogue transcripts without speaker labels (Tiedemann, 2012; Lowe et al., 2015). This paper attempts to identify the six main characters in the dialogues occurring in the first 8 seasons of the TV show, *Friends*. The minor characters in the show are to be identified collectively as *Other*.

For each episode, we first withhold the identity of the speaker to each utterance in its transcript, and have prediction models label the speakers. The accuracy and the F1 score of the labeling against the gold labels are used to measure the model performance. Our best model using multi-document convolutional neural network shows an accuracy of 31.06% and a macro average F1 score of 29.72, exhibiting promising performance on the text-based speaker identification task. We believe that the application of text-based speaker identification is extensive since data collected from online chatting and social media contains no acoustic information. Building accurate speaker identification models will enable the prediction of speaker labels in such datasets.

2 Related Work

Reynolds and Rose (1994) introduced the Gaussian Mixture Models (GMM) for robust text independent speaker identification. Since then, GMM has been applied to a number of datasets and achieved great results (Fine et al., 2001; Campbell et al., 2006). Knyazeva et al. (2015) proposed to perform sequence labeling and structured prediction in TV show speaker identification, and achieved better performance on sequential data. Despite the potential of text-based speaker identification in targeted

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 49–55
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3009

Speaker	Utterance
Monica	No . Not after what happened with Steve .
Chandler	What are you talking about ? We love **Schhteve** ! **Schhteve** was **schhexy** !.. Sorry .
Monica	Look , I do n't even know how I feel about him yet . Just give me a chance to figure that out .
Rachel	Well , then can we meet him ?
Monica	Nope . **Schhorry** .

Table 1: An excerpt from the transcripts to the TV show *Friends*.

Internet surveillance, research into this area has been scant. So far, there have been only a handful of attempts at text-based speaker identification. Kundu et al. (2012) proposed to use the K Nearest Neighbor Algorithm, Naive Bayes Classifier and Conditional Random Field to classify speakers in the film dialogues based on discrete stylistic features. Although their classification accuracies increase significantly from the random assignment baseline, there remains significant room for improvement. Serban and Pineau (2015) proposed their text-based speaker identification approach using Logistic Regression and Recurrent Neural Network (RNN) to learn the turn changes in movie dialogues. Their task is fundamentally different from the task of this paper, as their main focus is on the turn changes of dialogues instead of the identities of speakers. To the best of our knowledge, it is the first time the multi-document CNN has been applied to the speaker identification task.

3 Corpus

The Character Mining project provides transcripts to the TV show *Friends*; transcripts to the first 8 seasons of the show are publicly available in JSON format. Moreover, the first 2 seasons are annotated for the character identification task (Chen and Choi, 2016). Each season contains a number of episodes, and each episode is comprised of separate scenes.[1] The scenes in an episode, in turn, are divided at the utterance level. An excerpt from the data is shown in Table 1. In total, this corpus consists of 194 episodes, 2,579 scenes and 49,755 utterances. The utterance distribution by speaker is shown in Figure 1. The percentages for major speakers are fairly consistent. However, the *Other* speaker has a larger percentage in the dataset than any of the major speakers. The frequencies of interactions between pairs of speakers exhibit significant variance. For instance, *Monica* talks with *Chandler* more often than any other speaker, whereas *Phoebe* does not talk with *Rachel* and *Joey* very frequently. It will be of interest to note whether the variance of interaction rates can affect the performance of our identification model.

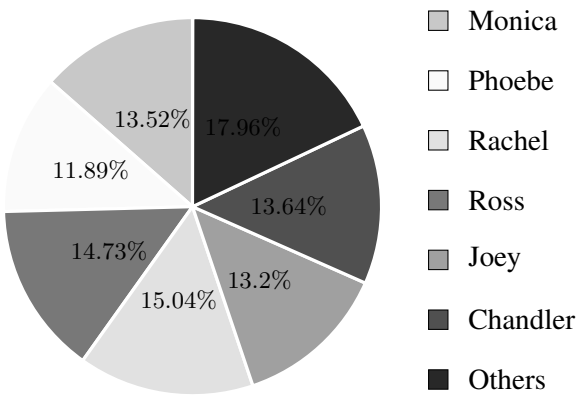

Figure 1: Data distribution

The first dataset is structured such that each utterance is considered as one discrete sample. To test the prediction performance for samples of greater lengths, all utterances of the same speaker in a scene are concatenated together as one single sample in the second dataset. Additional summary of the dataset is presented in Table 2.

4 Approaches

4.1 K Nearest Neighbor

In Kundu et al. (2012), the best result is reported using the K Nearest Neighbor algorithm (KNN), which is selected as the baseline approach for this paper, and implemented according to the original authors' specifications. Each utterance is treated as one sample, and 8 discrete stylistic features defined in the original feature template are extracted from each sample. Cosine similarity is used to locate the 15 nearest neighbors to each utterance. Majority voting of the neighbors, weighted by cosine similarity, is used to make predictions.

[1] http://nlp.mathcs.emory.edu/character-mining

4.2 Recurrent Neural Network

A recurrent neural network (RNN) model is also considered in the course of the experiments, where each utterance in the transcripts is handled separately. The RNN model treats the speaker identification task as a variant of sequence classification. For each instance, the concatenation of word embedding vectors is fed into the model, with a dense layer and softmax activation to model the probability for each speaker. The model is unable to demonstrate significantly above random accuracy on labeling, achieving a maximal accuracy of 16.05% after training. We conclude that a simple RNN model is unable to perform speaker identification based on textual data. Variations on the hyperparameters, including the dimension of the RNN, the dimension of word embeddings, and dropout rate, produced no appreciable improvements.

4.3 Convolutional Neural Network

Widely utilized for computer vision, Convolutional Neural Network (CNN) models have recently been applied to natural language processing and showed great results for many tasks (Yih et al., 2014; Kim, 2014; Shen et al., 2014). Speaker identification can be conceptualized as a variant of document classification. Therefore, we elected to use the traditional CNN for our task. The model is a minor modification to the proposal of Kim (2014), which consists of a 1-dimensional convolution layer with different filter sizes, a global max pooling layer, and a fully connected layer. Each utterance is treated as one sample and classified independently.

One of the challenges is the large number of misspellings and colloquialisms in the dataset as a result of the mistakes in the human transcription process and the nature of human dialogues. It is unlikely for these forms to appear in pre-trained word embeddings. The bold instances in Table 1 provide a glimpse into these challenges. It should also be noted that these irregularities oftentimes only deviate slightly from the standard spellings. A character-aware word embedding model is expected to produce similar vectors for the irregular forms and the standard spellings. Most of the colloquialisms appear frequently in the dataset, and the challenge they pose can be resolved by a pre-trained character-aware word embedding model, such as `fastText` (Bojanowski et al., 2016). The word embeddings used in this paper are trained on a dataset consisting of the *New York Times* corpus,

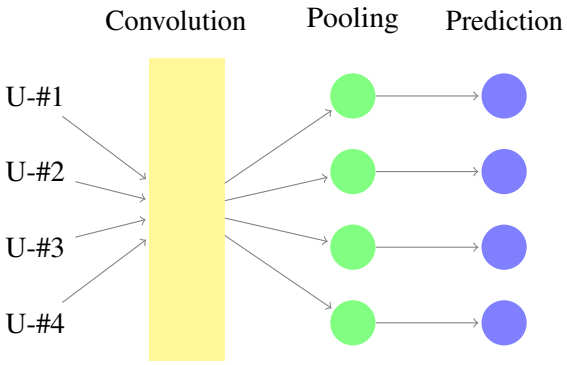

Figure 2: The baseline CNN model.

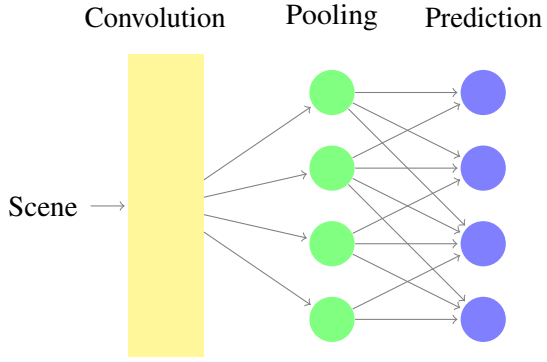

Figure 3: The multi-document CNN model.

the Wikipedia text dump, the Amazon Book Reviews,[2] and the transcripts from several TV shows.

4.4 CNN with Surrounding Utterance

Unlike other types of short documents such as movie reviews, where each sample is independent from the others, dialogues within a typical TV show are highly structured (Knyazeva et al., 2015). Every utterance is highly related to its prior and subsequent utterances, and it is important to take sequential information into account in predicting the speakers. However, contextual information is completely ignored by the basic CNN model. Each batch of input to the model consists of discrete utterances from different episodes and seasons, as shown in Figure 2.

To remedy the loss of contextual information, the CNN model is modified in a manner similar to the one proposed by Lee and Dernoncourt (2016). After the global max pooling layer, each utterance vector is concatenated with both the previous two utterances and the subsequent utterance in the same scene. Then, the vector is fed into the fully con-

[2] `snap.stanford.edu/data/web-Amazon.html`

Dataset	M	P	R_1	R_2	J	C	O	Total
Training	5,017	4,349	5,308	5,527	4,738	5,268	7,006	37,213
Development	898	799	1,092	821	930	846	931	6,317
Evaluation	810	769	1,082	983	909	673	999	6,225
Training	919	840	934	931	943	994	1,199	6,760
Development	151	148	151	131	148	139	159	1,027
Evaluation	116	116	137	132	129	119	156	905

Table 2: Dataset distribution by speakers. **M**: Monica, **P**: Phoebe, $\mathbf{R}_1$: Rachel, $\mathbf{R}_2$: Ross, **J**: Joey, **C**: Chandler, **O**: Other. Non-main speakers (all the others), are collectively grouped as the *Other* speaker.

nected layer. In this model, some information on the original dialogue sequence is preserved. Each scene is padded to the maximal sequence length, and fed into the model as one batch for both training and decoding. Figure 3 illustrates the structure of the model. Although topics within the scene are closely related, any single utterance is usually only relevant to its surrounding utterances. Based on this observation, including additional utterances in the prediction process can result in noisy input to the prediction model.

In a typical TV show, only a subset of characters are present in any particular scene. To further boost our model's ability to distinguish between speakers, the model optionally considers the set of speakers appearing in the scene. At the decoding stage, the Softmax probabilities for speakers absent from the scene are set to 0. The model benefits from the restrictions on its prediction search space. Such restrictions are applicable in the domain of targeted surveillance, where a vast number of speakers can be precluded from consideration during the identification process. For instance, speaker identification on a surveilled dialogue inside a criminal syndicate need only consider the members of the organization. In the majority of cases, however, the set of possible speakers may be difficult to ascertain. Therefore, we exclude this information in the determination of the best performing model.

4.5 CNN with Utterance Concatenation

Many individual utterances appearing in the dataset are fairly laconic and generic, as exemplified by the last utterance shown in Table 1, rendering them challenging to classify even with the help of contextual information. The proposed solution is to group multiple utterances together as one sample. Specifically, all of the utterances for each speaker in one scene are concatenated in the original dialogue order. We assign consistent unknown labels to all speakers in this dataset so that all the utterances in a single scene maintain their trackable provenances from the same speakers. The concatenated individual utterances can be fairly reasonable and consistent speech. As documents increase in length, it becomes easier for the CNN model to capture the speech pattern of each speaker. Once again, this model also optionally restricts its prediction search space to the set of speakers appearing in the scene for each batch of input.

5 Experiments

In the KNN experiment, the transcript to season 8 of *Friends* is used as evaluation data, and the first 7 seasons as training data. In the rest of the experiments, season 8 is used as evaluation data, and season 7 is used as the development set. The first 6 seasons are used as the training dataset. In each experiment, the F1 scores for the speakers, the average F1 score for major speakers, the average F1 score for all speakers, and the accuracy are reported in Tables 3 and 4.

In Kundu et al. (2012), the highest accuracy achieved by the KNN approach on the paper's film dialogue dataset was 30.39% , which is comparable to the best result of this paper. In contrast, the KNN approach did not perform well on the *Friends* dataset. Upon further examination of the KNN model's prediction process, we observe that the cosine similarities between any given utterance and its 15 nearest neighbors are consistently above 98%. The speaker labels are not linearly separable due to the low dimensionality of the feature space. The basic CNN model is able to outperform the baseline by almost 9% because the highly differing n-grams frequencies in the dataset enabled the model to distinguish between speakers. It is also worth noting that when the surrounding utterances

Model	Individual F1 Score							MF1	F1	ACC
	M	P	R_1	R_2	J	C	O			
KNN	13.30	12.13	17.34	19.23	14.68	14.61	19.23	15.22	15.80	16.18
RNN	17.87	15.22	14.98	17.51	17.42	13.48	12.02	15.39	15.50	16.05
CNN	20.55	17.52	24.20	24.70	28.15	14.05	31.81	21.36	22.86	25.01
Multi-Document-CNN	20.65	25.20	29.67	35.76	37.29	23.93	35.55	28.75	29.72	**31.06**
CNN-Concatenation	29.35	28.49	33.11	30.05	44.18	26.20	39.42	31.90	32.97	34.19

Table 3: Model performance. **MF1**: Average of F1 scores for major speakers, **F1**: Average of F1 scores for all speakers, **ACC**: Accuracy

Model	Individual F1 Score							MF1	F1	ACC
	M	P	R_1	R_2	J	C	O			
Multi-Document-CNN-2	28.13	29.59	41.49	48.15	45.72	36.06	46.98	38.19	39.45	41.36
CNN-Concatenation-2	36.43	33.16	50.09	45.03	53.67	39.90	51.02	43.05	44.19	46.48

Table 4: Model performance where the prediction labels are restricted to speakers present in each scene.

are taken into account, identification accuracy increases significantly from that achieved by the simple CNN. With more contextual information, the model is able to identify speakers with higher accuracy, as individual speakers react differently in comparable situations.

The experiment on the utterance concatenation dataset yields a relatively high identification accuracy, corroborating our theory that the prediction model can better capture different speech patterns on longer documents. When prediction labels are restricted to the speakers present in a scene, accuracy boosts of 10% and 12% are achieved on the two datasets, respectively.

Table 5 shows the confusion matrix produced by the multi-document CNN, i.e., the best performing model. The speakers for whom the model produces higher accuracies (*Ross* and *Other*) are also confused by the model more often than other speakers. The cause can be accounted for by the model's overzealousness in assigning these two labels to utterances due to their relatively large percentages in the training data. In addition, *Monica* and *Chandler* are often confused with each other. Due to their romantic relationship, it is possible that there is a convergence between their idiolectal styles. On the other hand, the confusion rates between *Phoebe* and *Rachel*, and between *Phoebe* and *Joey* are both fairly low. Such results confirm the observation that the frequency of interactions between speaker pairs correlates with the rate of confusion.

6 Conclusion

This paper presents a neural network-based approach to speaker identification in multiparty dialogues relying only on textual transcription data. The promising experimental results confirm the value of textual features in speaker identification on multiparty dialogues. The improvements produced by the consideration of neighboring utterances in the CNN's prediction process indicate that contextual information is essential to the performance of text-based speaker identification. Prior to this paper, Serban and Pineau (2015) used scripted dialogues to identify turn-taking and differences in speakers, where the actual identities of the speakers are irrelevant. However, this paper enables an identification where the names of the speakers are associated with their own utterances, a novel attempt in text-based speaker identification. Because of the ability of the model to uncover speaker identities in the absence of audio data, applications and interests in the intelligence and surveillance community are expected.

Although speaker verification based on acoustic signals is a helpful tool, it can conceivably be defeated by voice modulating algorithms. Whereas text-based speaker identification can discern the involuntary and unconscious cues of speakers. It is of interest to incorporate text-based features in a larger system of speaker identification to enhance its security. Several dialogue emotion recognition systems have incorporated both acoustic and

S \ P	M	P	R_1	R_2	J	C	O
M	**22.10**	8.40	6.17	17.90	8.27	20.37	16.79
P	12.09	**20.16**	6.63	18.21	4.81	17.95	20.16
R_1	13.22	6.38	**18.39**	19.50	4.25	15.90	22.37
R_2	8.75	4.17	4.68	**45.17**	5.70	14.95	16.58
J	9.13	4.51	6.38	19.03	**29.59**	16.61	14.74
C	18.28	6.39	3.42	10.85	6.84	**34.92**	19.32
O	13.21	3.80	5.01	12.71	4.80	15.02	**45.45**

Table 5: Confusion Matrix between speakers. **S**: true speaker, **P**: predicted speaker.

textual features, and resultant performances show improvements upon previous systems which rely only on one kind of features(Chuang* and Wu, 2004; Polzehl et al., 2009). Similarly, integration of acoustic and textual information in the speaker identification task can result in improved performance in future works.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquilloe. 2006. Support vector machines for speaker and language recognition. *Odyssey 2004: The speaker and Language Recognition Workshop* page 210–229.

Henry 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*. Los Angeles, CA, SIGDIAL'16, pages 90–100. http://www.sigdial.org/workshops/conference17/.

Ze-Jing Chuang* and Chung-Hsien Wu. 2004. Multi-Modal Emotion Recognition from Speech and Text . *The Association for Computational Linguistics and Chinese Language Processing* .

Shai Fine, Jirí Navrátil, and Ramesh A. Gopinath. 2001. A hybrid gmm/svm approach to speaker identification. In *ICASSP*.

Timothy J. Hazen, Douglas A. Jones, Alex Park, Linda C. Kukolich, and Douglas A. Reynolds. 2003. Integration of speaker recognition into conversational spoken dialogue systems]. *EUROSPEECH* .

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *In Proceedings of EMNLP* .

Elena Knyazeva, Guillaume Wisniewski, Herve Bredin, and Francois Yvon. 2015. Structured Prediction for Speaker Identification in TV Series. *LIMSI – CNRS – Rue John Von Neumann, Orsay, France. Université Paris Sud* .

Amitava Kundu, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Speaker identification from film dialogues. *IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction* .

Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. *In Proceedings of the NAACL Conference* .

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *In Proceedings of the SIGDIAL Conference* .

Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* pages 475–500.

Tim Polzehl, Shiva Sundaram, Hamed Ketabdar, Michael Wagner, and Florian Metze. 2009. Emotion classification in children's speech using fusion of acoustic and linguistic features. In *in Proc. InterSpeech*.

Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: one year of let's go! experience. *In INTERSPEECH* .

D.A. Reynolds and R.C. Rose. 1994. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing* pages 73–82.

Seyed Omid Sadjadi, Jason W. Pelecanos, and Sriram Ganapathy. 2016. The ibm speaker recognition system: Recent advances and error analysis. *arXiv preprint arXiv:1605.01635* .

Iulian V. Serban and Joelle Pineau. 2015. Text-Based Speaker Identification For Multi-Participant Open-Domain Dialogue System. *Department of Computer Science and Operations Research, Université de Montréal* .

Y. Shen, X. He, L. Deng J. Gao, and G. Mesnil. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. *In Proceedings of WWW* .

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. *In Proceedings of LREC* pages 2214–2218.

W. Yih, X. He, and C. Meek. 2014. Semantic Parsing for Single-Relation Question Answering. *In Proceedings of ACL 2014.* .