The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards Neural Speaker Modeling in Multi-Party Conversation: The Task, Dataset, and Models

Zhao Meng,[1,2] **Lili Mou,**[1,3] **Zhi Jin**[1]
[1]Key Laboratory of High Confidence Software Technologies, MoE; Software Institute, Peking University
[2]Department of Computer Science, ETH Zurich
[3]David R. Cheriton School of Computer Science, University of Waterloo
zhmeng@student.ethz.ch, doublepower.mou@gmail.com, zhijin@sei.pku.edu.cn

Abstract

In this paper, we address the problem of *speaker classification* in multi-party conversation, and collect massive data to facilitate research in this direction. We further investigate temporal-based and content-based models of speakers, and propose several hybrids of them. Experiments show that speaker classification is feasible, and that hybrid models outperform each single component.[1]

Introduction

Speaker modeling is important to dialog systems, and has been studied in traditional dialog research. However, existing methods are usually based on hand-crafted statistics and *ad hoc* to a certain application (Lin and Walker 2011). Recently, neural networks have become a prevailing technique in both task-oriented and open-domain dialog systems. After single-turn and multi-turn dialog studies, a few researchers have realized the role of speakers in neural conversational models. Li et al. (2016) show that, with speaker identity information, a sequence-to-sequence neural dialog system tends to generate more coherent replies. In their approach, a speaker is modeled by a learned vector (also known as an *embedding*). Such method, unfortunately, requires massive conversational data for a particular speaker to train his/her embedding, and thus does not generalize to rare or unseen speakers.

In this paper, we propose a *speaker classification* task that predicts the speaker of an utterance. It serves as a surrogate task for general speaker modeling, similar to *next utterance classification* (Lowe et al. 2015, NUC) being a surrogate task for dialog generation. The proposed task could also be useful in applications like *speech diarization*, which aims at answering "who spoke when" (Anguera et al. 2012).

We further propose a neural model that combines temporal and content information with interpolating or gating mechanisms. We investigate different strategies for combination, ranging from linear interpolation to complicated gating mechanisms.

Task Formulation and Data Collection

Assume that we have segmented a multi-party conversation into several parts by speakers; each segment comprises one or a few consecutive sentences $u_1, u_2, \cdots, u_N$, uttered by a particular speaker. A candidate set of speakers $\mathcal{S} = \{s_1, s_2, \cdots, s_k\}$ is also given. In our experiments, we assume $u_1, u_2, \cdots, u_N$'s speaker s_i is in $\mathcal{S}$. The task of speaker classification is to identify the speaker s_i of $u_1, \cdots, u_N$.

We represent the current utterance(s) as a real-valued vector $\boldsymbol{u}$ with recurrent neural networks. Speakers are also represented as vectors $\boldsymbol{s}_i, \cdots, \boldsymbol{s}_k$. The speaker classification is accomplished by a softmax-like function

$$\widetilde{p}_i = \exp\left\{\boldsymbol{s}_i^\top \boldsymbol{u}\right\}, \quad p(s_i) = \frac{\widetilde{p}_i}{\sum_j \widetilde{p}_j} \tag{1}$$

Because the number of candidate speakers may vary, the "weights" of softmax are not a fixed-size matrix, but the distributed representations of candidate speakers, $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_k$.

To facilitate the speaker classification task, we crawled transcripts of more than 8,000 episodes of TV talk shows, comprimising more than 200,000 individual sentences. We assumed that the current speaker is within the nearest k speakers. ($k = 5$, but at the beginning, k may be less than 5.) Since too few utterances do not provide much information, we required each speaker having at least 3 previous utterances, but kept at most 5. Samples failing to meet the above requirements were filtered out during data preprocessing. We split train/val/test sets according to TV show episodes instead of sentences; therefore no utterance overlaps between training and testing.

Methodology

We use a hierarchical recurrent neural network to model the current utterances $u_1, \cdots, u_N$ (Figure 1a), and obtain the utterance embedding $\boldsymbol{u}$, which is used in Equation 1.

Prediction with Content Information. Figure 1b illustrates the content-based model: a hierarchical RNN yields a vector $\boldsymbol{s}_i$ for each speaker, based on his or her nearest several utterances. The speaker vector $\boldsymbol{s}_i$ is multiplied by current utterances' vector $\boldsymbol{u}$ for softmax-like prediction (Equation 1). During prediction, we pick the candidate speaker that has the highest probability.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] Full and future version(s) can be found at https://arxiv.org/pdf/1708.03152.pdf Code and data available at https://sites.google.com/site/neuralspeaker/

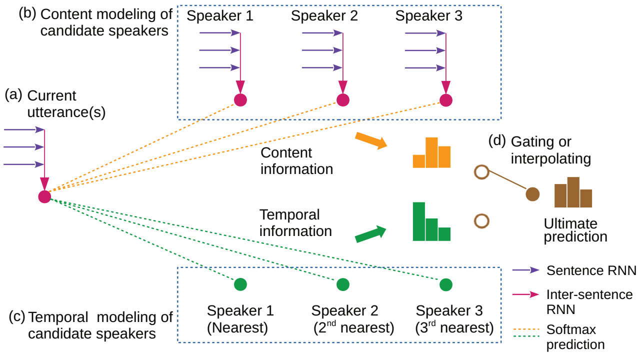

Figure 1: Hybrid content- and temporal-based speaker classification with a gating mechanism.

	Model	MacroF_1	WeightF_1	MicroF_1	Acc.	MRR.
	Random guess	19.93	34.19	27.53	27.53	N/A
	Majority guess	21.26	62.96	74.01	74.01	N/A
	Hybrid random/majority guess	25.26	61.99	69.29	69.29	N/A
	Temporal information	26.07	63.60	73.99	73.99	84.85
	Content information	42.61	65.04	61.82	58.58	74.86
	+ static att.	42.50	65.28	61.79	58.99	74.89
	+ sentence-by-sentence att.	42.56	65.96	62.86	59.81	75.58
Hybrid	Interpolating after training	**44.25**	**71.35**	**76.10**	**75.84**	**85.73**
	Interpolating while training	41.30	70.10	75.57	75.31	85.20
	Self-adaptive gating	39.45	69.55	74.11	74.09	84.85

Table 1: Model performance (in percentage). Validation was accomplished by each metric itself because different metrics emphasize different aspects of model performance.

Prediction with Temporal Information. We sort all speakers in a descending order according to the last time he or she speaks, and assign a vector (embedding) for each index in the list. Each speaker vector is randomly initialized and optimized as parameters during training. The predicted probability of a speaker is also computed by Equation 1.

Combining Content and Temporal Information. As both content and temporal provide important evidence for speaker classification, we propose to combine them by interpolating or gating mechanisms (illustrated in Figure 1d). In particular, we have

$$\boldsymbol{p}^{\text{(hybrid)}} = (1-g)\cdot \boldsymbol{p}^{\text{(temporal)}} + g \cdot \boldsymbol{p}^{\text{(content)}} \qquad (2)$$

Here, g is known as a *gate*, balancing these two aspects. We investigate three strategies to compute the gate. (a) Interpolating after training. We train two predictors separately, and interpolate after training by validating the hyperparameter g. (b) Interpolating while training. We could also train the hybrid model as a whole with cross-entropy loss directly applied to Equation 2. (c) Self-adaptive gating. Inspired by hybrid content- and location-based addressing in Differentiable Neural Computers (Graves and others 2016, DNCs), we design a learnable gate in hopes of dynamically balancing temporal and content information. Formally

$$g = \operatorname{sigmoid}\left(w \cdot \operatorname{std}[\, p^{\text{(content)}}\,] + b\right) \qquad (3)$$

where we compute the standard deviation (std) of p. w and b are parameters to scale $\operatorname{std}[\, p^{\text{(content)}}\,]$ to a sensitive region of the sigmoid function.

Experimental Results

Neural layers were set to 100d. The batch size was 10. Dropout rate and early stop were also applied by validation. Table 1 compares the performance of different models. Majority-class guess results in high accuracy, showing that the dataset is screwed. Hence, we choose macro F_1 as the main metric, which addresses minority classes more than other metrics. Other metrics including accuracy, mean reciprocal ranking (MRR), and micro/weighted F_1 are presented as additional evidence.

As shown, the content-based model achieves higher performance in macro F_1 than majority guess, showing the effectiveness of content information. We also tried attention mechanisms, following Rocktäschel et al. (2016). However, they bring little improvement (if any).

All hybrid models achieve higher performance compared with either content- or temporal-based prediction in terms of most measures, which implies content and temporal information sources capture different aspects of speakers.

Among different strategies of hybrid models, the simple approach "interpolating after training" surprisingly outperforms the other two. A plausible explanation is that training a hybrid model as a whole leads to optimization difficulty in our scenario; that simply interpolating well-trained models is efficient yet effective. Interested readers are referred to Sha et al. (2018) for detailed discussion. In our experiments, the hyperparameter g is sensitive and only yields high performance for certain values of g. Thus, the learnable gating mechanism could also be useful in some scenarios, as it is self-adaptive.

Acknowledgments

This research is supported by the National Basic Research Program of China (the 973 Program) under Grant No. 2015CB352201, and the National Natural Science Foundation of China under Grant Nos. 61232015 and 61620106007.

References

Anguera, X.; Bozonnet, S.; Evans, N.and Fredouille, C.; Friedland, G.; and Vinyals, O. 2012. Speaker diarization: A review of recent research. *ASLP* 20(2):356–370.

Graves, A., et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538(7626):471–476.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. In *ACL*, 994–1003.

Lin, G. I., and Walker, M. A. 2011. All the world's a stage: Learning character models from film. In *AIIDE*, 46–52.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 285–294.

Rocktäschel, T.; Grefenstette, E.; Hermann, K. M.; Kočiskỳ, T.; and Blunsom, P. 2016. Reasoning about entailment with neural attention. In *ICLR*.

Sha, L.; Mou, L.; Liu, T.; Poupart, P.; Li, S.; Chang, B.; and Sui, Z. 2018. Order-planning neural text generation from structured data. In *AAAI*.