The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

Analysis of Utterance Combinations for Emotion Recognition in Conversation

Nang Su Lin Nwe *1 Yanan Wang *2 Jianming Wu *2 Gen Hattori *2 Aye Thida *1

*1 University of Computer Studies, Mandalay, Myanmar *2KDDI Research Inc.,

Emotion recognition in conversation is an important step for developing empathetic systems in diverse areas such as healthcare, education, and business. Recent work demonstrates that utterance-level conversational context modeling leads to high-performance emotion recognition. In this paper, we combine the utterances of conversations where the same person speaks continuously with the same emotions, and train with DialogueRNN to further improve emotion recognition performance. According to comparison of DialogueRNN models that trained by combining different numbers of utterances, the combination of three utterances achieves the highest F1-score of 63.90%, and improved the F1-score of 1.63% compared to baseline.

1. Introduction

Emotion recognition in conversation (ERC) has been gaining attention from the research community due to the proliferation of conversational data opened on social media platforms such as Facebook, Twitter. ERC is also a key step in developing empathy systems such as health-care chatbots that generate empathetic responses based on user emotions.

Some studies have focused on modeling the conversational context to achieve high-performance emotion recognition (Hazarika et al., 2018, Poria et al., 2018). As shown in Figure 1, person A's emotions are influenced not only by herself but also by person B in the conversation. It makes sense that analyzing conversational context from two aspects of self and inter-speaker influences can accelerate emotion understanding (Liu, 2014)(Hazarika et al., 2018, Poria et al.,2018). DialogueRNN is an architecture that applies global, emotion and party Gated Recurrent Units (GRUs)(chung, 2014) to model self and inter-speaker influences(Majumder et al., 2019). Although DialogueRNN can be adapted to model the conversational context, short utterances like "Yeah", "Okay" or "Oh" do not contain any emotional representations and are difficult to be trained corresponding different emotion labels.

In this paper, in order to improve the F1-score of DialogueRNN, we combine the utterances of conversations where the same person speaks continuously with the same emotions, and train it on the DialogueRNN. According to comparison of DialogueRNN models that trained by combining different numbers of utterances, the combination of three utterances achieves the highest F1-score of 63.90%, which can improve the F1-score of 1.63% compared to baseline.

2. Utterance combination strategy

We propose the utterance combination strategy to modify IEMOCAP dataset (Busso et al. 2008) by combining utterances. The IEMOCAP dataset consists of two-way conversation videos of 10 unique speakers. These videos are segmented into utterances and these utterances are annotated with one of six emotion labels of happy, sad, neutral, angry, excited, and frustrated. We define the utterance combination strategy showed as Algorithm 1. Given that the number of utterances to be combined has been determined, the utterances have the same emotion label and are combined when the same person speaks. In contrast, the utterances that do not have the same label, or that are not spoken by the same person, use the original utterances. We combine the utterances from all videos in the IEMOCAP dataset, and train the DialogueRNN model on the modified IEMOCAP dataset. An example of comparison between the before and after adaptation of the algorithm 1 is shown at Table 1 and 2.

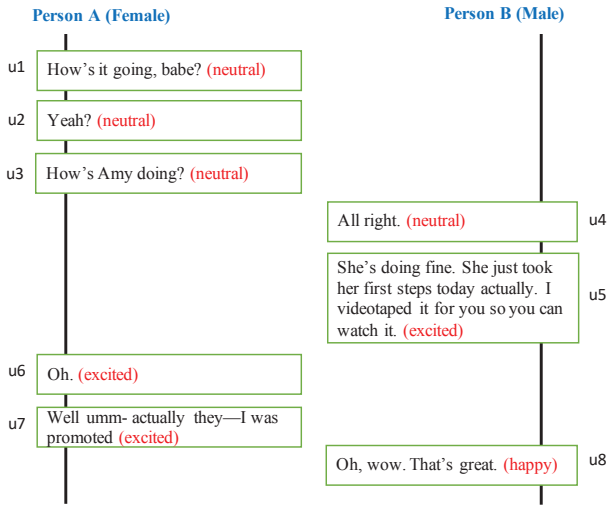

Figure 1: A abridged dialogue from IEMOCAP dataset. Person A (wife) is working away from person B (husband). At first, A and B emotions are driven by their own emotional inertia (u1, u2, u3, u4). But finally, both person's emotions were influenced by their counterpart (u6, u7, u8).

Contact: Yanan Wang, KDDI Research Inc., 2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502 JAPAN,+81-70-4560-2745, wa-yanan@kddi-research.jp

Id	Person	Utterance	Emotion
u1	Male	Hey sweetie.	sad
u2	Male	Hey.	sad
u3	Female	Um.	happy
u4	Female	How's it going?	happy
u5	Male	Uh, -um.	sad
u6	Female	Why the sad face?	happy
u7	Male	They need me out there sooner than I though.	sad
u8	Male	They want to deploy me next week.	sad
u9	Male	For–	sad
u10	Male	They said anywhere from two to three years.	sad
u11	Female	Honey–	frustrate

Table 1: Original utterances in IEMOCAP dataset before the algorithm is adapted.

Id	Person	Utterance	Emotion
u1 u2	Male	Hey sweetie. Hey	sad
u3 u4	Female	Um. How's it going?	happy
u5	Male	Uh, -um	sad
u6	Female	Why the sad face?	happy
u7 u8	Male	They need me out there sooner than I though. They want to deploy me next week	sad
u9 u10	Male	For– They said anywhere from two to three years.	sad
u11	Female	Honey–	frustrate

Table 2: Utterances obtained after the adaptation of algorithm (combine two utterances).

3. Experiment

In this section, we train dalogueRNN on the different modified IEMOCAP dataset showed at Table 4, and demonstrate the best utterance combinations based on the F1-score of emotion recognition.

3.1 Feature extraction

We follow (Majumder et al, 2019) to extract textual features from utterance. Each token were initialize with pre-trained 300-dimensional GloVe word vectors (Pennington, 2014) and feed them to 1D convolutional neural network (CNN). Outputs are then subjected to max-pooling which is followed by rectified linear unit (ReLU) activation. These activations are then passed through 100-dimensional dense layer, which is regarded as the textual representation of the utterance.

Algorithm 1 Utterance Combination Strategy

for $i = 1, ..., K$ **do** ▷ K videos
  $tempUtterance \leftarrow []$
  $FeatureToTrain \leftarrow []$
  for $j = 1, \ldots, M$ **do** ▷ M utterances in a video
    if currentSpeaker=nextSpeaker **and** currentLabel=nextLabel **then**
      $tempUtterance \Leftarrow textFeature[j]$
    else if currentSpeaker=previousSpeaker **and** currentLabel=previousLabel **then**
      $tempUtterance \Leftarrow textFeature[j]$
      for $k = 1, ..., N$ **do** ▷ N=utterances in tempUtterance
        if $lengthOftempUtterance > NumberOfUtteranceToCombine$ **then**
          $FeatureToTrain \leftarrow sum(tempUtterance[k : k + step])$
          $k \leftarrow k + step$
        else
          for All remaining utterances in tempUtterance **do**
            $FeatureToTrain \leftarrow$ All remaining utterances in tempUtterance
          end for
        end if
      end for
    else
      $FeatureToTrain \leftarrow textFeature[j]$
    end if
  end for
end for

3.2 Baseline model

DialogueRNN is an attentive RNN for emotion recognition in conversation, which models the two emotional influences by using three types of GRUs. The global and emotion GRUs are for inter-speaker influence and the party GRU is for self-speaker influence. The detail of DialogueRNN is shown as following:

- The utterances are fed to global and party GRU to update the context and speaker state, respectively.
- The speaker state of party GRU is updated based on 1) the utterance, 2) the speaker's previous state and 3) the conversational context obtained by attending the global GRU.
- The updated speaker state and the emotion label of the previous utterance are fed to the emotion GRU to decode the emotion representation of the utterance which is used for emotion classification.
- Attending over the emotion representation captures the context from the other utterances spoken by different speakers.

Combination	Emotions						
	Happy	Sadness	Neutral	Angry	Excited	Frustrate	Weighted average
Original (baseline)	33.19	77.83	58.72	**66.07**	67.86	**60.74**	62.27
Two	34.42	**79.32**	**59.93**	65.86	72.90	60.08	63.64
Three	36.28	79.29	58.81	64.48	75.04	60.58	**63.90**
Four	36	79.74	58.7	62.58	72.58	69.80	63.32
Five	33.96	74.45	58.15	64.83	**76.39**	59.70	62.89
Six	**43.60**	77.75	58.54	62.91	67.75	60.64	62.76
Seven	36.62	75.23	58.59	62.87	72.37	59.80	62.43

Table 3: Test-set weighted F1-score results of DialogueRNN for different combinations of utterances number.

Combination	Train	Validation	Test	Total
Original (baseline)	5283	527	1623	7433
Two	4376	401	1361	6138
Three	4776	438	1477	6691
Four	5028	458	1551	7037
Five	5103	487	1571	7161
Six	5176	479	1578	7233
Seven	5211	485	1593	7289

Table 4: Size of modified IEMOCAP datasets (contain original one). All datasets are partitioned into train, validation and test set with the ratio of 7:1:2. Due to the size of the dataset depends on the number of utterances we decide, the larger number of utterances we decide, the smaller size of data that fit the proposed utterance combination strategy. For instance, we can combine total of 1295 utterances with decided number of two. In contrast, only total of 144 utterances can combine with a decided number of seven.

3.3 Result

As shown in table 3, all the F1-scores of combined utterances are higher than the original method (baseline). Specially, the combination of three utterances get the highest F1-score of 63.9% with 1.63% surpasses the original F1-score. In addition, we also compare the score of individual labels. The result shows that the combined utterances can recognize emotion better in four out of six emotion classes. For angry and frustrated classes, the model lag behind DialogueRNN by 0.21% and 0.1% respectively. We consider that the improvement contributes to obtaining meaningful emotional information.

4. Conclusion

In this work, we proposed an utterance combination strategy applied to combine the utterances where the same person speaks continuously with the same emotions based on the number of combinations to improve emotion recognition accuracy of DialogueRNN. The results on the IEMOCAP dataset demonstrated that the combination of utterances indeed improve the F1-score of DialogueRNN. Considering some utterances contain meaningful emotional information enough that not need to combine, we plan to explore a different algorithm that can automatically decide whether to combine the utterances.

References

[Hazarika 2018] Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In Proceedings of the NAACL-HLT, Volume 1 (Long Papers), 2122–2132. 2018

[Hazarika 2018] Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R.; "Icon: Interactive conversational memory network for multimodal emotion detection," in Proceedings of the 2018 Conference on EMNLP, pp. 2594–2604. 2018

[Liu 2014] Liu, F.; and Maitlis, S. Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings. Journal of Management Studies, 51(2):202–234. 2014

[Majumder 2019] Majumder, N.; Poria, S.; Haz-arika, D.; Mihalcea, R.; Gelbukh, A.; and ErikCambria. DialogueRNN: An attentive RNNfor emotion detection in conversations.Thirty-Third AAAI Conference on Artificial Intelligence. 2019

[Chung 2014] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.Presented in NIPS 2014 Deep Learning and Representation Learning Workshop. 2014

[Pennington 2014] Pennington, J.; Socher, R.; and Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2014

[Busso 2014] Busso, C.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Mower,E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S.IEMOCAP: Interactive emotional dyadic motioncapture database. Language resources and evaluation42(4):335–359. 2008