Multilingual Emotion Recognition in Conversation

Junyin Chen Hanshu Ding Zoe Fang Yifan Jiang
{junyinc, hsding99, zoekfang, yfjiang}@uw.edu
Department of Linguistics
University of Washington

Abstract

TBD

1 Introduction

TBD

2 Task Description

2.1 Primary Task

Our primary task is emotion recognition in conversation (ERC) task on the text modality of the Multimodal EmotionLines Dataset (MELD). The dataset is in English and contains dialogues and utterances from TV series scripts . We will predict the emotion for each utterance from dialogues involving multiple speakers.

2.1.1 Dataset

Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) [1] dataset is a multimodal emotional conversational dataset built on EmotionLines dataset (Hsu et al., 2018) with three modalities: audio, visual, and text. The dataset contains about 13,000 utterances from 1,433 dialogues, which are collected from the TV-series Friends. Each utterance is annotated with Ekman's basic emotions plus neutral and sentiment labels.

2.2 Adaptation Task

Our adaptation task is to adapt our model to Chinese dialogues in the Multi-party Dialogue Dataset (MPDD). Other dimensions for this task remain the same as the primary task.

2.2.1 Dataset

Multi-party Dialogue Dataset (MPDD) (Chen et al., 2020) [2] is a Chinese emotional conversational dataset. The dataset contains a total of 25, 548 utterances from 4, 142 dialogues, which are collected from five TV series scripts from www.juban108.com. Each uttrance is annmotated with three types of labels: emotion, relation, and target listener. In particular, the emotion labels are consistent with those in the Emotionlines dataset.

3 System Overview

3.1 Design

Instead of appending the correspondent speaker tag with the utterance for training, we integrate the ideology of having a multi-task deep neural network that shares the lower layer across multiple single-sentence and pairwise text classification tasks.

With the benefit of multi-task deep neural network, we enable our model to gain context awareness with multiple context related tags, such as speaker, past and future utterance.

3.2 Methodology

We want to determine the emotion of a specific utterance in a dialogue group M, which contains multiple speakers in speaker group S. The dialogue group M can be expressed as a list of vectors: $dialogue = [x_1, x_2, ..., x_{n-1}, x_n]$, where each utterance contains multiple words. The correspondent speaker list can be expressed as another list of vectors: $speaker = [s_1, s_1, ..., s_2, s_3]$, where s_n is in the speaker group S. Since this is a supervised setup, we will utilize the manually labeled emotion label e_n corresponding to each utterance x_n in the dialogue group M.

The simplest solution is to have a function f that takes each utterance x_t as input and output the correct label e_t. For our setup, we decide on a multi-tasking function f to output correspondent speaker s_t, as we want to take the speaker into account. Furthermore, we add past utterance $[x_1, x_2, ..., x_{t-1}]$ and even future utterance $[x_{t+1}, x_{t+2}, ..., x_n]$, as we anticipate adding more context around the utterance for analysis will further improve the result.

[1] https://affective-meld.github.io/
[2] http://nlg.csie.ntu.edu.tw/nlpresource/MPDD

3.3 Algorithm

For each task in the multi-task deep neural network, we use the pretrained "Roberta-base" model Liu et al. (2019). We choose BERT like algorithms as they both have simple structure, and support more than one segment for tokenization. We choose to use RoBERTa as the main algorithm for training, facilitating result comparison with results listed in Kim and Vossen (2021). Even though pre-trained BERT and RoBERTa models do not expect more than two segments as inputs, both Kim et al. and us show that having more than two segments improves evaluation results.

We will have either two or three segments, if we want to include both past and future utterances.

4 Approach

4.1 Problem Statement

MELD provides information on the speaker and turn (in the dialogue) of each utterance. We want to take these two factors into account and build a model that learns speaker-specific features and the context of an utterance. At the same time, we want to build a multi-task model to accomplish both tasks at the same time.

4.2 Model Architecture

We build a Multi-Task Deep Neural Networks (MP-DNN) Model Liu et al. (2019) to train both Emotionlines and MPDD data so that the model can learn shared knowledge between two datasets. We use a pre-trained RoBERTa base encoder transformer for the initial text embeddings. On top of the shared pre-trained transformer encoder, we build two different task-specific models for each task.
To manipulate the inputs, we either add some context or provide the speaker information. The caveat of training a MP-DNN on top of RoBERTa is that the latter is previosuly trained with inputs no longer than two sentences, where each two sentence has a BOS <*s*> token and an EOS </*s*> token. When we concatenate more than one past utterance to the to-predict utterance, we strip the EOS token of the past utterances and keep only one that exists between the context and the to-predict utterance. Algorithm 1 shows our preprocessing algorithm of adding context information.

4.3 Evaluation

Both tasks will be evaluated using weighted F1 to account for the imbalance of the dataset. We will also evaluate the error rate between predicted the true emotion labels that are labelled manually by the dataset curators.

Algorithm 1 Add All Past Utterance(s)

Given an $Uttr$ and its idx in dialogue
if $idx \neq 0$ **then**
    $i \leftarrow 1$
    while $idx - i \geq 0$ **do**
        $Uttr_{past} \leftarrow Data_{idx-i}$
        $Speaker_{past} \leftarrow Uttr_{Past_{speaker}}$
        $str \leftarrow Speaker_{past} + Uttr_{Past}$
        $Context \leftarrow Context + str$
        $i \leftarrow i + 1$
$Uttr_{curr} \leftarrow Context + EOS + Uttr$

5 Results

past utterance	future utterance	weighted F1
0	0	60.17
0	6	61.96
6	0	62.46
6	6	62.56

Table 1: Weighted F1 when number of past or future utterances are added as context.

Using our evaluation metric, the model currently yields a 62.56 percent weighted F1 score. This accuracy fluctuates within a small range when we change the inputs to the model. That is, accuracy on test data can vary by in 1.2 percent when the model receives inputs with(out) speaker information or past/future utterances. Table 1 shows the result under different model setups

emotion	true count	total count	accuracy
neutral	1026	1256	81.69
joy	247	402	61.44
anger	156	281	55.52
surprise	173	345	50.14
sadness	66	208	31.73
disgust	5	68	7.35
fear	1	50	2.00

Table 2: Weighted F1 when number of past or future utterances are added as context.

The prediction output and accuracy is shown as below in Table 2. It is obvious that the prediction for emotion 'neutral' is easier compared to other types. Also, it is interesting that the accuracy for

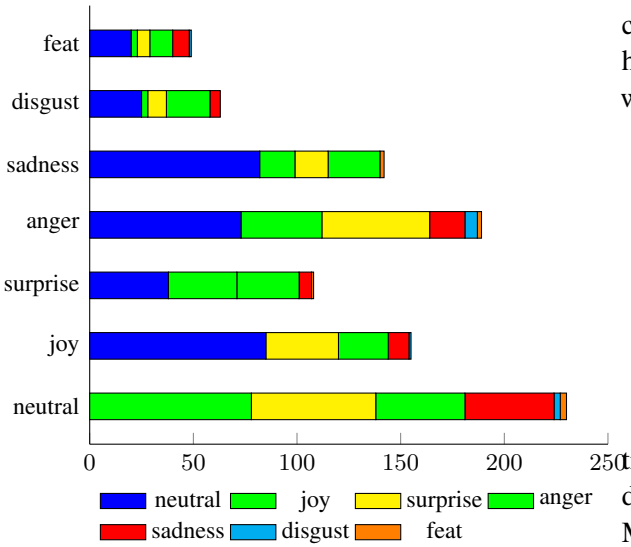

Figure 1: The distribution of mis-predicted emotions

the prediction of specific emotions decreases with the decreasing number of learning instances for the corresponding emotions.

Besides, we also noticed that there was a high tendency for non-neutral emotions to be predicted as 'neutral' emotions (323 out of 936 misprediction). Oppositely, for real neutral emotions, they are likely to be predicted as 'joy' or 'surprise'.

6 Discussion

6.1 Pre-trained model decision

Using RoBERTA-base pre-trained model, we achieve similar result illustrated in Kim et al. However, using RoBERTa-full pre-trained model used by Kim et al., weighted F1 score reduced by as much as 30 percent, regardless of number of past or future utterances included in the training dataset.

6.2 Number of past and future utterance.

Adding future utterance alone does not greatly improve the result. The result of adding future utterance alone is similar of having no future or past utterance. Adding both past and future utterance, preforms almost the same as adding only past utterance.

6.3 Qualitative Analysis

We did a qualitative analysis with on 10 correctly and 10 incorrectly classified random sample from the test split. For all the incorrectly classified random sample, the speaker tags are all incorrectly classified. For the 10 correctly classified random sample, 40 percent of the speaker tags are correctly classified. This shows that having speaker tag does help improve the result, but not as substantial as we hopped.

> <s>Monica:Is that too much to ask after six year?!Monica:I mean, all I'm asking for is just a little emotio! Chandler:And you're upset because you didn't make your best friend cry?</s></s>I mean what?</s>

An incorrectly classified example. The prediction is surprise while the truth is anger. The predicted speaker is Rachel while the true speaker is Monica.

> <s>Rachel:Oh, that sounds great.Others:How does that sound?Others:Well, I've got a project for you that's a lot more related to fashion.Others:Well, don't think I haven't noticed your potential.Rachel:Oh, you got me.Others:Eh.</s></s>Come on over here, sweetheart.</s>

A correctly classified example. Both the prediction and truth are neutral. Both the predicted speaker and the true speaker are Others.

We initially anticipate that the speaker tag will be utilized for sentiment classification. However, both the incorrectly and correctly classified example, the <s> token in the last layer mainly focus on punctuation marks. This phenomenon far deviates from Kim et al.'s result, that <s> token mainly focus on the speaker in correctly classified examples. This proves that separating speaker and utterance for speaker detection does not benefit the model to utilize speaker context for sentiment analysis.

Furthermore, the RoBERTa pre-trained tokenizer tokenized Mon from Monica and Ch from Chandler. We suspect incorrect tokenization of the Speaker may further deteriorate the result, as the model may not correctly utilize the speaker tags in past or future utterance for current utterance's sentiment analysis.

7 Conclusion

TBD

References

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv e-prints*, page arXiv:2108.12009.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.