EmotionX-JTML: Detecting emotions with Attention

Johnny Torres
ESPOL University / Guayaquil, Ecuador
jomatorr@espol.edu.ec

Abstract

This paper addresses the problem of automatic recognition of emotions in text-only conversational datasets for the EmotionX challenge. Emotion is a human characteristic expressed through several modalities (e.g., auditory, visual, tactile), therefore, trying to detect emotions only from the text becomes a difficult task even for humans. This paper evaluates several neural architectures based on Attention Models, which allow extracting relevant parts of the context within a conversation to identify the emotion associated with each utterance. Empirical results the effectiveness of the attention model for the *EmotionPush* dataset compared to the baseline models, and other cases show better results with simpler models.

1 Introduction

With technology increasingly present in people's lives, human-machine interaction needs to be as natural as possible, including the recognition of emotions. Emotions are an intrinsic characteristic of humans, often associated with mood, temperament, personality, disposition or motivation (Averill, 1980). Moreover, emotions are inherently multimodal, as such, we perceived them in great detail through vision or speech (Jain and Li, 2011).

Detecting emotions from text poses particular difficulties. For instance, an issue that arises from working with conversational text data is that the same utterance (message) can express different emotions depending on its context. The table 1 illustrate the issue with some utterances expressing different emotions with the same word from the challenge datasets (Chen et al., 2018).

Chandler	I guess it must've been some movie I saw. (Neutral)
Chandler	What do you say? (Neutral)
Monica	***Okay! (Joy)***
Chandler	Okay! Come on! Let's go! All right! (Joy)
Rachel	Oh okay, I'll fix that to. What's her e-mail address? (Neutral)
Ross	Rachel! (Anger)
Rachel	All right, I promise. I'll fix this. I swear. I'll-I'll- I'll-I'll talk to her. (Non-neutral)
Ross	***Okay! (Anger)***
Rachel	Okay. (Neutral)

Table 1: Two dialogs from Friends TV scripts. The word "Okay!" denote different emotions depending of the context.

Despite improvements with neural architectures, given an utterance in a conversation without any previous context, it is not always obvious even for human beings to identify the emotion associated. In many cases, the classification of utterances that are too short is hard. For instance, the utterance *'Okay'* can be either an *Agreement* or indicative of *Anger*, for such cases the context plays an essential role at disambiguation. Therefore, using context information from the previous utterances in a conversation flow is a crucial step for improving DA classification.

In this paper, we explore the use of AMs to learn the context representation, as a manner to differentiate the current utterance from its context as well as a mechanism to highlight the most relevant information while ignoring unnecessary parts for emotion classification. We propose and compare different neural-based methods for context representation learning by leveraging a recurrent neu-

Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media , pages 56–60,
Melbourne, Australia, July 20, 2018.

ral network architecture with LSTM (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRUs) (Chung et al., 2014) in combination with AMs.

2 Related Work

The identification of emotions is an essential task for understanding natural language and building conversational systems. Previous works on recognizing emotion in text documents consider three categories: keyword-based, learning-based, and hybrid recommendation approaches (Kao et al., 2009).

In recent years, learning methods based on neural architectures have achieved great success. Emotion recognition can be framed as a sentences classification task and has been addressed using various traditional statistical methods, such as Markov Models (HMM) (Stolcke et al., 2000), conditional random fields (CRF) (Zimmermann, 2009) and support vector machines (SVM) (Henderson et al., 2012). Recent work has shown advances in text classification using deep learning techniques, such as convolutional neural networks (CNN) (Kalchbrenner and Blunsom, 2013; Lee and Dernoncourt, 2016), recurrent neural networks (RNNs) (Lee and Dernoncourt, 2016; Ji et al., 2016) and short-term long memory models (LSTM) (Shen and Lee, 2016).

Recent previous works have suggested utilizing context as possible prior knowledge for utterance classification (Lee and Dernoncourt, 2016; Shen and Lee, 2016). Contextual information from preceding utterances has been found to improve the classification performance, but it depends on the specific aspect of the dataset Ortega and Vu (2017). These works highlight that such information should be differentiable from the current utterance information; otherwise, the contextual information could have a negative impact.

Attention mechanisms (AMs) introduced by Bahdanau et al. (2014) have contributed to significant improvements in many natural language processing tasks, for instance machine translation (Bahdanau et al., 2014), sentence classification (Shen and Lee, 2016) and summarization (Rush et al., 2015), uncertainty detection (Adel and Schütze, 2016), speech recognition (Chorowski et al., 2015), sentence pair modeling (Yin et al., 2015), question-answering (Golub and He, 2016), document classification (Yang et al., 2016) and entailment (Rocktäschel et al., 2015) . AMs let the model decide what parts of the input to pay attention to according to the relevance of the task.

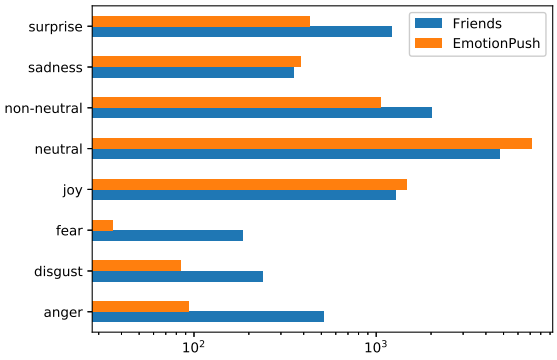

Figure 1: Label distribution of the datasets in the challenge.

3 Data

Conversational datasets with utterance information are accessible such as movies, television scripts or chat records. Although, despite the importance of emotion detection in conversational systems, most datasets do not have emotion tags, so it is not possible to use such data directly to train models to identify emotions.

The EmotionX challenge provides two annotated datasets with emotions tags. The first, denoted *Friends*, contains the scripts of seasons 1 to 9 of *Friends* TV shows[1]. The second, denoted *EmotionPush*, consist of private conversations between friends on Facebook Messenger collected by the app*EmotionPush* (2016).

Each utterance in the datasets has the same format: the user, the message, and the emotion label. The labels are one of six primary emotions anger, disgust, fear, happiness, sadness, surprise, and neutral defined in (1987). EmotionPush dataset has more skewed label distribution than Friends dataset as shown in Fig.1.

Both Friends and EmotionPush datasets contain 1,000 dialogues. The length distribution of utterances in EmotionPush dataset is much shorter than the length of those of TV show scripts (10.67 vs. 6.84). The EmotionPush dataset is anonymized to hide users' details such as names of real people, locations, organizations, and email addresses. Ad-

[1]Scripts of seasons 1-9 of "Friends": http://www.livesinabox.com/friends/scripts.shtml

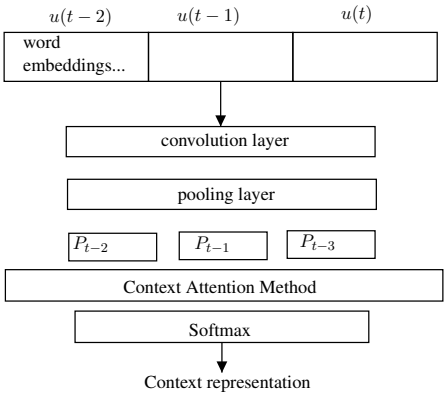

Figure 2: An overview of the architecture of the model based on Attention for classifying emotions in the conversation context.

ditional steps were applied to ensure the privacy of users as described in the dataset paper (Chen et al., 2018).

4 Model

The architecture of the model considers two main parts: the CNN-based utterance representation and the attention mechanism for context representation learning. The Figure 2 shows an overview of the model. The model feeds the context representation into a softmax layer which outputs the posterior of each context utterances given the current utterance.

4.1 Utterance Representation

The proposed architecture uses CNNs for the representation of each utterance. For the emotion classification task, the input matrix represents an utterance and its context (i.e.,n previous utterances). Each column of the matrix stores the embeddings of the corresponding word, resulting in d dimensional input matrix $M \in \mathbb{R}^{M \times d}$. The weights of the word embeddings use the 300-dimensional GloVe Embeddings pre-trained on Common Crawl data (Pennington et al., 2014).

The model performs a discrete 1D convolution on an input matrix with a set of different filters of width $|f|$ across all embedding dimensions d, as described by the following equation:

$$(w*f)(x,y) = \sum_{i=1}^{d} \sum_{j=-|f|/2}^{|f|/2} w(i,j) \cdot f(x-i, y-j) \quad (1)$$

After the convolution, the model applies a max pooling operation that stores only the highest activation of each filter. Additionally, the model applies filters with different window sizes 3-5 (multi-windows), which span a different number of input words. Then, the model concatenates all feature maps to one vector which represents the current utterance and its context.

4.2 Attention Layer

The model applies an attention layer to different sequences of input vectors, e.g., representations of consecutive utterances in a conversation. For each of the input vectors $u(t-i)$ at time step $t-i$ in a conversation, the model computes the attention weights for the current time step t as follows:

$$\alpha_i = \frac{exp(f(u(t-i)))}{\sum_{0<j<m} exp(f(u(t-j))} \quad (2)$$

where f is the scoring function. In the model, f is the linear function of the input $u(t-i)$

$$f(u(t-i)) = W^T u(t-i) \quad (3)$$

where W is a trainable parameter. The output $attentive\_u$ after the attention layer is the weighted sum of the input sequence.

$$attentive\_u = \sum_i \alpha_i u(t-i) \quad (4)$$

4.3 Context Modeling

This paper evaluates different methods to learn the context representation using AMs.

Max This method applies max-pooling on top of the utterance representations which spans all the contexts and the embedding dimension.

Input This method applies the attention mechanism directly on the utterance representations. The weighted sum of all the utterances represents the context information.

GRU-Attention This method uses a sequential model with GRU cells on top of the utterance representations to learn the relationship between the context and the current utterance over time. The output of the hidden layer of the last state is the context representation.

		WA	UWA	Neu	Joy	Sad	Fea	Ang	Sur	Non
NB	**Friends**	54.9	**57.4**	51.4	57.5	50.0	-	100.0	76.3	36.8
	EmotionPush*	67.3	**57.3**	68.7	76.2	87.5	-	-	100.0	26.7
CNN	**Friends**	59.2	45.2	64.3	60.2	41.2	21.9	46.6	61.5	20.6
	EmotionPush*	71.5	41.7	80.8	46.9	43.7	0.0	27.0	53.8	40.0
CNN-BiLSTM	**Friends**	**63.9**	43.1	74.7	61.8	45.9	12.5	46.6	51.0	8.8
	EmotionPush*	77.4	39.4	87.0	60.3	28.7	0.0	32.4	40.9	26.7
GRU-Attention	**Friends**	57.1	33.4	85.2	46.0	-	3.1	45.1	51.8	30.0
	EmotionPush*	**78.2**	46.8	91.4	65.7	29.9	-	-	58.3	47.1

Table 2: Weighted and unweighted accuracy on Friends and EmotionPush

5 Experiments

For the experiments, neural architectures apply an end-to-end learning approach, i.e., with minimum text preprocessing. For cross-validation, the splitting strategy divides them by the dialogues, similar to (Chen et al., 2018).

The challenge evaluates the performance using the metrics weighted accuracy (WA) and unweighted accuracy (UWA), as defined in equations 5 and 6.

$$\text{WA} = \sum_{l \in C} s_l a_l \tag{5}$$

$$\text{UWA} = \frac{1}{|C|} \sum_{l \in C} a_l \tag{6}$$

where a_l denotes the accuracy of emotion class l and s_l denotes the percentage of utterances in emotion class l.

The Table 2 shows the experimental results including baselines for the emotion detection task. This paper evaluated a Multinomial Naive Bayes (NB) model and the proposed Attention Model (AM). Surprisingly, NB model outperforms neural models for UWA metric in both datasets with 57.4% and 57.3%. This result could be related to the size of the dataset since neural architectures take advantage of learning on large-scale datasets.

The attention model performs well on the EmotionPush dataset but fails to improve on the Friends datasets for WA metric. Further evaluation of the results as depicted in the Fig. 3, show that the label imbalance for *neutral* emotion affects the predictions of other labels.

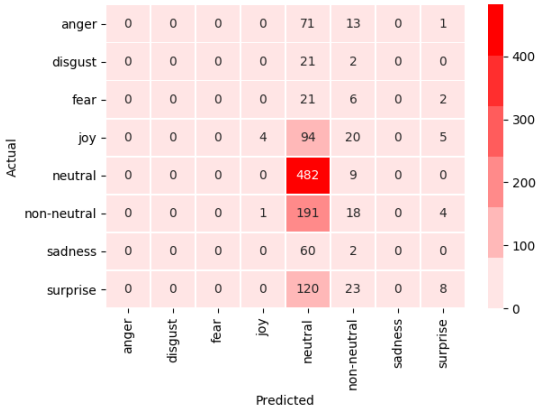

Figure 3: Confusion matrix for the results of Attention Model on the Friend dataset.

6 Conclusions and Future Work

This paper presents a neural attention model for the EmotionX challenge. Attention models take advantage of the context information in conversational datasets for recognizing emotions. The results obtained through several experiments outperformed the baseline methods in some metrics in the emotionPush dataset and was less effective on the Friends dataset.

Despite the promising results with Attention Models, the model struggles to accurately detect ambiguous utterances in the Friend dataset due to the label imbalance and the small scale of it. As such, large-scale conversational corpus with annotated data becomes crucial for pushing the frontiers in emotion recognition.

Attention methods have the potential to provide improved accuracy in detecting emotions in conversational datasets, and future work can explore additional strategies for Attention Models.

References

Heike Adel and Hinrich Schütze. 2016. Exploring different dimensions of attention for uncertainty detection. *arXiv preprint arXiv:1612.06549* .

James R Averill. 1980. A constructivist view of emotion. In *Theories of emotion*, Elsevier, pages 305–339.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* .

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*. pages 577–585.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53(4):712.

David Golub and Xiaodong He. 2016. Character-level question answering with attention. *arXiv preprint arXiv:1604.00727* .

Matthew Henderson, Milica Gašić, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young. 2012. Discriminative spoken language understanding using word confusion networks. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, pages 176–181.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Anil K Jain and Stan Z Li. 2011. *Handbook of face recognition*. Springer.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913* .

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584* .

Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. 2009. Towards text-based emotion detection a survey and possible improvements. In *Information Management and Engineering, 2009. ICIME'09. International Conference on*. IEEE, pages 70–74.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827* .

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. *arXiv preprint arXiv:1708.02561* .

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* .

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* .

Sheng-syun Shen and Hung-yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *arXiv preprint arXiv:1604.00077* .

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–373.

Shih-Ming Wang, Chun-Hui Li, Yu-Chun Lo, Ting-Hao K Huang, and Lun-Wei Ku. 2016. Sensing emotions in text messages: An application and deployment study of emotionpush. *arXiv preprint arXiv:1610.04758* .

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1480–1489.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193* .

Matthias Zimmermann. 2009. Joint segmentation and classification of dialog acts using conditional random fields. In *Tenth Annual Conference of the International Speech Communication Association*.