PKUSE at SemEval-2019 Task 3: Emotion Detection with Emotion-Oriented Neural Attention Network

Luyao Ma[1,2*], **Long Zhang**[1,2*], **Wei Ye**[1], **Wenhui Hu**[1]
[1]National Engineering Research Center for Software Engineering, Peking University
[2]School of Software and Microeconomics, Peking University
{maluyao, zhanglong418, wye, huwenhui}@pku.edu.cn

Abstract

This paper presents the system in SemEval-2019 Task 3, "EmoContext: Contextual Emotion Detection in Text". We propose a deep learning architecture with bidirectional LSTM networks, augmented with an emotion-oriented attention network that is capable of extracting emotion information from an utterance. Experimental results show that our model outperforms its variants and the baseline. Overall, this system has achieved 75.57% for the microaveraged F1 score.

1 Introduction

With the rapid development of social media platforms like Twitter, a huge number of textual dialogues has increasingly emerged. It is a challenge for chat bots to generate responses based on user emotions which can avoid inappropriate conversations. Emotion detection in text (Chatterjee et al., 2019) is a research area within Natural Language Processing which is aim to detect the emotion of user expressed in text.

Many techniques have been proposed, Wang et al.,Hasan et al.,Liew and Turtle used feature engineering to extract features manually. In this area, deep learning-based approches have performed well in recent years. Some methods (Wöllmer et al., 2010; Metallinou et al., 2012; Poria et al., 2017; Chernykh et al., 2017) used recurrent neural network to model the sequence of utterances for emotion detection. However, those models did not highlight the emotion-related parts. We use attention mechanism to locate the parts expressing emotions in the utterance.

The Task3 in Semeval-2019 is to detect contextual emotions in text. For this task, we propose a deep learning approach which is a combination of Long Short-Term Memory network and attention mechanism.

The rest of the paper is organized as follows: Section 2 provides system overview. Section 3 describes our approach in detail. Our experiment is discussed in Section 4. We conclude our work in Section 5.

2 System Overview

2.1 Text Preprocessing and Word Embedding

We use word embeddings as input to the model. Word embeddings are distributed vector presentations of words (Mikolov et al., 2013), capturing their syntactic and semantic information. A good word embedding can get a better classification performance. After comparison, we find that the effect of the GloVe (Pennington et al., 2014) is the best, but when we turn the word into a word vector, we find a lot of cases that are out of vocabulary(oov). In view of that, we preprocess the data as follows:

- The emoji used in the chat can better express human emotions, so we turn them into corresponding emotion words and add them to the sentence, which not only solves the oov, but also increases the emotion information in the sentence
- Several emoticons are replaced by the tokens "happy", "sad", "angry"
- All words are lowercased

2.2 Long Short-Term Memory

LSTM is a special form of threshold RNN (Hochreiter and Schmidhuber, 1997), which is designed to deal with sequential data by sharing its internal weights across the sequence. Different from the structure of RNN, LSTM has three gates:

*These authors have contributed equally to this work.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 287–291
Minneapolis, Minnesota, USA, June 6–7, 2019.

an input gate i_t, a forget gate f_t, an output gate o_t and a memory cell c_t. Their effect is to allow the network to store and retrieve information over long periods of time.

In our approach, we use the bidirectional LSTM model to better capture the contextual information in sentences. Schuster and Paliwal shows that the bidirectional structure has better performance in classification experiments. In order to better handle the relations among the utterances of a dialogue, we use the bc-LSTM architecture(Poria et al., 2017) to process the dialogue-level classification. The architecture preserves the sequential order of utterances when constructing the dialogue representation.

2.3 Attention Mechanism

The attention mechanism was originally applied to image recognition (Itti and Koch, 2001; Mnih et al., 2014), mimicking the focus of the eye moving on different objects when the person viewed the image. Similarly, when people read an article, their attention to each part of the text is different. The attention mechanism imitates human behavior, giving each feature different weights. With the weight of a feature being greater, the contribution of this to current recognition becomes greater. Neural networks with attention mechanism have been applied in many tasks of NLP, including machine translation (Bahdanau et al., 2014; Luong et al., 2015) text summarization (Rush et al., 2015) text classification (Yang et al., 2016) sentiment classification (Chen et al., 2016) and stance classification (Du et al., 2017).

When learning the representations of text sequences, word embeddings are the most effective intermediate representations for capturing semantic information. We embed the classification label and word into the same semantic space, and then construct the semantic relatedness of them according to the similarity of word embeddings. Our model obtains the attention weights of the words through the emotion-oriented attention network, which highlights the emotion words, thus improving the performance of the emotion classification.

3 Model Description

Our model has two steps as follows: 1. Extract the features of each utterance in the dialogue 2. Construct the representation of the dialogue by the features of three utterances for emotion classification.

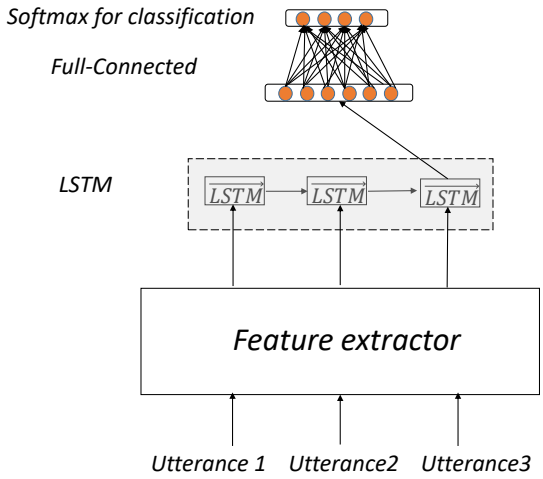

Figure 1: Architecture for emotion classification.

In the feature extraction step: the embedding of each utterance is fed into the BiLSTM layer to construct the word representation of each word; meanwhile we obtain the attention weight of the corresponding word by the emotion-oriented attention network. We use the inner product of them to represent the word, and then feed it into the BiLSTM layer. Finally, we get the representation of each utterance after the pooling operation(Fig. 2).

In the classification step: the features of the three utterances obtained from the previous step are fed into the LSTM layer as timing information for emotion classification(Fig. 1).

3.1 Embedding Layer

An input sequence $\boldsymbol{X}$ of length T is composed of word tokens: $\boldsymbol{X} = \{x_1, ..., x_T\}$. Each token x_t is replaced with the corresponding vocabulary index $V(t)$. The embedding layer transforms the token into vector $e_t \in \mathbb{R}^d$ which is selected from the embedding matrix E according to the index, where d is the dimensionality of the embedding space.

In order to highlight the emotion words in the sequence, we append the word embedding vector of "emotion" to the embedding of each word in original text. The emotion-augmented embedding of a word t is the concatenation of the embedding vector e_t and the emotion representation e_z,

$$e_t^z = e_t \| e_z \tag{1}$$

where $\|$ denotes the concatenation operation, and then the dimention of e_t^z is $2d$.

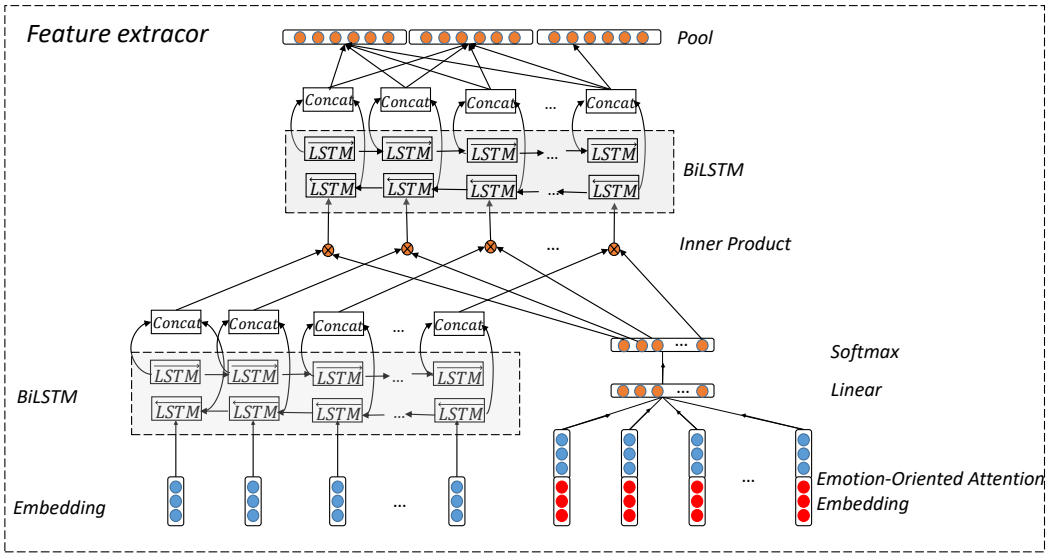

Figure 2: Feature extractor of an utterance with emotion-oriented attention network.

3.2 BiLSTM Layer

The LSTM reads the sequence $\boldsymbol{X}$ only in one direction. We use a bidirectional LSTM to get annotations of words by summarizing the contextual information from both directions. A bidirectional LSTM consists of a forward LSTM $\overrightarrow{f}$ that reads the sentence from x_1 to x_T and a backward LSTM $\overleftarrow{f}$ that reads the sentence from x_T to x_1. We obtain the annotation h_t for each word x_t, by concatenating the forward hidden state $\overrightarrow{h}_t$ and the backward one $\overleftarrow{h}_t$,

$$h_t = \overrightarrow{h}_t \| \overleftarrow{h}_t \tag{2}$$

3.3 Emotion-Oriented Attention Network

In the task, the emotion words in the conversation are vital for classification, which cannot be captured by the BiLSTM. In order to highlight the emotion-related words in the utterance, we design an attention mechanism which increases the weight of the important words on the basis of the BiLSTM and contributes more to the classification decision.

We apply a linear layer to convert the emotion-augmented embedding of a word e_t^z to a scalar value u_t, and then get a normalized importance weight α_t through a softmax function. This weight is producted with the word representation h_t to get a weighted word representation v_t for each word.

$$u_t = W_u e_t^z + b_u \tag{3}$$

$$\alpha_t = \frac{e^{u_t}}{\sum_{i=1}^{T} e^{u_i}} \tag{4}$$

$$v_t = \alpha_t h_t \tag{5}$$

3.4 Pooling Layer

From the idea of network in network, we use global maxpooling, global averagepooling and last tensor for the matrix f output of the BiLSTM layer. Maxpooling can get the most important features of all features (Scherer et al., 2010). Averagepooling can get the most common features of all features. The last tensor output $\bar{l}$ of the matrix f can obtain the semantic information of the sentence in forward and backward through BiLSTM.

The utterance representation z is obtained by the concatenation of the max vector $\overline{m}$, the average vector $\overline{a}$ and the last vector $\bar{l}$.

$$z = \overline{m} || \overline{a} || \bar{l} \tag{6}$$

3.5 Emotion Classification

We use the three utterance representations obtained by feature extractor shown in Figure 2 to construct the dialogue representation. The three utterance representations $[z_1, z_2, z_3]$ are fed into the LSTM, and the last time-step hidden state h_3 of the LSTM is regarded as the dialogue representation r. We pass it to a fully-connected network with a softmax activation function. This

layer obtains a normalized four-dimensional vector through the nonlinear transformation function of the input vector.

$$p = \text{softmax}(W_f h_3 + b_f) \quad (7)$$

where W_f and b_f are the weights and bias terms of the fullly-connected layer.

4 Evaluation

4.1 Data

The datasets are provided by Semeval-2019 Task 3. Table 1 gives an overview of the datasets. All the conversations are collected from twitter. The conversations consist of user 1's tweet, user 2's response to the tweet and user 1's response to user2. The label is the emotion of the third turn that human judges mark after considering the context of three rounds of dialogue.

Dataset	Happy	Sad	Angry	Others	Total
Training	4243	5463	5506	14948	30160
Validation	425	547	551	1495	3018
Test	284	250	298	4677	5509

Table 1: Datasets for Semeval-2019 Task 3.

4.2 Experiments

The model is implemented using Keras 2.0 (Chollet et al., 2017). We experiment with Stanford's GloVe 300 dimensional word embeddings trained on 840 billion words from Common Crawl. Our model is trained with Adam Optimizer (Kingma and Ba, 2014) with initial learning rate of 0.001 and batch size of 64. We use BiLSTMs with hidden state size 256, with dropout rate 0.5 on the first BiLSTM layer and dropout rate 0.3 on the second one to prevent our neural network from overfitting (Srivastava et al., 2014).

In our task, the size of samples for each class is not balanced, which will result in the model tending to be biased toward the majority class with poor accuracy for the minority class. For this, we adjust the parameter 'class_weight' to weight the loss function of each class during training. This can be useful to tell the model to "pay more attention" to samples from an under-represented class. In this case, we set the parameter 'class_weight' (Happy : 2, Sad : 1, Angry : 1, Others : 4)

4.3 Result and Analysis

In order to evaluate the effect of the emotion-oriented attention network and the balanced class weights, we compare our approach with its variants and the baseline.

Variant1: The variant does not adjust the parameter 'class_weight'.

Variant2: The variant changes the emotion-oriented attention network with the attention machanism used in (Yang et al., 2016)

Variant3: The variant removes the emotion-oriented attention network from the model.

model	Happy	Sad	Angry	MicroF1
Baseline	0.5461	0.6149	0.5945	0.5861
Variant1	0.7082	0.7574	0.7175	0.7264
Variant2	0.6967	0.7683	0.7375	0.7330
Variant3	0.7051	**0.8113**	0.7182	0.7423
Our Model	**0.7138**	0.8088	**0.7500**	**0.7557**

Table 2: Performance of our system and its variants.

Table 2 shows that our model outperforms the other variants which are all above the baseline 0.5861 for the micro-averaged F1 score. Variant3 has the best performance on 'Sad' class and our model has the best performance on two classes and micro-averaged F1 score.

To validate that our model has the ability to capture the emotion-related parts of an utterance, we visualize the weights of attention for the following three dialogues. Figure 3 shows that the emotion words are highlighted in the dialogues, such as 'Haha', 'funny', 'cool', 'like', 'hate', 'felt', 'bad', 'SORRY', but the model also highlights some trivial words, such as 'Give'.

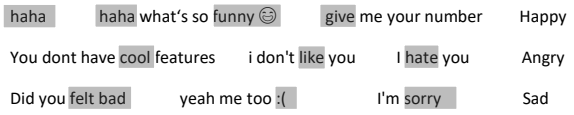

Figure 3: Visualization of attention of examples.

5 Conclusion

In this paper, we proposed an emotion-oriented neural attention network for Semeval-2019 Task 3. The network use the attention mechanism to select the emotion-related parts in the utterances. The classification performance of our model is better than its variants and the baseline. Meanwhile, the visualization shows that the model has captured more decision-making information in the dialogue.

References

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

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext: Contextual emotion detection in text. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659.

Vladimir Chernykh, Grigoriy Sterling, and Pavel Prihodko. 2017. Emotion recognition from speech with recurrent neural networks. *arXiv preprint arXiv:1701.08071*.

François Chollet et al. 2017. Keras https://github.com/fchollet/keras.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD Workshop on Health Informatics, New York, USA*.

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

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*, pages 73–80.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. 2012. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

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.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

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

Dominik Scherer, Andreas Müller, and Sven Behnke. 2010. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365.

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.