MoonGrad at SemEval-2019 Task 3: Ensemble BiRNNs for Contextual Emotion Detection in Dialogues

Chandrakant Bothe and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg,
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
www.informatik.uni-hamburg.de/WTM/
{bothe, wermter}@informatik.uni-hamburg.de

Abstract

When reading "I don't want to talk to you any more", we might interpret this as either an angry or a sad emotion in the absence of context. Often, the utterances are shorter, and given a short utterance like "Me too!", it is difficult to interpret the emotion without context. The lack of prosodic or visual information makes it a challenging problem to detect such emotions only with text. However, using contextual information in the dialogue is gaining importance to provide a context-aware recognition of linguistic features such as emotion, dialogue act, sentiment etc. The SemEval 2019 Task 3 EmoContext competition provides a dataset of three-turn dialogues labeled with the three emotion classes, i.e. *Happy*, *Sad* and *Angry*, and in addition with *Others* as none of the aforementioned emotion classes. We develop an ensemble of the recurrent neural model with character- and word-level features as an input to solve this problem. The system performs quite well, achieving a microaveraged F1 score (F1μ) of 0.7212 for the three emotion classes.

1 Introduction

Humans might interpret text wrongly when reading sentences in the absence of context, so machines might too. When reading the following utterance,

`Why don't you ever text me?`

it is hard to interpret the emotion where it can be either a sad or an angry emotion (Chatterjee et al., 2019; Gupta et al., 2017). The problem becomes even harder when there are ambiguous utterances, for example, the following utterance:

`Me too!`

one cannot really interpret the emotion behind such an utterance in the absence of context. See Table 1 where the utterance "Me too!" is used in many emotional contexts such as *sad*, *angry*, and *happy* and also in the class *"others"* where none of aforementioned emotions is present.

Analyzing the emotion or sentiment of text provides the opinion cues expressed by the user. Such cues could assist computers to make better decisions to help users (Kang and Park, 2014) or to prevent potentially dangerous situations (O'Dea et al., 2015; Mohammad and Bravo-Marquez, 2017; Sailunaz et al., 2018). Character-level deep neural networks have recently showed outstanding results on text understanding tasks such as machine translation and text classification (Zhang et al., 2015; Kalchbrenner and Blunsom, 2013).

Usually, the utterances are short and contain mis-spelt words, emoticons, and hashtags, especially in the textual conversation. Hence, using character-level language representations can theoretically capture the notion of such texts. On the other hand, the EmoContext dataset is collected from the social media, and so the character language model used in our experiments is also trained on such a corpus (Radford et al., 2017).

We propose a system that encapsulates character- and word-level features and is modelled with recurrent and convolution neural networks (Lakomkin et al., 2017). We used our recently developed models for the context-based dialogue act recognition (Bothe et al., 2018). Our final model for EmoContext is an ensemble average of the intermediate neural layers, ended with a fully connected layer to classify the contextual emotions. The system performs quite well and we ranked on the public leaderboard (MoonGrad team) on CodaLab[1] in the top 35% of the systems (at the time of writing this paper Feb 2019) achieving the microaveraged F1 score (F1μ) of 0.7212 for the three emotion classes.

[1] https://competitions.codalab.org/competitions/19790

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 261–265
Minneapolis, Minnesota, USA, June 6–7, 2019.

id	User 1 turn1	User 2 turn2	User 1 turn3	label
2736	I don't hate you. you are just an AI	i don't hate anyone	me too	angry
2867	everything is bad	whats bad?	me too	sad
4756	I am very much happy :D	Thank you, I'm enjoying it :)	Me too	happy
8731	How r uh	am fine dear and u?	Me too	others

Table 1: Examples from training dataset, where *turn3* is mostly the same while contextual emotion is different.

Label	Train	Dev	Test
	30160	2755	5509
happy	4243	142	284
sad	5463	125	250
angry	5506	150	298
others	14948	2338	4677

Table 2: EmoContext Data Distribution; first row represents the total number of conversations in dataset.

2 Approach

The final model used for the submission to the EmoContext challenge is shown in Figure 1. It is an average ensemble of four variants of neural networks. *Net1* and *Net2* use the input from a pre-trained character language model; *Net3* and *Net4* use GloVe word embeddings as input. All models are trained with Adam optimizer at a learning rate of 0.001 (Kingma and Ba, 2014).

The dataset provided by the EmoContext organizers consists of the 3-turn dialogues from Twitter, where *turn1* is a tweet from user 1; *turn2* is a response from user 2 to that tweet, and *turn3* is a back response to user 2 (Gupta et al., 2017). The data distribution is presented in Table 2. We do not perform any special pre-processing except converting all the data into plain text.

2.1 Character-level RNN Model

The character-level utterance representations are encoded with the pre-trained recurrent neural network model[2] which contains a single multiplicative long short-term memory (mLSTM) (Krause et al., 2016) layer with 4,096 hidden units, trained on ∼80 million Amazon product reviews as a character-level language model (Radford et al., 2017). *Net1* and *Net2* are fed the last vector (LM) and the average vector (AV) of the mLSTM respectively. It is shown in (Lakomkin et al., 2017) that the AV contains effective features for emotion detection. The character-level RNN models (*Net1* and *Net2*) are identical and consist of two stacked bidirectional LSTMs (BiLSTM) followed by an average layer over the sequences computed by final BiLSTM.

2.2 Word-level RNN and RCNN Model

The word embeddings are used to encode the utterances. We use pre-trained GloVe embeddings (Pennington et al., 2014) trained on Twitter[3] with 200d embedding dimension. The average length of the utterances is 4.88 (i.e. ∼5 words/utterance on average) and about 99.37% utterances are under or equal to 20 words. Therefore, we set 20 words as a maximum length of the utterances. *Net3* is stacked with two levels of BiLSTM plus the average layer while *Net4* consists of a convolutional neural network (Conv). Conv in *Net4* over the embedding layer captures the meaningful features followed by a max pooling layer (max), with the kernel size of 5 with 64 filters and all the kernel weights matrix initialized with Glorot uniform initializer (Glorot et al., 2011; Kim, 2014; Kalchbrenner and Blunsom, 2013). The max pooling layer of pool size 4 is used in this setup, the output dimensions are shown in Figure 1. We build a recurrent-convolutional neural network (RCNN) model by cascading the stack of LSTMs and the average layer to model the context.

2.3 Ensemble Model

The overall model is developed in such a way that the outputs of all the networks (*Net1*, *Net2*, *Net3*, and *Net4*) are averaged and a fully connected layer (FCL) is used with *softmax* function over the four given classes. The complete model is trained end-to-end so that, given a set of three turns as an input, the model classifies the emotion labels.

[2] https://github.com/openai/generating-reviews-discovering-sentiment

[3] https://nlp.stanford.edu/projects/glove/

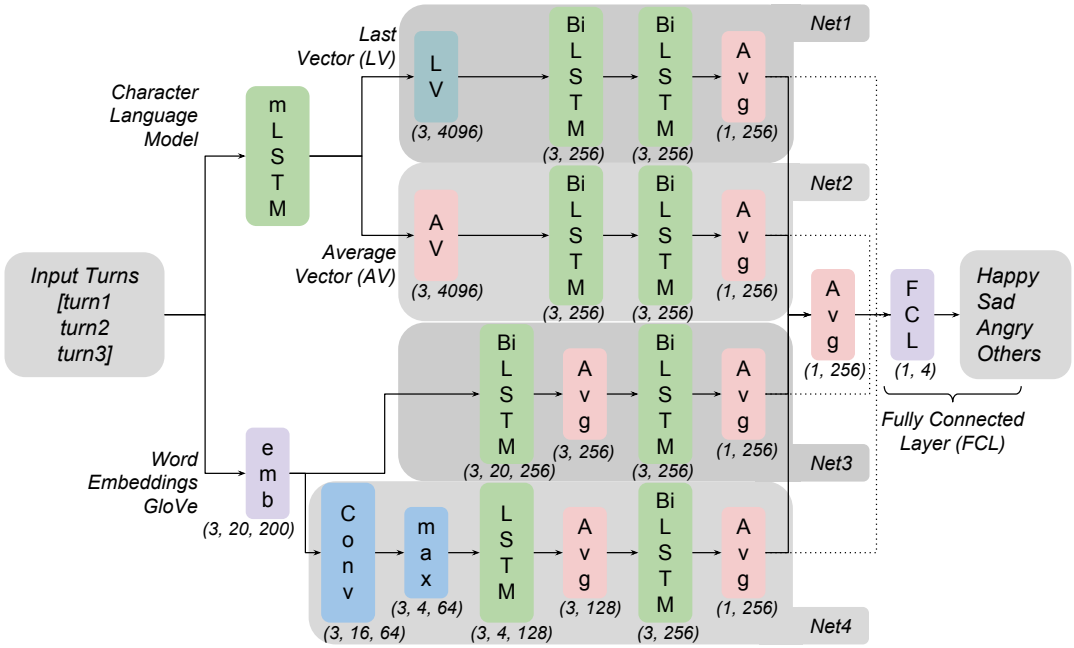

Figure 1: The overall architecture of the contextual emotion detection.

Models	F1μ
Baseline model (organizers)	0.5838
Our proposed model	**0.7212**
happy	0.6893
sad	0.7485
angry	0.7287

Table 3: Result as microaveraged F1 score (F1μ) compared to baseline and F1 score for each emotion.

3 Experiments and Results

The final submitted result to the challenge is shown in Table 3. The metric used for the challenge is the microaveraged F1 score (F1μ) for the three emotion classes, i.e. *Happy*, *Sad* and *Angry*. Our model performance was able compete quite well with the participating teams in the challenge. The main goal to present these experiments is to explore the features used for contextual emotion detection. For the comparison of different language features (character and word), we consider calculating the accuracy over all four classes, in addition to F1μ. The experimental setup developed and each network is tested individually and in an ensemble way. The results are reported in Table 4. When the models train individually, the output of the model being trained is directly connected to the FCL as shown in dotted line in Figure 1. From the results, it is clear that the average vector Char-LM AV Model outperforms the four individual networks. As this model performs well, we also train a single FCL to see the effect of the absence of context. The ensemble models, Char-LM Models (*Net1* and *Net2*) and Word Embs Models (*Net3* and *Net4*) show a clearer pick up on accuracy than individuals. The final ensemble model clearly improves the overall performance. However, we also ensemble the output predictions of all the networks trained individually, and average them at the end. Such ensembling is also effective for the overall improvement in the performance.

Models	Acc (%)	F1μ
Char-LM LV Model (*Net1*)	88.12	0.655
Char-LM AV Model (*Net2*)	89.87	0.694
Char-LM AV Model (*No Context*)	86.25	0.603
Word Embs Model (*Net3*)	88.27	0.665
Word Embs Model (*Net4*)	88.80	0.653
Char-LM Models (*Net1* and *Net2*)	89.59	0.688
Word Embs Models (*Net3* and *Net4*)	87.91	0.692
Final Ensemble Model	**91.63**	**0.721**
Avg. Ensemble Model *(outputs of individual nets)*	**91.71**	**0.721**

Table 4: Results comparing our experimental setups.

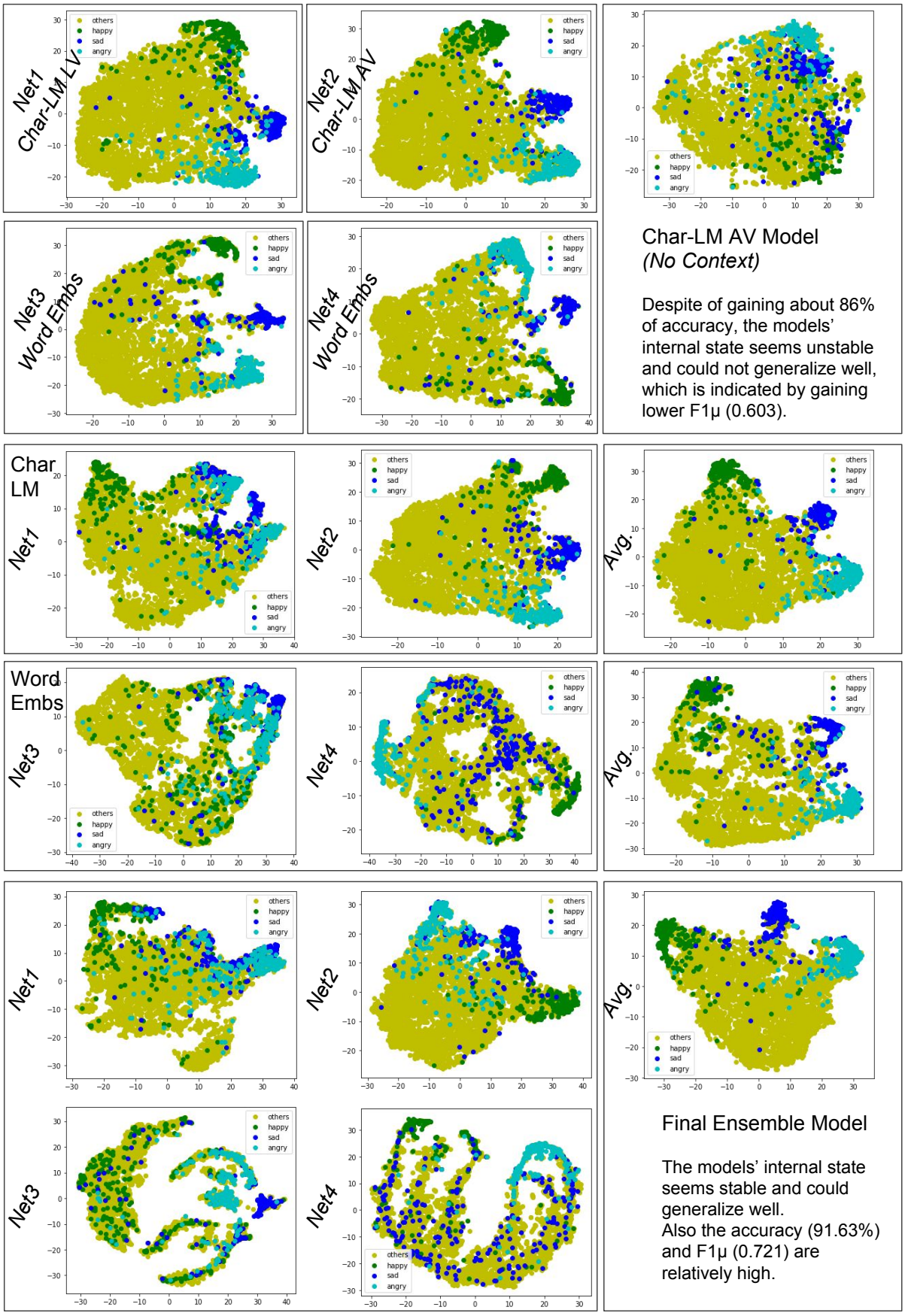

Figure 2: Clustering the intermediate representations of different networks and their average (Avg.) ensembled representations. EmoContext test data is used to generate these representations.

In Figure 2, we demonstrate the intermediate representations taken at the last average layers of the networks on test data and plotted against four classes. We use t-SNE algorithm that converts multi-dimensional (in our case 256) to 2-dimensional arrays. We can notice that the *Net2* Char-LM AV model is quite consistent while other models are a bit unstable in clustering for the given emotions classes. For the final ensemble model, surprisingly, word models become too cluttered, but still contribute to the improvement.

4 Conclusion

The contextual emotion detection is a crucial step towards conversational analysis where emotion can aid the natural language understanding in socio-linguistic studies. Especially in the absence of facial expression and prosodic features, context becomes an important asset for emotion detection in the text. As we can see from the results our model could compete and provide insight to explore different feature representations. The ensemble modelling and transfer learning are effective tools for such a challenging task, specifically, when the given data is small and the labels are not balanced over all the samples.

Acknowledgments

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement No 642667 (SECURE).

References

Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2018. Conversational Analysis using Utterance-level Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the International Conference INTERSPEECH 2018*.

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)*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR*, volume 15, pages 315–323. PMLR.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations. *Proceedings of the Neu-IR 2017 SIGIR Workshop on Neural Information Retrieval*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, ACL*, pages 119–126.

Daekook Kang and Yongtae Park. 2014. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, 41(4):1041–1050.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the Conference on EMNLP*, pages 1746–1751.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. 2016. Multiplicative LSTM for sequence modelling. *Workshop track of Proceedings of the International Conference on Learning Representations*.

Egor Lakomkin, Chandrakant Bothe, and Stefan Wermter. 2017. GradAscent at EmoInt-2017: Character and Word Level Recurrent Neural Network Models for Tweet Emotion Intensity Detection. In *Proceedings of the 8th Workshop WASSA at the Conference EMNLP*, pages 169–174. ACL.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics (*Sem)*, Vancouver, Canada.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2(2):183–188.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the Conference on EMNLP*, pages 1532–1543.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to Generate Reviews and Discovering Sentiment. *arXiv: 1704.01444*.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion Detection from Text and Speech - A Survey. *Social Network Analysis and Mining*, 8(1):28.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, pages 649–657.