In: Proc. of the 26th Int. Conference on Artificial Neural Networks (ICANN), Alghero, Italy, September 11-14, 2017
The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-68612-7_54

Dialogue-based neural learning to estimate sentiment of next upcoming utterance

Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter

Knowledge Technology Institute
Department of Informatics, Universität Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{bothe,magg,weber,wermter}@informatik.uni-hamburg.de
www.informatik.uni-hamburg.de/WTM/

Abstract. In a conversation, humans use changes in a dialogue to predict safety-critical situations and use them to react accordingly. We propose to use the same cues for safer human-robot interaction for early verbal detection of dangerous situations. Due to the limited availability of sentiment-annotated dialogue corpora, we use a simple sentiment classification for utterances to neurally learn sentiment changes within dialogues and ultimately predict the sentiment of upcoming utterances. We train a recurrent neural network on context sequences of words, defined as two utterances of each speaker, to predict the sentiment class of the next utterance. Our results show that this leads to useful predictions of the sentiment class of the upcoming utterance. Results for two challenging dialogue datasets are reported to show that predictions are similar independent of the dataset used for training. The prediction accuracy is about 63% for binary and 58% for multi-class classification.

1 Introduction

In human-robot interaction, one of the primary concerns is safety. In this paper, we address safety as the condition of being protected from or unlikely to cause danger or injury. A mobile robot serving a wrong drink, coffee instead of water, in a cup might be an acceptable mistake, whereas serving the drink in a broken cup might be an unacceptable risk. When the robot is verbally instructed to perform this action, most probably the user also tells the robot that there is a danger or a chance of risky situation.

Early recognition of hazards is crucial for safety-related control systems, such as protective or emergency stop, which is an essential feature for personal care robots [21]. The main goal of our research is to study the early detection of safety-related cues through language processing. In the case of a wrong robot action, the user might respond with an utterance which, although often not easily understandable for the robot, carries feedback information for the last action performed, which can help to understand the situation [12, 23].

A possible conversation is shown in Figure 1, the robot (R) perceives a sentence from the person (P) with neutral sentiment and responds with a query

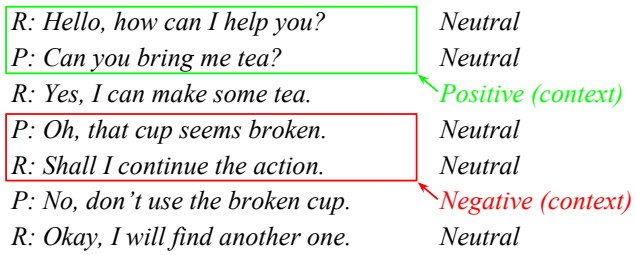

Fig. 1. Example for preparing the contexts: labeled by sentiment analyser, previous two utterances of the positive and negative class are taken as context.

whether this means it should continue. Expecting a positive reply in case everything is ok, the next utterance has a negative sentiment. Without understanding the meaning of a sentence, the robot can stop or revert the last action just on the basis of a failed response sentiment prediction. Furthermore, an estimate of the user's response sensitivity is necessary when the robot needs to ask safety-critical questions [7].

Our goal is, as a first step, to learn from spoken language dialogues to predict the sentiment of the next upcoming utterance. As shown in Figure 1, we use two utterances as context, capturing a sequence with both speakers, to predict the next utterance sentiment from the first speaker. Long short-term memory networks (LSTM) have shown good performance on the text-classification tasks (e.g. [2]) learning long-term dependencies. Since we want to extend our model to longer contexts, we choose those networks and show that they could successfully learn to estimate the sentiment of the next upcoming utterance.

2 Related work

Responses from humans in an interaction have been used in various ways in human-robot scenarios. In student/teacher learning scenarios, to facilitate learning, a teacher gives positive and negative feedback depending on the success of the student [12]. Weston [23] has shown that the positive-negative sentiment in the teacher's response helps to guide the learning process. Other work [20] describes context-sensitive response generation in the field of language understanding and generation. They report that there is a lack of reflecting the agents intent and maintaining the consistency with the sentiment polarity. This consistency of polarity means that unpredicted changes in polarity may be cues for changing situations, so monitoring the sentiment over a dialogue can not only be used for simple feedback signals but give evidence on, maybe not yet otherwise perceivable, changes in the environment.

Sentiment analysis is an important aspect of the decision-making process [17] and thus has received much attention in the scientific community. With vast amounts of data available for analysis, many methods have been explored

recently, e.g. [10, 22]. Deep learning has given rise to some new methods for the sentiment analysis task, outperforming traditional methods [19, 5]. Different NLP tasks have been performed independently and in a unified way using deep neural networks [4]. Especially in the field of text classification, the strength of neural network approaches is evident, e.g. convolutional neural networks [11] or recursive and recurrent neural networks [2, 19]. A fixed-size context window can solve the problem of the variable length of language text sequences, but this fails to capture the dependencies longer than the window size. Recurrent neural networks have the ability to use variable sequence length, and especially LSTM networks have shown good performance [5].

The accessibility of large unlabelled text data can be utilised to learn the meaning of words and the structure of sentences and this has been attempted by word2vec [16]. The learned word embeddings are used for creating lexicons and have a reduced dimensionality compared to traditional methods. This approach has also been used for learning sentiment-specific word embedding for sentiment classification [14]. Our approach utilises word embeddings to feed an LSTM network similar to [2] in order to learn sentiment prediction.

3 Approach

3.1 Datasets

We have used two spoken interaction corpora for training our model from two very different sources, child-adult interaction and movie subtitles. The first is the child language component of the TalkBank system, called CHILDES[1] [15], where different child and adult speakers converse on daily issues. In this dataset, we selected the conversations with children of age 12 and above, which have significant verbal interaction and less grammatical mistakes [3]. The other corpus is the Cornell Movie-Dialogues corpus [6], which is more structured, i.e. it is more grammatically correct, and is also larger than the child-interaction corpus.

As our goal is to predict sentiment from a context as shown in Figure 1, we need sentiment annotation of the utterances. The child-interaction corpus (CHI) already has word-level sentiment annotation, while the movie dialogues corpus (MDC) has none. We thus used the natural language toolkit's [13] Vader sentiment analysis tool [9] to create sentiment labels for each utterance. To avoid imbalanced classes in our data, we empirically adjusted the thresholds of the sentiment level to 0.2 and 0.6 on the scale of 0 to 1 for both positive and negative classes. Data samples were now extracted by selecting an utterance with a given sentiment as ground truth and saving the previous two utterances as context. We have created datasets for two experiments, creating contexts from utterances with either negative/positive, or negative/neutral/positive classes. The dataset details are shown in Table 1. While taking the previous utterances for each sample, we have the overlapping of utterances in the contexts, i.e. one utterance may appear in two contexts. The two data-sets are processed for binary (pos-neg) and multi-class (pos-neu-neg) classification.

[1] `http://childes.talkbank.org` or `http://childes.psy.cmu.edu`

Table 1. Dataset details

Datasets	CHI	MDC
Raw utterances	11.1k	304k
Contexts (pos-neg)	4.1k	189k
Contexts (pos-neu-neg)	6.2k	283k

3.2 Model

For the experiments, we used the well-established recurrent long short-term memory (LSTM) neural network [8], a special form of recurrent neural network, shown in Figure 2(a). The sequence of the words represented by their numeric indices in a dictionary, is first fed into the embedding layer which is implemented as standard MLP layer, as shown in Figure 2(b). The embedding layer randomly initializes the normalised vectors, or can utilize already pretrained embeddings, to represent each word index by a real-valued vector of a given size of the embedding dimension which is then fed into the LSTM layer.

The LSTM unit receives an embedded word x as an input and outputs a sentiment prediction y. It maintains a hidden vector h and a memory vector in cell c responsible for controlling state updates and outputs. The LSTM consists of a memory cell c, an input gate i, a forget gate f, and an output gate o, which are updated at time step t as follows:

$$f_t = \sigma\left(W_f * h_{t-1} + I_f * x_t + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_i * h_{t-1} + I_i * x_t + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o * h_{t-1} + I_o * x_t + b_o\right) \tag{3}$$

$$\tilde{c}_t = tanh\left(W_c * h_{t-1} + I_c * x_t + b_c\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + \tilde{c}_t \tag{5}$$

$$h_t = o_t \odot tanh\left(c_t\right) \tag{6}$$

where σ is the *sigmoid* function, W_f, W_i, W_o, W_c are recurrent weight matrices, I_f, I_i, I_o, I_c are the corresponding projection matrices and b_f, b_i, b_o, b_c are learned biases. The weight-projection matrices and bias vectors are initialized randomly and learned during training. The gating functions of the LSTM helps this RNN to mitigate the vanishing and exploding gradient problems and to train the model smoothly. As an output, we get a hidden vector representation (h) of the entire sequence of words which is then used as an input to a classifier. In the sequence classification setup as shown in Figure 2(b), given the current activation function in the hidden state h_t, the RNN generates the output according to the following equation:

$$y_t = g\left(W_{out} * h_t\right) \tag{7}$$

where $g(.)$ denotes an output activation function, in our case a *softmax* function that gives the normalized probability distribution over the possible classes, and W_{out} is an output weight matrix which can be stored to make the predictions.

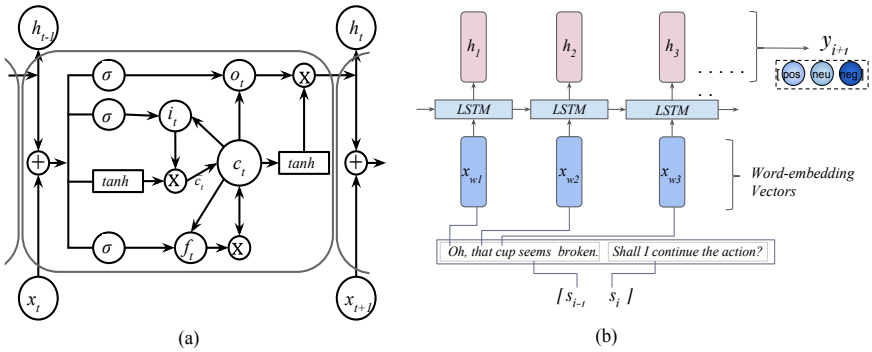

Fig. 2. (a) The long short-term memory (LSTM) unit with (b) our classification setup. Biases are ignored for simplicity.

3.3 Experiments and Results

Our aim is to recognise the sentiment polarity of the upcoming utterance, given the recent utterances as the context. We have trained our classifier by concatenating the context utterances and using the label of the utterance following this context as the training signal. The utterances have been labeled by the sentiment analysis for either binary or multi-class classification as shown in Table 1. The input to the network was always concatenated utterances and the prediction for the upcoming utterance was taken from the classified output of an LSTM at the end of the input sequence. The model was implemented using the Keras Python library and Theano [1]. The input sequence length was fixed to the maximum length in the utterances and padding was used to make them of the same length.

The training was done using categorical crossentropy as the loss function, using stochastic gradient descent as the optimization method. Learning rate and the number of hidden units were empirically determined for both datasets. The hidden layer dimension was 64 for CHILDES and 512 for Movie-Dialogues corpus. We randomly initialized the word embedding vectors with the dimension of 10 and 100 for CHILDES and 100 for the other, and we also used the pre-trained GloVe vectors of dimension 100 [18]. We trained the model on both the datasets as described before and for two different set-ups. Each dataset was split into training, validation and test data with a 60%-20%-20% split. The summary of the test data prediction accuracies is shown below in Table 2.

Table 2. Prediction accuracy on test data

Different setups	Random guess	Trained embeddings			GloVe embeddings (100d)	
		CHI(10d and 100d)		MDC(100d)	CHI	MDC
Binary	50.00%	59.30%	59.06%	52.44%	63.36%	54.97%
Multi-class	33.33%	54.60%	54.56%	48.36%	58.13%	51.71%

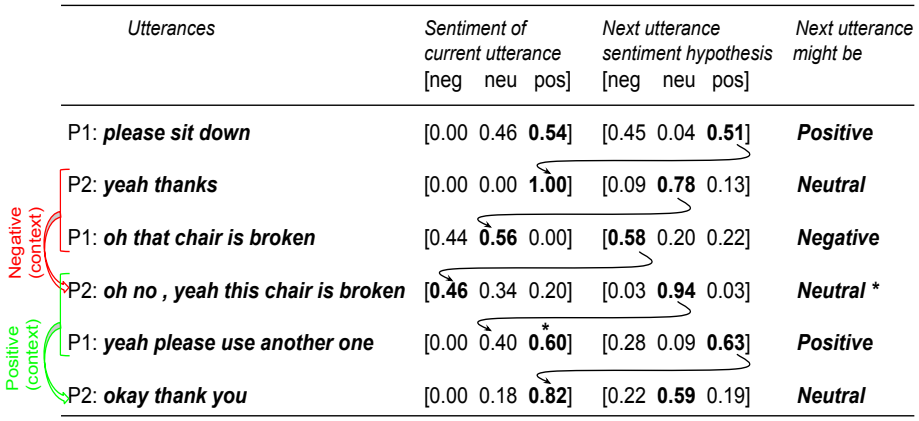

Fig. 3. Test example: prediction on some utterances
* indicates that the sentiment recognition does not match the actual.

The use of pre-trained embedding shows more accuracy than the random initialization, also using different embedding dimensions produced very similar results. We also implemented a simple chat-bot in Python, that receives the utterances sequentially, to evaluate the trained model on a dialogue and monitor the changing hypothesis of the sentiment of the upcoming utterances.

In Figure 3, we present an example from test data. The utterances from the conversation are processed one by one, and the progression of the statements is shown with the predicted hypothesis and the ground-truths. Bold values in the array [neg neu pos] represent the detected class, for the sentiment hypothesis of the current and the next utterance. We also show two related contexts, positive (green) and negative (red). For example, the utterance *"oh no, yeah this chair is broken"* has a negative sentiment label and the model has the correct prediction hypothesis. We can also see that the model failed to predict the positive class for the utterance *"yeah please use another one"*.

Looking at the details of the distributions, the unpredicted increase in negative sentiment for the sentence *"oh that chair is broken"*, although overall still classified as neutral (negative), could have been used already to detect a change in sentiment and thus be aware of a possible change in the environment, the safety situation, or just the user's perception of the robot's current action. The same can be said for the misclassified utterance where P2 perceived a negative situation and might have no solution, interpreting the suddenly positive sentiment of P1 in the next utterance to understand that the situation has a solution or has been solved and nothing bad has happened. Overall, the results show that it is possible to derive valuable cues by estimating the sentiment of the next upcoming utterance, and the model can learn to keep track of the sentiment through dialogues. The corpora used were auto-annotated with the standard sentiment analysis tool which led to comprehensible results, although a human-annotated corpus might still lead to better results.

4 Conclusion and future work

We have presented a learning approach to estimate the sentiment of the next upcoming utterance within a dialogue. We have shown that the model can predict the sentiment of an upcoming utterance to a certain degree, taking into account that the used corpora are noisy and no system would be able to reliably predict the upcoming sentiments simply due to the changing nature of human dialogues. Detecting safety-related cues as early as possible is crucial, and a number of false-positives can be accepted (or quickly resolved through a query within the dialogue) if dangers can be avoided when they occur. We think that tracking even a noisy sentiment through a dialogue can have a positive impact on the safety of a robot, especially when combined with a multi-modal system.

While this work focuses on keeping track of the sentiment in dialog-based context learning, our aim is to extend this to different language features containing safety-related cues. Using not only simple auto-annotated sentiment as labels but including annotations based on prosodic features might lead to a better prediction since humans often involuntarily change their voice when perceiving a dangerous situation while speaking. This work presents already a promising step towards the main goal and can provide useful dialogue-based information regarding the current safety context in human-robot interaction.

Acknowledgement

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

References

1. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y.: Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop (2012)
2. Biswas, S., Chadda, E., Ahmad, F.: Sentiment Analysis with Gated Recurrent Units. Advances in Computer Science and Information Technology (ACSIT) 2(11), 59–63 (2015)
3. Clark, E.V.: Awareness of Language: Some Evidence from what Children Say and Do. pp. 17–43. Springer Berlin Heidelberg (1978)
4. Collobert, R., Weston, J.: A unified architecture for natural language processing. In: Proceedings of the 25th International Conference on Machine Learning - ICML '08. vol. 20, pp. 160–167 (2008)
5. Dai, A.M., Le, Q.V.: Semi-supervised Sequence Learning. In: Neural Information Processing Systems (NIPS). pp. 3079–3087. No. 28, Curran Associates, Inc. (2015)
6. Danescu-Niculescu-Mizil, C., Lee, L.: Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL (2011)

7. Fong, T., Thorpe, C., Baur, C.: Collaboration, Dialogue, and Human-Robot Interaction. 10th International Symposium of Robotics Research (Springer Tracts in Advanced Robotics), 255–266 (2003)
8. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9, 1735–1780 (1997)
9. Hutto, C.J., Gilbert, E.: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Association for the Advancement of Artificial Intelligence (Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media), 216–225 (2014)
10. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. COLING '04 Proceedings of the 20th International Conference on Computational Linguistics p. 1367 (2004)
11. Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the Conference on EMNLP pp. 1746–1751 (2014)
12. Latham, A.S.: Learning through feedback. Educational Leadership 54(8), 86–87 (1997)
13. Loper, E., Bird, S.: NLTK: the Natural Language Toolkit. Proceedings of the ACL-2 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics 1, 63–70 (2002)
14. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics pp. 142–150 (2011)
15. MacWhinney, B.: The CHILDES project: Tools for analyzing talk. Lawrence Erlbaum Associates, Inc (1991), `http://childes.psy.cmu.edu/`
16. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013) pp. 1–12 (2013)
17. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(12), 1–135 (2008)
18. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. Proceedings of the Conference on EMNLP pp. 1532–1543 (2014)
19. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In: Proceedings of the Conference on EMNLP. pp. 1631–1642. Association for Computational Linguistics (2013)
20. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. Association for Computational Linguistics (Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL), 196–205 (2015)
21. Tadele, T.S., de Vries, T., Stramigioli, S.: The Safety of Domestic Robotics: A Survey of Various Safety-Related Publications. IEEE Robotics & Automation Magazine 21(3), 134–142 (2014)
22. Wang, S., Manning, C.D.: Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. pp. 90–94 (2012)
23. Weston, J.: Dialog-based Language Learning. In: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems (NIPS) 29. pp. 829–837. Curran Associates, Inc. (2016)