In: *INTERSPEECH 2018*, pp. 996-1000, September 2-6, 2018 in Hyderabad, India. doi:10.21437/Interspeech.2018-2527

Conversational Analysis using Utterance-level Attention-based Bidirectional Recurrent Neural Networks

Chandrakant Bothe, Sven Magg, Cornelius Weber 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,magg,weber,wermter}@informatik.uni-hamburg.de

Abstract

Recent approaches for dialogue act recognition have shown that context from preceding utterances is important to classify the subsequent one. It was shown that the performance improves rapidly when the context is taken into account. We propose an utterance-level attention-based bidirectional recurrent neural network (Utt-Att-BiRNN) model to analyze the importance of preceding utterances to classify the current one. In our setup, the BiRNN is given the input set of current and preceding utterances. Our model outperforms previous models that use only preceding utterances as context on the used corpus. Another contribution of our research is a mechanism to discover the amount of information in each utterance to classify the subsequent one and to show that context-based learning not only improves the performance but also achieves higher confidence in the recognition of dialogue acts. We use character- and word-level features to represent the utterances. The results are presented for character and word feature representations and as an ensemble model of both representations. We found that when classifying short utterances, the closest preceding utterances contribute to a higher degree.

Index Terms: spoken language understanding, conversational analysis, dialogue act classification

1. Introduction

Conversational discourse analysis is an important task for natural language understanding and for building a spoken dialogue system. A conversation consists of several utterances in a sequence. Discourse analysis of the conversation can be conducted by using speech acts where a speech act defines the performative function of an utterance [1]. However, speech acts are context-sensitive, where the context provides information for appropriate interpretation of the speech act [2]. Once the context is taken into account, the question is how many utterances in the context contribute to the current utterance and how do context-utterances affect the interpretation [1, 2, 3, 4]?

We attempt to answer these questions in this research. We propose an utterance-level attention mechanism using a bidirectional recurrent neural network (Utt-Att-BiRNN) for context-based learning in conversational analysis [5, 6, 7, 8]. The proposed model is intended to not only model context-based learning but also to analyze the amount of contributing information in the utterances for the dialogue act (DA) recognition task. We assess the model performance on the Switchboard Dialogue Act (SwDA) corpus [9].

We previously found a significant improvement of using the context-utterances against the simple utterance-level DA classification. As a result, we now investigate the discourse analysis in a conversation with context-based learning using the Utt-Att-BiRNN model. We show that context-based learning is important for the conversational analysis improving the performance by 5% to 8% accuracy over utterance-level classification.

We also show that the proposed model not only improves the performance but also provides higher confidence in the predicted classes. We report the amount of information contributed by the context-utterances. We experiment with two models: utterance-level model and Utt-Att-BiRNN model. We discover that there are many instances that were detected wrongly with both models. These instances are also reported in the results and discussion section along with the samples where the simple utterance-level model fails to predict correctly, as opposed to the Utt-Att-BiRNN model, for the SwDA corpus test set. With this investigation, we might be able to find ambiguously or wrongly annotated utterances.

2. Related work

Previous work in the field of conversational discourse analysis has been attempting to model utterance-level classification of the dialogue acts [10, 11, 12, 13]. However, classifying the DA classes at a single-utterance level might fail when it comes to DA classes where the utterances share similar lexical and syntactic cues (words and phrases) like the *backchannel (b)*, *no-answer (nn)*, *yes-answer (ny)* and *accept/agree (aa)* DA classes. Stolcke et al., 2000 [10] achieve about 71% of accuracy with hidden Markov models on the SwDA test set. Many recent works show that context-based learning, which takes the preceding utterances into account, improves the performance of the proposed models to achieve state-of-the-art results [14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

The context-based learning approach was first proposed to model discourse within a conversation using RNNs. The DA of the current utterance was calculated using the preceding utterances as a context, achieving state-of-the-art results of about 74% of accuracy on SwDA [15, 20]. Kalchbrenner and Blunsom, 2013 [15] represent the utterance as a compressed vector of word embeddings using convolutional neural networks (CNN) and use these utterance representations to model discourse within a conversation using RNNs. Lee and Dernoncourt, 2016 [22] also use recent techniques such as RNNs and CNNs with word-level feature embeddings and achieve about 73% of accuracy. Ortega and Vu, 2017 [20] also use CNNs and RNNs and achieve about 74% of accuracy.

In another line of research, the context-based learning approach processes the whole set of utterances in a conversation, where the model can see past and future utterances to calculate the DA of the current utterance [16, 17]. Ji et al. 2016

[16] use discourse annotation for the word-level language modelling on the SwDA corpus and achieve about 77% of accuracy but also highlight a limitation that this approach is not scalable to large data. On the other hand, this work suggests that a domain-independent language model which is trained on big data might be a solution. In some approaches, a hierarchical convolutional and recurrent neural encoder model are used to learn utterance representations by processing a whole conversation [17, 19]. The utterance representations are further used to classify DA classes using the conditional random field (CRF) as a linear classifier. However, these models might fail in a dialogue system where one can perceive the past utterances, but cannot see future ones.

In a dialogue system for example in human-machine interaction, one can only perceive the preceding utterance as a context but does not know the upcoming utterances. The DA corpus is also annotated by looking at the preceding utterances. Therefore, we use a context-based learning approach where only preceding utterances are considered and regard the 73.9% accuracy [15, 20] on the SwDA corpus as a current state-of-the-art result for this particular task.

3. Experimental setup

3.1. Dataset

Discourse analysis is a very important task in the field of natural language processing and hence there are many dialogue act corpora available [24]. We use the Switchboard Dialogue Act (SwDA[1]) corpus which is annotated with the Dialogue Act Markup in Several Layers (DAMSL) tag set [9, 25]. SwDA is annotated with 42 DA classes. The corpus consists of 1,115 conversations (196,258 utterances) in the training and 19 conversations (4,186 utterances) in the test set [10, 15].

3.2. Utterance representations

We represent each utterance with two different speech-language features: characters and words.

Character representations: The character-level utterance is encoded with a pre-trained character-level language model (LM[2]) [26]. This model consists of a single multiplicative long-short-term memory (mLSTM) network [27] layer with 4,096 hidden units. The mLSTM is composed of an LSTM and a multiplicative RNN and considers each possible input in a recurrent transition function. It is trained as a character language model on ∼80 million Amazon product reviews [26]. We sequentially input the characters of an utterance to the mLSTM and get the hidden vector obtained after the last character and also average the states over all characters in the utterance. We use the average feature vector representations for each utterance in the experiments as it was shown that the average vector over all characters in the utterance works better for dialogue act recognition [23] and for emotion detection [28].

Word representations: Word-level features are important for analyzing the short sentences in a conversation. We use the word-embeddings distributed as part of ConceptNet 5.5[3] as it is designed to represent the general knowledge involved in understanding language and allows the application to better understand the meanings behind the words people use [29]. It has a knowledge graph that connects words and phrases of natural language with labelled edges. The embedding dimension is 300 and averaged over all tokens in the utterance. These embeddings provide the out-of-vocabulary instance rate close to 10 percent and mostly for infrequent words.

[1] https://github.com/cgpotts/swda

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

[3] https://github.com/commonsense/conceptnet-numberbatch

3.3. Utterance-level attention-based BiRNN

First, we present our baseline model as shown in Figure 1(a), it is a simple utterance-level classifier which classifies the utterances with their respective labels (dialogue acts) using a simple feed-forward neural network with backpropagation. The Utt-Att-BiRNN model is shown in Figure 1(b), for which the main components are the bidirectional recurrent neural network (*BiRNN*) and *Attention* mechanism.

3.3.1. Bidirectional recurrent neural network

A BiRNN is an extended form of an unidirectional RNN [30], introducing one extra hidden layer [5, 6]. The hidden to hidden layer connections flow into the opposite temporal direction. The model provides forward and backward states with corresponding directions of the hidden layers, as shown in Figure 1(b), and the final result is calculated as follows:

$$h_t^f = f\left(W_h^f h_{t-1}^f + W_u^f u_t + b_h^f\right) \quad (1)$$

$$h_t^b = f\left(W_h^b h_{t-1}^b + W_u^b u_t + b_h^b\right) \quad (2)$$

$$P(y_t|\{u_t, u_{t-1}, ...u_{t-n}\}) = g\left(W_y^f h_t^f + W_y^b h_t^b + b_y\right) \quad (3)$$

where n is the number of utterances in the context for time instance t. W and h are the corresponding weight matrices and hidden vectors, where the superscripts f and b represent forward and backward hidden layer directions respectively. In our scenario, we want the model to learn the context, thus the input consists of the current utterance and the preceding context. If we use a unidirectional RNN model, there might be a chance that the model becomes more attentive to the current utterance only, as sequential information is compressed to the final state. The bidirectional RNN model, on the other hand, exploits the information in all given input utterances by looking back and forth through them. Therefore, our goal is to treat all utterances equally and learn how much each contribute to the final result.

3.3.2. Attention mechanism

Attention mechanism is loosely based on visual attention found in humans, and broadly used in image recognition and tracking [31, 32]. But recently, attention mechanism with RNNs are being used for several natural language processing tasks, such as machine translation and comprehension, speech recognition [7, 33, 34]. We propose the attention mechanism to compute the contribution weights of the utterances for predicting the corresponding class. Given the number (n) of preceding utterances in an input sequence $u = \{u_t, u_{t-1}, ...u_{t-n}\}$, the BiRNN provides the respective hidden vectors $h = \{h_t, h_{t-1}, ...h_{t-n}\}$. The attention layer computes the weights $a = \{a_t, a_{t-1}, ...a_{t-n}\}$ as the contribution for every corresponding input utterance in u using the respective hidden representations h, as depicted in Figure 1(b). Hence, the final utterance representation u_{final} of the utterance sequence in u is formed by a weighted sum of h and a:

$$m = tanh\left(W_h h\right) \quad (4)$$

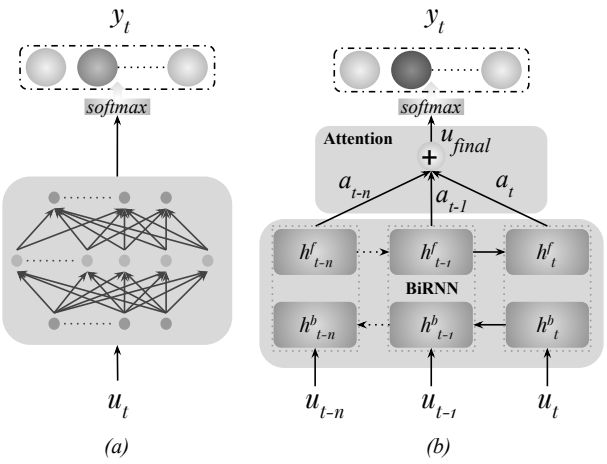

Figure 1: *(a) Our baseline model, (b) Utt-Att-BiRNN model.*

$$a = softmax\left(W_m^T m\right) \tag{5}$$

$$u_{final} = tanh\left(ha^T\right) \tag{6}$$

where W is a trained parameter while W^T being its transpose. We use the $softmax$ function to compute the weights which provides $\sum_{n=0}^{n} a_{t-n} = 1$. It is important for the utterance-level attention mechanism that we normalize a to interpret the amount of contribution for each utterance in u.

3.3.3. Training the model

In the baseline model and the Utt-Att-BiRNN model settings, we use a $softmax$ function to predict a discrete set of classes y on top of the learned u_{final} representations. We use a set of 5 utterances in u, with the current utterance and 4 utterances in the context. A similar study performed in [23] shows the effect of the number of utterances in the context. It was shown that three utterances provide sufficient context, however, we use four context-utterances to provide a large enough window for bidirectional exploration by the RNN, hence $n = 4$.

In all learning cases, we minimize the categorical cross-entropy as we have multiple classes in the DA recognition task. For the baseline model, we use 2 hidden layers with 300 and 100 hidden units respectively. For the proposed model, we use 64 hidden units with the dropout regularizer [35] in the BiRNN hidden layer. As a result, we get 128 hidden units as a concatenation of the h_t^f and h_t^b hidden units. These are the only parameters determined empirically for the classification tasks but all other parameters were learned during training.

The Adam optimizer [36] was used with an initial learning rate 1*e*-4, which decays during training. Early stopping was used to avoid over-fitting of the network, and 15% of training samples were used for validation. We wait for at least 5 iterations over which the accuracy on the validation set does not improve. Typically, both models, baseline and Utt-Att-BiRNN, took about 20 to 30 interations.

Table 1: *Accuracies (in %) on the SwDA test set, baseline with no context (NC) and Utt-Att-BiRNN model with context (WC)*

Models	**NC**	**WC**
Prior related work		
Most common class baseline	31.50	
Stolcke et al., 2000 [10]	71.00	
Kalchbrenner and Blunsom, 2013 [15]		73.90
Lee and Dernoncourt, 2016 [22]		73.10
Ortega and Vu, 2017 [20]		73.80
Our work		
Character LM rep.	71.84	76.47
Word-embeddings mean rep.	71.73	75.43
Concatenated rep.	70.83	76.15
Average char-word-level predictions	71.85	76.84
Average char-word-level & concatenated rep. predictions	71.97	77.42

4. Results and discussion

The baseline and Utt-Att-BiRNN models are trained and tested using both the utterance representations explained in Section 3.2. We report the accuracies on a test set of the SwDA corpus in Table 1. Character LM and word-embeddings mean utterance representations perform quite well for this task. Surprisingly, the word-embeddings mean representations of the utterances used from the ConceptNet seem to show good results given the fact of the low dimensionality of the embeddings (300) compared to character LM (4096) size.

We also experiment with a combined model of these representations in two ways: first by concatenating both representations and use them as an input, and second by averaging the output predictions from both models. Averaging the predictions has shown the best results, and we even found that the average of prediction of models trained with character LM, word-embeddings mean, and concatenated representations give the best of the performance. We can see that context-based learning shows a performance improvement of about 5% on this discourse analysis task.

We examined the SwDA corpus test set and found that there are many instances that were predicted wrongly with both models. The dominant DA classes in the SwDA corpus are Statement-non-opinion (*sd*) and Statement-opinion (*sv*). Table 2 shows the number of samples (*Num*) and the percentage (*pct.*) out of 4,186 utterances. The examples of utterances show that it might be difficult for humans to identify the correct DA class. It shows the ambiguity in the two DA classes *sd* and *sv*, which accounts about 6% of accuracy reduction for both of the models only with these two classes. We also show the effectiveness of the pragmatic model which predicts the correct class when the

Table 2: *The test samples from the SwDA corpus where both classifiers, simple utterance-level and Utt-Att-BiRNN, failed to correctly predict classes (the majority classes, Statement-non-opinion (sd) and Statement-opinion (sv), are reported here). Where Num is a number of samples, GT stands for ground truth, and pct. for percentage.*

GT	**NC**	**WC**	**Num**	**pct.**	**Example of utts**
sv	*sd*	*sd*	198	4.73	Uh, the problem is here But they don't have We're hearing the same
sd	*sv*	*sv*	51	1.22	They're certainly legal, Real long legs, And time consuming,

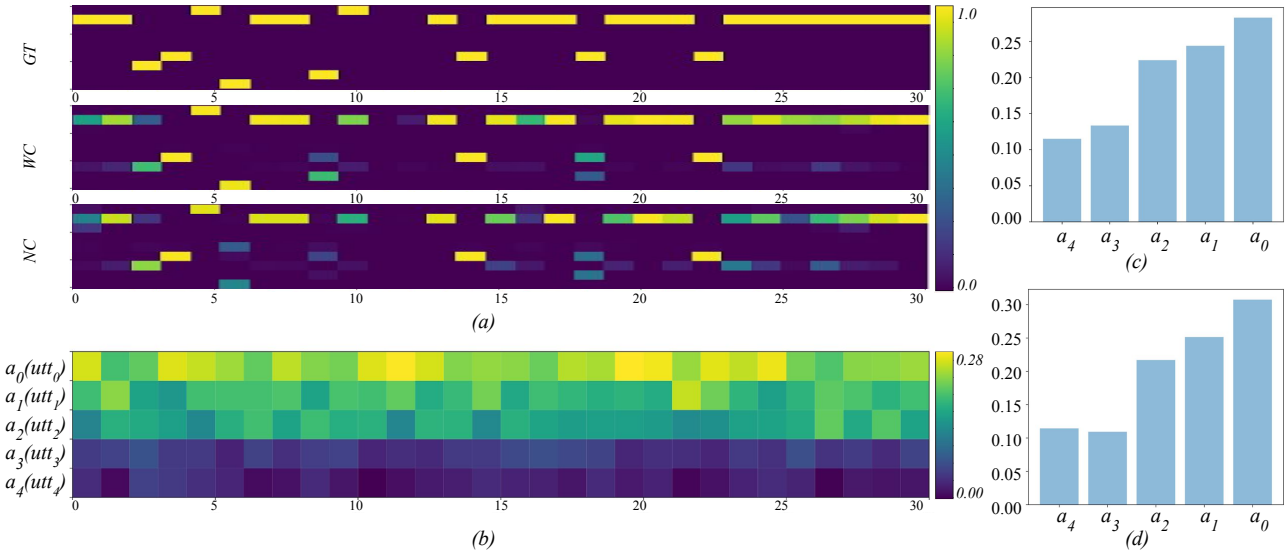

Figure 2: *Effectiveness of the context. (a) Prediction confidence for a batch of 30 sets of utterances: the first row is the ground truth (GT), the second row the predictions with context (WC), and the third row the predictions with no context (NC). We show only 8 of the 42 classes for simplicity on the y-axis and the set of utterances on the x-axis. (b) The contribution of utterances $utt_0, utt_1, ...utt_4$ as the attention weights $a_0, a_1, ...a_4$. (c) The average weight of utterances and (d) in addition averaged over 10 runs to show robustness.*

context is important, see Table 3. For example, if the utterances like "Yes", "Yeah" etc. are followed by Yes-No Question (*qy*), the probability that the second utterance belongs to Yes-Answer (*ny*) is higher than being in Backchannel (*b*) or Abandoned (%). Similar utterances to the *ny* class are used in the Agree/Accept (*aa*) class, but they are usually followed by *sv*, *sd*, *b*, or some other classes. In total, we found 330 samples which constitute around 7.88% of the samples that were correctly recognized by the Utt-Att-BiRNN model but not by the utterance-level model.

However, we also found that the prediction confidence of the Utt-Att-BiRNN model is higher than the utterance-level classifier. Figure 2(a) shows three rows for 30 batches of utterance sets in the DA recognition task: first ground truth, second the predictions of the Utt-Att-BiRNN model, and third the predictions of the utterance-level classifier. The predictions of the Utt-Att-BiRNN model show higher confidence when compared to the predictions of the utterance-level model.

With the help of Utt-Att-BiRNN model we also computed the amount of contribution of the context utterances. As discussed in Section 3.3.2, the attention weights $(a_t, a_{t-1}, ...a_{t-n})$ can be interpreted as the contribution of the utterances, as the u_{final} of the utterance sequence in u is formed by a weighted sum of h and a. Figure 2(b) shows the attention weights $(a_0, a_1, ...a_4)$ that represent the contribution of the corresponding utterances $(utt_0, utt_1, ...utt_4)$. It is clear that the current utterance utt_0 contributes more than others, however, the closest preceding utterances seem to contribute substantially. In Figure 2(c) and 2(d), we can see the average of the weights for the corresponding utterances.

Table 3: *The test samples from the SwDA corpus where the Utt-Att-BiRNN model correctly predict as opposed to the simple utterance-level classifier.*

GT	NC	WC	Num	pct.
ny	*b*	*ny*	33	0.79
aa	*b*	*aa*	29	0.69
aa	*sd*	*aa*	12	0.28
b	*aa*	*b*	23	0.55
b	*%*	*b*	16	0.38

5. Conclusions and future research

In this article, we have presented the Utt-Att-BiRNN model for conversational analysis. We demonstrated that our model allows not only to model context-based pragmatic learning but also to compute the amount of information used from the context. Our model achieves a state-of-the-art result on the SwDA corpus of about 77% of accuracy, using only preceding utterances in the context. We showed that our model correctly predicted a significant number of the instances on a DA recognition task. We also show that the context-based learning approach shows higher confidence on the classification task compared to simple utterance-level classification. We have investigated different aspects of the conversational analysis and tested on an important task: dialogue act recognition.

In this research, we only analyzed the utterance representations based on transcripts. However, we plan to use audio features in addition which could provide better representations. Furthermore, it would also help to analyze and mitigate the influence of transcription errors. We investigated the DA annotations by reviewing the predictions of different models, but we could extend it to find out a reliable metric to assess the model performance.

6. Acknowledgements

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

7. References

[1] J. L. Austin, *How to Do Things with Words*. Oxford University Press, 1962.

[2] M. Sbisà, "Speech acts in context," *Language & Communication*, vol. 22, no. 4, pp. 421–436, 2002.

[3] J. R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, 1979.

[4] S. Wermter and M. Löchel, "Learning dialog act processing," in *Proc. of the 16th Conference on Computational Linguistics*, vol. 2. Association for Computational Linguistics, 1996, pp. 740–745.

[5] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 273–278.

[6] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proc. of the International Conference on Learning Representations*, 2015.

[8] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based Bidirectional Long Short-term Memory Networks for Relation Classification," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2016, pp. 207–212.

[9] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1992, pp. 517–520.

[10] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[11] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue act classification using a Bayesian approach," in *Proc. of the 9th Conference Speech and Computer (SPECOM)*, 2004, pp. 495–499.

[12] M. Tavafi, Y. Mehdad, S. R. Joty, G. Carenini, and R. T. Ng, "Dialogue Act Recognition in Synchronous and Asynchronous Conversations," in *Proc. of the Conference of the Special Interest Group on Discourse and Dialogue*. ACL, 2013, pp. 117–121.

[13] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network," in *Proc. of the International Conference on Computational Linguistics*, 2016, pp. 2012–2021.

[14] V. K. R. Sridhar, S. Narayanan, and S. Bangalore, "Modeling the Intonation of Discourse Segments for Improved Online Dialog Act Tagging," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 5033–5036.

[15] N. Kalchbrenner and P. Blunsom, "Recurrent Convolutional Neural Networks for Discourse Compositionality," in *Proc. of the Workshop on Continuous Vector Space Models and their Compositionality, ACL*, 2013, pp. 119–126.

[16] Y. Ji, G. Haffari, and J. Eisenstein, "A Latent Variable Recurrent Neural Network for Discourse Relation Language Models," in *Proc. of the Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 332–342.

[17] H. Kumar, A. Agarwal, R. Dasgupta, S. Joshi, and A. Kumar, "Dialogue Act Sequence Labeling using Hierarchical encoder with CRF," *arXiv:1709.04250v2*, 2017.

[18] Q. H. Tran, I. Zukerman, and G. Haffari, "Preserving Distributional Information in Dialogue Act Classification," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2017, pp. 2141–2146.

[19] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using Context Information for Dialog Act Classification in DNN Framework," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2017, pp. 2160–2168.

[20] D. Ortega and N. T. Vu, "Neural-based Context Representation Learning for Dialog Act Classification," in *Proc. of the Conference of the Special Interest Group on Discourse and Dialogue*, 2017, pp. 247–252.

[21] Z. Meng, L. Mou, and Z. Jin, "Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection," in *Proc. of the ACM Conference on Information and Knowledge Management*, 2017, pp. 2203–2206.

[22] J. Y. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," *arXiv:1603.03827*, 2016.

[23] C. Bothe, C. Weber, S. Magg, and S. Wermter, "A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks," in *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ERLA), 2018, pp. 1952–1957.

[24] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A Survey of Available Corpora for Building Data-Driven Dialogue Systems," *arXiv:1512.05742*, 2015.

[25] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard Dialog Act Corpus," International Computer Science Inst. Berkeley CA, Tech. Rep., 1997.

[26] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to Generate Reviews and Discovering Sentiment," *arXiv: 1704.01444*, 2017.

[27] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative LSTM for sequence modelling," *Workshop track of Proc. of the International Conference on Learning Representations*, 2016.

[28] E. Lakomkin, C. Bothe, and S. Wermter, "GradAscent at EmoInt-2017: Character and Word Level Recurrent Neural Network Models for Tweet Emotion Intensity Detection," in *Proc. of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis at the Conference EMNLP*. ACL, 2017, pp. 169–174.

[29] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4444–4451.

[30] J. L. Elman, "Finding Structure in Time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[31] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. of the Conference on Advances in Neural Information Processing Systems*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1243–1251.

[32] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning Where to Attend with Deep Architectures for Image Tracking," *Neural Computation*, vol. 24, no. 8, pp. 2151–2184, 2012.

[33] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a Foreign Language," in *Proc. of the Conference on Advances in Neural Information Processing Systems*, 2015, pp. 2773–2781.

[34] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in *Proc. of the Conference on Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[35] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving Neural Networks by Preventing Co-adaptation of Feature Detectors," *arXiv:1207.0580*, 2012.

[36] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of the 3rd International Conference on Learning Representations*, 2014.