Dialogue Act Classification with Context-Aware Self-Attention

Vipul Raheja **Joel Tetreault**
Grammarly
firstname.lastname@grammarly.com

Abstract

Recent work in Dialogue Act classification has treated the task as a sequence labeling problem using hierarchical deep neural networks. We build on this prior work by leveraging the effectiveness of a context-aware self-attention mechanism coupled with a hierarchical recurrent neural network. We conduct extensive evaluations on standard Dialogue Act classification datasets and show significant improvement over state-of-the-art results on the Switchboard Dialogue Act (SwDA) Corpus. We also investigate the impact of different utterance-level representation learning methods and show that our method is effective at capturing utterance-level semantic text representations while maintaining high accuracy.

1 Introduction

Dialogue Acts (DAs) are the functions of utterances in dialogue-based interaction (Austin, 1975). A DA represents the meaning of an utterance at the level of illocutionary force, and hence, constitutes the basic unit of linguistic communication (Searle, 1969). DA classification is an important task in Natural Language Understanding, with applications in question answering, conversational agents, speech recognition, etc. Examples of DAs can be found in Table 1. Here we have a conversation of 7 utterances between two speakers. Each utterance has a corresponding label such as *Question* or *Backchannel*.

Speaker	Utterance	DA label
A	Okay.	Other
A	Um, what did you do this weekend?	Question
B	Well, uh, pretty much spent most of my time in the yard.	Statement
B	[Throat Clearing]	Non Verbal
A	Uh-Huh.	Backchannel
A	What do you have planned for your yard?	Question
B	Well, we're in the process of, revitalizing it.	Statement

Table 1: A snippet of a conversation sample from the SwDA Corpus. Each utterance has a corresponding dialogue act label.

Early work in this field made use of statistical machine learning methods and approached the task as either a structured prediction or text classification problem (Stolcke et al., 2000; Ang et al., 2005; Zimmermann, 2009; Surendran and Levow, 2006). Many recent studies have proposed deep learning models for the DA classification task with promising results (Lee and Dernoncourt, 2016; Khanpour et al., 2016; Ortega and Vu, 2017). However, most of these approaches treat the task as a text classification problem, treating each utterance in isolation, rendering them unable to leverage the conversation-level contextual dependence among utterances. Knowing the text and/or the DA labels of the previous utterances can assist in predicting the current DA state. For instance, in Table 1, the *Answer* or *Statement* dialog acts often follow *Question* type utterances.

This work draws from recent advances in NLP such as self-attention, hierarchical deep learning models, and contextual dependencies to produce a dialogue act classification model that is effective across multiple domains. Specifically, we propose a hierarchical deep neural network to model different levels of utterance and dialogue act semantics, achieving state-of-the-art performance on the Switchboard Dialogue Act Corpus. We demonstrate how performance can improve by leveraging context at different levels of the model: previous labels for sequence prediction (using a CRF), conversation-level context with self-attention for utterance representation learning, and character embeddings at the word-level. Finally, we explore different ways to learn effective utterance repre-

Proceedings of NAACL-HLT 2019, pages 3727–3733
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

sentations, which serve as the building blocks of our hierarchical architecture for DA classification.

2 Related Work

A full review of all DA classification methods is outside the scope of the paper, thus we focus on two main classes of approaches which have dominated recent research: those that treat DA classification as a text classification problem, where each utterance is classified in isolation, and those that treat it as a sequence labeling problem.

Text Classification: Lee and Dernoncourt (2016) build a vector representation for each utterance, using either a CNN or RNN, and use the preceding utterance(s) as context to classify it. Their model was extended by Khanpour et al. (2016) and Ortega and Vu (2017). Shen and Lee (2016) used a variant of the attention-based encoder for the task. Ji et al. (2016) use a hybrid architecture, combining an RNN language model with a latent variable model.

Sequence Labeling: Kalchbrenner and Blunsom (2013) used a mixture of sentence-level CNNs and discourse-level RNNS to achieve state-of-the-art results on the task. Recent works (Li and Wu, 2016; Liu et al., 2017) have increasingly employed hierarchical architectures to learn and model multiple levels of utterance and DA dependencies. Kumar et al. (2018), Chen et al. (2018) and Tran et al. (2017) used RNN-based hierarchical neural networks, using different combinations of techniques like last-pooling or attention mechanism to encode sentences, coupled with CRF decoders. Chen et al. (2018) achieved the highest performance to date on the two datasets for this task.

Our work extends these hierarchical models and leverages a combination of techniques proposed across these prior works (CRF decoding, contextual attention, and character-level word embeddings) with self-attentive representation learning, and is able to achieve state-of-the-art performance.

3 Model

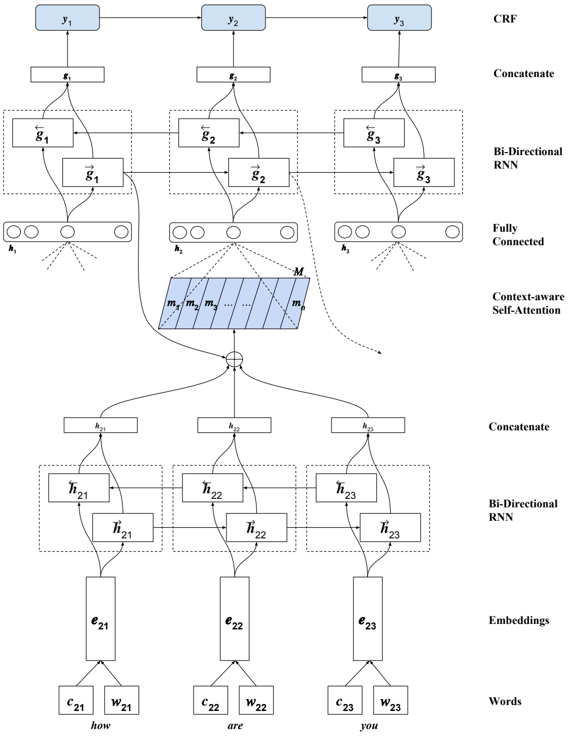

Figure 1: Model Architecture

The task of DA classification takes a conversation C as input, which is a varying length sequence of utterances $U = \{u_1, u_2, ...u_L\}$. Each utterance $u_i \in U$, in turn, is a sequence of varying lengths of words $\{w_i^1, w_i^2, ..., w_i^{N_i}\}$, and has a corresponding target label $y_i \in Y$. Hence, each conversation (i.e. a sequence of utterances) is mapped to a corresponding sequence of target labels $Y = \{y_1, y_2, ..., y_L\}$, which represents the DAs associated with the corresponding utterances.

Figure 1 shows the overall architecture of our model, which involves three main components: (1) an utterance-level RNN that encodes the information within the utterances at the word and character-level; (2) a context-aware self-attention mechanism that aggregates word representations into utterance representations; and (3) a conversation-level RNN that operates on the utterance encoding output of the attention mechanism, followed by a CRF layer to predict utterance labels. We describe them in detail below.

3.1 Utterance-level RNN

For each word in an utterance, we combine two different word embeddings: GloVe (Pennington et al., 2014) and pre-trained ELMo representations (Peters et al., 2018) with fine-tuned task-specific parameters, which have shown superior performance in a wide range of tasks. The word embedding is then concatenated with its CNN-based 50-D character-level embedding (Chiu and Nichols, 2016; Ma and Hovy, 2016) to get the complete word-level representations. The motivation behind incorporating subword-level information is to infer the lexical features of utterances and named entities better.

The word representation layer is followed by a bidirectional GRU (Bi-GRU) layer. Concatenating the forward and backward outputs of the Bi-GRU generates the utterance embedding that serves as input to the utterance-level context-aware self-attention mechanism which learns the final utterance representation.

3.2 Context-aware Self-attention

Self-attentive representations encode a variable-length sequence into a fixed size, using an attention mechanism that considers different positions within the sequence. Inspired by Tran et al. (2017), we use the previous hidden state from the conversation-level RNN (Section 3.3), which provides the context of the conversation so far, and combine it with the hidden states of all the constituent words in an utterance, into a self-attentive encoder (Lin et al., 2017), which computes a $2D$ representation of each input utterance. We follow the notation originally presented in Lin et al. (2017) to explain our modification of their self-attentive sentence representation below.

An utterance u_i, which is a sequence of n words $\{w_i^1, w_i^2, ...w_i^n\}$, is mapped into an embedding layer, resulting in a d-dimensional word embedding for every word. It is then fed into a bidirectional-GRU layer, whose hidden state outputs are concatenated at every time step.

$$\overrightarrow{h_i^j} = \overrightarrow{GRU}(w_i^j, \overrightarrow{h_i^{j-1}}) \tag{1}$$

$$\overleftarrow{h_i^j} = \overleftarrow{GRU}(w_i^j, \overleftarrow{h_i^{j+1}}) \tag{2}$$

$$\mathbf{h_i^j} = concat(\overrightarrow{h_i^j}, \overleftarrow{h_i^j}) \tag{3}$$

$$H_i = \{\mathbf{h_i^1}, \mathbf{h_i^2}, ...\mathbf{h_i^n}\} \tag{4}$$

H_i represents the n GRU outputs of size $2u$ (u is the number of hidden units in a unidirectional GRU).

The contextual self-attention scores are then computed as follows:

$$S_i = W_{s2} tanh(W_{s1} H_i^T + W_{s3} \overrightarrow{g_{i-1}} + \mathbf{b}) \tag{5}$$

Here, W_{s1} is a weight matrix with a shape of $d_a \times 2u$, W_{s2} is a matrix of parameters of shape $r \times d_a$, where r and d_a are hyperparameters we can set arbitrarily, and W_{s3} is a parameter matrix of shape $d_a \times k$ for the conversational context, where k is another hyperparameter that is the size of a hidden state in the conversation-level RNN (size of $\overrightarrow{g_{i-1}}$), and $\mathbf{b}$ is a vector representing bias. Equation 5 can then be treated as a 2-layer MLP with bias, with d_a hidden units, W_{s1}, W_{s2} and W_{s3} as weight parameters. The scores S_i are mapped into a probability matrix A_i by means of a softmax function:

$$A_i = softmax(S_i) \tag{6}$$

which is then used to obtain a 2-d representation M_i of the input utterance, using the GRU hidden states H_i according to the attention weights provided by A_i as follows:

$$M_i = A_i H_i \tag{7}$$

This 2-d representation is then projected to a 1-d embedding (denoted as $\mathbf{h_i}$), using a fully-connected layer. The conversation-level GRU then operates over this 1-d utterance embedding, and hence, we can represent $\mathbf{g_i}$ as:

$$\overrightarrow{g_i} = \overrightarrow{GRU}(\mathbf{h_i}, \overrightarrow{g_{i-1}}) \tag{8}$$

$$\overleftarrow{g_i} = \overrightarrow{GRU}(\mathbf{h_i}, \overrightarrow{g_{i+1}}) \tag{9}$$

$$\mathbf{g_i} = concat(\overrightarrow{g_i}, \overleftarrow{g_i}) \tag{10}$$

$\mathbf{g_i}$ then provides the conversation-level context used to learn the attention scores and 2-d representation (M_{i+1}) for the next utterance in the conversation ($\mathbf{h_{i+1}}$).

3.3 Conversation-level RNN

The utterance representation $\mathbf{h_i}$ from the previous step is passed on to the conversation-level RNN, which is another bidirectional GRU layer used to encode utterances across a conversation. The hidden states $\overrightarrow{g_i}$ and $\overleftarrow{g_i}$ (Figure 1) are then concatenated to get the final representation $\mathbf{g_i}$ of each utterance, which is further propagated to a linear chain CRF layer. The CRF layer considers the correlations between labels in context and jointly decodes the optimal sequence of utterance labels for a given conversation, instead of decoding each label independently.

4 Data

We evaluate the classification accuracy of our model on the two standard datasets used for DA classification: the Switchboard Dialogue Act Corpus (SwDA) (Jurafsky et al., 1997) consisting of 43 classes, and the 5-class version of the ICSI Meeting Recorder Dialogue Act Corpus (MRDA) introduced in (Ang et al., 2005). For both datasets,

Dataset	|C|	|V|	Train	Validation	Test
MRDA	5	12k	78k	16k	15k
SwDA	43	20k	193k	23k	5k

Table 2: Number of utterances by dataset. |C| denotes number of classes and |V| is the vocabulary size.

we use the train, validation and test splits as defined in Lee and Dernoncourt (2016).

Table 2 shows the statistics for both datasets. They are highly imbalanced in terms of class distribution, with the DA classes `Statement-non-opinion` and `Acknowledge/Backchannel` in SwDA and `Statement` in MRDA making up over 50% of the labels in each set.

5 Results

5.1 Dialogue Act Classification

We compare the classification accuracy of our model against several other recent methods (Table 3).[1] Four approaches (Chen et al., 2018; Tran et al., 2017; Ortega and Vu, 2017; Shen and Lee, 2016) use attention in some form to model the conversations, but none of them have explored self-attention for the task. The last three use CRFs in the final layer of sequence labeling. Only one other method (Chen et al., 2018) uses character-level word embeddings. All models and their variants were trained ten times and we report the average test performance. Our model outperforms state-of-the-art methods by 1.6% on SwDA, the primary dataset for this task, and comes within 0.6% on MRDA. It also beats a TF-IDF GloVe baseline (described in Section 5.2) by 16.4% and 12.2%, respectively.

The improvements that the model is able to make over the other methods are significant, however, the gains on MRDA still fall short of the state-of-the-art by 0.6%. This can mostly be attributed to the conversation/context lengths and label noise at the conversation level. Conversations in MRDA (1493 utterances on average) are significantly longer than in SwDA (271 utterances on average). In spite of having nearly 12% the number of labels (5 vs 43) compared to SwDA, MRDA has 6 times the normalized label entropy in its data. Consequently, due to the noise in label dependencies, and hence, in the inherent conversational structure, the model is not able to yield as big of a gain on the MRDA as it does on the SwDA. Consequently, learning long-range dependencies is a challenge because of noisier and longer path lengths in the network. This is illustrated in Figures 2 and 3, which show for every class, the variation between the entropy of the previous label in a conversation, and the accuracy of that class. MRDA was found to have a high negative correlation[2] (-0.68) between previous label entropy and accuracy, indicating the impact of label noise, which was compounded by longer conversations. On the other hand, SwDA was found to have a low positive correlation (+0.22), which could be compensated by significantly shorter conversations.

Model	SwDA	MRDA
TF-IDF GloVe	66.5	78.7
Kalchbrenner and Blunsom (2013)	73.9	-
Lee and Dernoncourt (2016)	73.9	84.6
Khanpour et al. (2016)	75.8	86.8
Ji et al. (2016)	77.0	-
Shen and Lee (2016)	72.6	-
Li and Wu (2016)	79.4	-
Ortega and Vu (2017)	73.8	84.3
Tran et al. (2017)	74.5	-
Kumar et al. (2018)	79.2	90.9
Chen et al. (2018)	81.3	**91.7**
Our Method	**82.9**	91.1
Human Agreement	84.0	-

Table 3: DA Classification Accuracy

5.2 Utterance Representation Learning

One of the primary motivations for this work was to investigate whether one can improve performance by learning better representations for utterances. To address this, we retrained our model by replacing the utterance representation learning (utterance-level RNN + context-aware self-attention) component with various sentence representation learning methods (either pre-training them or learning jointly), and feeding them into the conversation-level recurrent layers in the hierarchical model, so that the performance is indicative of the quality of utterance representations.

There are three main categories of utterance representation learning approaches: (i) the baseline which uses a TF-IDF weighted sum of GloVe word embeddings; (ii) pre-trained on cor-

[1] Contemporaneous to this submission, (Li et al., 2018; Wan et al., 2018; Ravi and Kozareva, 2018) proposed different approaches for the task. We do not focus on them here per NAACL 2019 guidelines, however note that our system outperforms the first two. (Ravi and Kozareva, 2018) bypasses the need for complex networks with huge parameters but its overall accuracy is 4.2% behind our system, despite being 0.2% higher on SwDA.

[2] Pearson's r

Method	SwDA	MRDA
Baseline		
TF-IDF GloVe	66.5	78.7
Pre-trained on Corpus		
Skip Thought Vectors	72.6	82.8
Paragraph vectors	72.5	82.6
Joint Learning		
RNN-Encoder	74.8	85.7
Bi-RNN-LastState	76.2	85.4
Bi-RNN-MaxPool	77.6	86.7
CNN	76.9	84.5
Bi-RNN + Attention	80.1	87.7
+ Context	81.8	89.2
Bi-RNN + Self-attention	81.1	88.6
+ Context	82.9	91.1

Table 4: Performance of utterance representation methods when integrated with the hierarchical model

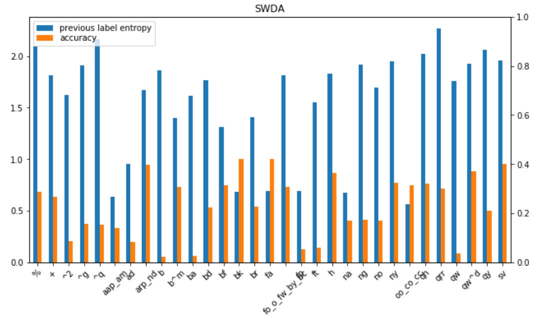

Figure 2: Previous Label Entropy vs. Accuracy on the SwDA Dataset

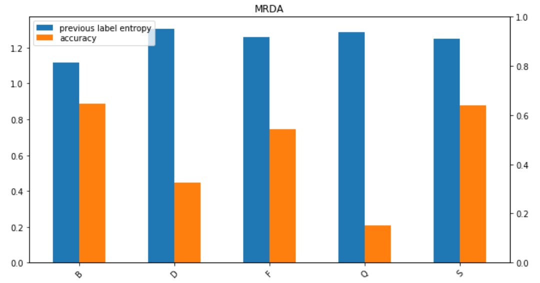

Figure 3: Previous Label Entropy vs. Accuracy on the MRDA Dataset

pus, where we first learn utterance representations on the corpus using Skip-Thought Vectors (Kiros et al., 2015) and Paragraph Vectors (Le and Mikolov, 2014), and then use them with the rest of the model; (iii) jointly learned with the DA classification task. Table 4 describes the performance of different utterance representation learning methods when combined with the overall architecture on both datasets.

Introducing the word-level attention mechanism (Yang et al., 2016) enables the model to learn better representations by attending to more informative words in an utterance, resulting in better performance (Bi-RNN + Attention). The self-attention mechanism (Bi-RNN + Self-attention) leads to even greater overall improvements. Adding context information (previous recurrent state of the conversation) boosts performance significantly.

A notable aspect of our model is how contextual information is leveraged at different levels of the sequence modeling task. The combination of conversation-level contextual states for utterance-representation learning (+ Context) and a CRF at the conversation level to further inform conversation sequence modeling, leads to a collective performance improvement. This is particularly pronounced on the SwDA dataset: the two variants of the context-aware attention models (Bi-RNN + Attention + Context and Bi-RNN + Self-attention + Context) have significant performance gains.

6 Conclusion

We developed a model for DA classification with context-aware self-attention, which significantly outperforms earlier models on the commonly-used SwDA dataset and is very close to state-of-the-art on MRDA. We experimented with different utterance representation learning methods and showed that utterance representations learned at the lower levels can impact the classification performance at the higher level. Employing self-attention, which has not previously been applied to this task, enables the model to learn richer, more effective utterance representations for the task.

As future work, we would like to experiment with other attention mechanisms such as multi-head attention (Vaswani et al., 2017), directional self-attention (Shen et al., 2018a), block self-attention (Shen et al., 2018b), or hierarchical attention (Yang et al., 2016), since they have been shown to address the limitations of vanilla attention and self-attention by either incorporating information from different representation subspaces at different positions to capture both local and long-range context dependencies, encoding temporal order information, or by attending to context dependencies at different levels of granularity.

Acknowledgements

The authors would like to thank Dimitris Alikaniotis, Maria Nadejde and Courtney Napoles for their insightful discussions, and the anonymous reviewers for their helpful comments.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05*, pages 1061–1064.

John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 225–234. ACM.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342. Association for Computational Linguistics.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado at Boulder Technical Report 97-02.

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*, pages 119–126. Association for Computational Linguistics.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021. The COLING 2016 Organizing Committee.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 3440–3447.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pages 1188–1196.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520. Association for Computational Linguistics.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2018. A dual-attention hierarchical recurrent neural network for dialogue act classification. *CoRR*, abs/1810.09154.

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1970–1979. The COLING 2016 Organizing Committee.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *International Conference on Learning Representations 2017 (Conference Track)*.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 247–252. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 2227–2237. Association for Computational Linguistics.

Sujith Ravi and Zornitsa Kozareva. 2018. Self-governing neural networks for on-device short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 887–893. Association for Computational Linguistics.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

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. In *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association*, pages 2716–2720.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018a. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 5446–5455.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. 2018b. Bi-directional block self-attention for fast and memory-efficient sequence modeling. In *International Conference on Learning Representations*.

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

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing*.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yao Wan, Wenqiang Yan, Jianwei Gao, Zhou Zhao, Jian Wu, and Philip S. Yu. 2018. Improved dynamic memory network for dialogue act classification with adversarial training. *CoRR*, abs/1811.05021.

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. Association for Computational Linguistics.

Matthias Zimmermann. 2009. Joint segmentation and classification of dialog acts using conditional random fields. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, pages 864–867.

A Supplementary Material

A.1 Training & Hyperparameters

All hyperparameters were selected by tuning one hyperparameter at a time while keeping the others fixed. Validation splits were used for the tuning process. The final set of hyperparameters were then used to train two different models, one each on SwDA and MRDA training splits. Table 5 lists the range of values for each parameter that we experimented with, and the final value that was chosen. Dropout was applied to the utterance embeddings. Early stopping was used on the validation set with a patience of 15 epochs.

Hyperparams	**Range of values**	**Final value**
Word Embeddings	GloVe 100D GloVe 200D GloVe 300D	GloVe 300D + ELMo 1024D
	Word2vec 300D Word2vec 200D	
	ELMo 1024D	
	GloVe 300D + ELMo 1024D	
	Word2Vec 300D + ELMo 1024D	
Sentence GRU Size (u)	20 - 300	50
Utterance GRU Size (k)	20 - 600	100
Learning Rate	0.01 - 2.0	0.015
Dropout	0.1 - 0.8	0.3
Optimizer	SGD, RMSProp, Adam	Adam

Table 5: Hyperparameter space and tuned values