Contextual Dialogue Act Classification with Long Short-Term Memory Networks

Noah Pflaum and Khyati Tripathi
University of California at Berkeley
npflaum@berkeley.edu and kntripathi@berkeley.edu

Abstract

Dialog Act Recognition (DAR) facilitates implementation of dialog systems. This paper utilizes context-based DAR with neural network models and operates on Switchboard Dialog Act (SwDA) Corpus datasets. Several neural network models are found to provide comparable performance for DAR. The Dialog Act is found to be conversation specific with short-term context sensitivities. Embed-dings generated by the Universal Sentence Encoder (USE) serve as the base embeddings upon which a novel model is constructed that achieves near state-of-the art performance. The proposed novel framework of exploiting changes in the speaker and previous Dialog Acts help achieve the classification accuracy of 77.5%.

1 Introduction

This research performs Dialog Act Recognition (DAR) as the classification task using supervised learning. DAR classifies a speaker's utterance into a Dialog Act such as "Question," "Answer," and "Opinion." Classification of dialogue acts can aid in natural language understanding and interpretation of a conversation, as well as identification of previous dialogue acts often inform the future acts. For example, it is often the case that an utterance with a tag of Yes-No-Question is followed by either No Answers or Yes Answers. However, that is not always the case as a question or other dialogue act may follow a question instead highlighting the contextual complexities of such a problem (Grosz, 1982). Consider Table 1 below.

Dialog Act	ID	Utterance
Yes-No-Question	Utt1	A: You never think about that do you?
Yes-Answer	Utt2	B: Yeah.
Statement-opinion	Utt3	A: I would think it would be harder to get up than it would be.
Backchannel	Utt4	B: Yeah.

Table 1: Dialog Act Examples.

Table 1 below shows the contextual complexity of discourse where the same utterance "Yeah" is classified differently based on its context. In case of the utterance ID Utt2, the utterance "Yeah" is classified as the Dialog Act of "Yes-Answer," because this utterance represents an answer to the question posed in the previous utterance. In contrast, the same utterance "Yeah" in Utt4 is classified as the Dialog Act of "Backchannel" because it is a response or affirmation to the previous utterance, which was "Statement-opinion." The example in Table 1 shows that it is quite challenging to accurately identify the speaker's intent.

DAR has various use cases in dialogue systems ranging from virtual personal assistants, chatbots and machine translation (Král and Cerisara, 2010). DAR can also be used to help summarize the type of conversation that has taken place as well as search use cases when trying to find utterances of specific DA types.

DAR can be viewed as a dialog interpretation problem, where semantic labels are attached to utterances and the speaker's intention is characterized. This research utilizes Switchboard Dialog Act (SwDA) Corpus, which is a collection consisting of about 2,400 two-way telephone

conversations among 543 speakers covering 70 conversational topics. There are 43 classifications of Dialog Acts. **The goal of this research is to produce a model that could be deployed on live dialog systems**.

Section 2 carries out a concise literature survey and summarizes the status of research on DAR. Section 3 describes the overall design procedure and the proposed novel DAR framework. Section 4 illustrates the results of the analysis. Finally, Section 5 summarizes our findings and outlines directions for further research.

2 Background

DAR is a well-studied dialog task with research regarding Dialog Act schemes (Austin, 1962) being developed as early as the 1960s and later refined to the Dialogue Act Markup in Several Layers (Allen and Core, 1997) - the scheme used for the SwDA dataset this paper explores. Moreover, classification models have been widely explored starting with statistical, non-contextual models (Reithinger et al., 1997). Beyond that, Hidden Markov Models (HMM) with various language features (Stolcke et al., 2000) provided the state-of-the-art performance at around 71% accuracy. However, as noted by Chen et al. (2017), most of these models use feature engineering in a non-scalable manner relying on information specific to the dataset of concern. These models provided their best performance using an SVM-HMM hybrid model (Tavafi et al. 2013) with a peak accuracy of 74.32%.

With the emergence of deep learning, various neural network-based approaches have been explored. Lee and Dernoncourt proposed a novel approach using recurrent neural networks (RNNs) and convolutional neural networks (CNNs) with the GloVe pretrained embeddings (Pennington et al., 2014) to achieve accuracy of 71.4%. Since then, most state-of-the-art approaches have used Long Short-Term Memory (LSTM) networks with Conditional Random Fields (CRFs) in some form to achieve state of the art results. This is seen with Kumar et al. (2017) achieving an accuracy of 79%, followed by Chen et al. (2017) achieving an accuracy of 81.3% by extending this type of architecture with an Attentive Structured Network (ASN). These models, however, are forward looking, classifying the utterances at the conversation level. This obviously limits their use in live dialog systems. When considering only backward-looking models, the current state-of-the-art stands at an accuracy of 77.3% as reported by Bothe et al. (2018).

3 Methods

3.1 Design Methodology: A Brief Overview

Figure 1 depicts the overall methodology pursued in this research.

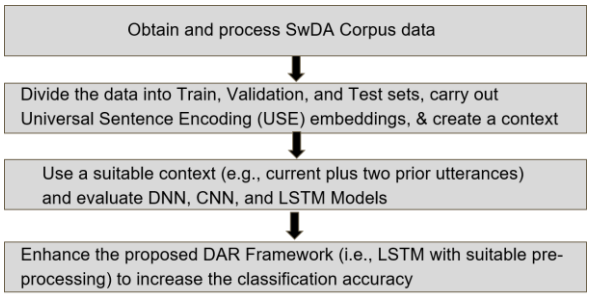

Figure 1: Overall DAR Methodology.

The SwDA dataset was obtained from this location: https://github.com/cgpotts/swda.git .

The dataset was divided into 1003 Train conversations, 112 Validation conversations, and 19 Test conversations.

The Universal Sentence Encoding (USE) was used to represent each utterance (i.e., a sentence) of a conversation by a 512-element vector. A context was created by using the current utterance along with a certain number of prior utterances (e.g., two prior utterances).

To choose a specific neural network model for a detailed exploratory analysis, three models were evaluated- a Deep Neural Network (DNN) model, a Convolutional Neural Network (CNN) model, and Long Short-Term Memory (LSTM) model. The findings for the DNN model and the CNN model are summarized in Section 3.2, while the detailed analysis for the LSTM model is discussed in Section 4.

Based on the performance of the candidate models- DNN, CNN, and LSTM, the LSTM model was chosen for a more detailed analysis and experimentation of innovations such as the speaker change and the previous Dialog Act.

3.2 Comparison of Candidate Neural Network Models

The DNN model was evaluated with varying numbers of training epochs (e.g., 5 to 30) and the number of hidden layer neurons (e.g., 150, 250, and 300). Figure 2 displays the trends in the Train and Validation accuracies as a function of epochs for the case of 250 neurons. When the Validation set accuracy was the highest, the Training, Validation, and Test accuracies of 76.2%, 73.4%, and 72.0%, were achieved, respectively.

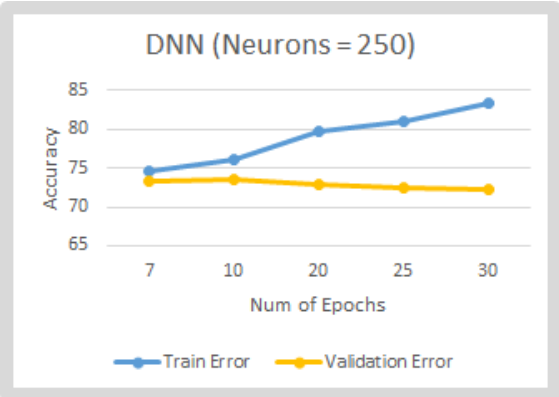

Figure 2: DNN Model Performance.

Figure 3 shows the trends in the Train and Validation accuracies as a function of epochs for the filter length of 300 for the CNN model. When the Validation set accuracy was the highest, the Training, Validation, and Test accuracies of 77.1%, 73.5%, and 71.7%, were achieved, respectively.

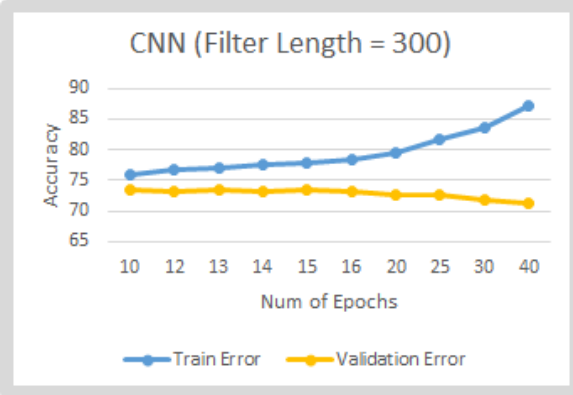

Figure 3: CNN Model Performance.

The comparison of Figure 2 and Figure 3 indicates that the DNN and the CNN provide comparable performance. The LSTM model provides a slightly better Test accuracy of 72.3% and was chosen for further analysis and refinements.

3.3 Proposed DAR Framework with the Selected LSTM Model

The proposed methodology only considers the utterances, speaker and historical dialogue acts of the dataset as it was determined that these would be accessible in a live system. The model can be split into two main parts: representation of the input data and the model architecture for classification.

For the representation of the input data, a composite embedding of the utterance, speaker and dialogue act were utilized. For the dialogue act and speaker, a one-hot encoding was used. For the utterances' text, the Universal Sentence Encoder (USE) (Cer et al., 2018) was used to provide sentence level embeddings as a vector with length 512. A decision to use the USE-based embeddings was made because it has been pretrained on various classification tasks. The USE transformer-based model was used due to its better performance than the deep averaging network variant on various tasks. The embedding for each timestep can be represented as follows:

$$e_t = \text{concat}(da_{t-1}, s_t, \text{USE}(u_t)),$$

where USE is the USE embedding function, ut is the current utterance, s_t is the current speaker and da_{t-1} is the previous timestep's Dialog Act. Each conversation is padded with a known token to keep a fixed context length for the LSTM.

The model architecture is a unidirectional single layer LSTM with a fully connected layer and a softmax for classification as shown in Figure 4. The architecture displayed is for a single sample where a number of embeddings in context are supplied as input to the LSTM, and the final cell state is used by the fully connected layer.

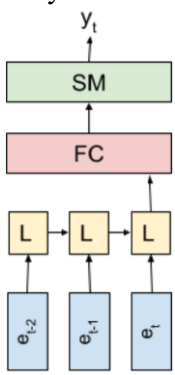

Figure 4: CNN Model Performance.

While not shown in this paper, initial investigations included using multiplicative LSTMs and Bidirectional LSTMs. However, since no benefit was observed, this simpler architecture was chosen in the interest of parsimony.

3.3 Model Training with Google Colab

For training, Google Colab was used with a GPU for the runtime with 26 GB of RAM on the CPU and 16 GB of RAM on the GPU. This configuration was sufficient to hold the data in memory allowing for quick training and testing with a typical epoch taking tens of seconds for the entire dataset.

4 Results and Discussion

To evaluate the models, hyperparameters were tested at a range of values varying one hyper parameter at a time to find appropriate values. Table 2 below represents the values explored and the final values used.

Hyperparameter	Ideal	Explored
LSTM Output Dim.	256	32 - 2048
Dropout rate	0.2	0.05 - 0.4
Fully Connected Layer Dim.	512	64 - 2048
Utterances in context	3	2, 3, 5, 10
Optimizer	Nesterov Adam	Adagrad, Adam and Nesterov Adam

Table 2: LSTM Hyperparameters.

As noted, USE was employed for the embedding layer resulting in each utterance being represented by a vector of 512 dimensions. It was determined that USE provided a reasonable language model given that simply adding a softmax classification layer on top of it provided 68% accuracy - see table 3 below. This served as a baseline for performance to measure against although choosing the most common class and a simple DNN and CNN are also provided as reference.

Model	Accuracy
Most common class	31.5%
Embedding + Softmax	68.2%
Deep Neural Network (3 utt. in context)	72.0%
CNN (3 utt. in context)	71.7%
LSTM (2 utt. in context)	71.9%
LSTM (3 utt. in context)	72.3%
LSTM (5 utt. in context)	72.2%
LSTM (2 utt. + speaker)	73.2%
LSTM (3 utt. + speaker)	74.8%
LSTM (5 utt. + speaker)	74.4%
LSTM (2 utt. + speaker + DA)	75.8%
LSTM (3 utt. + speaker + DA)	77.5%
LSTM (5 utt. + speaker + DA)	77.5%

Table 3: Summary of Performance Results.

The first iteration considered the effects of the number of utterances in context with exploration of 2, 3 and 5 total utterances including the current utterance. In these models it was found that including a greater amount of context benefited the performance of the model. It was also noticed that the benefit of context is fairly short ranged with performance peaking with 3 total utterances and including 5 utterances didn't improve performance. Although not shown, 10 utterances were explored with similar results to 5 utterances.

Beyond the textual utterances, providing the speaker identity as a one hot encoded vector allowed the model to learn when speaker changes occurred. As hypothesized, this improved performance of the system as speaker changes can be informative of utterance intent. Similarly, this configuration provided the best accuracy with 3 utterances in context with an accuracy of 74.8%.

The third model includes providing the previous contexts dialogue act. Again, this improved performance up to 77.5% with 3 and 5 utterances in context. It is perhaps suggestive that dialogue acts follow certain sequences based on the context of the utterances.

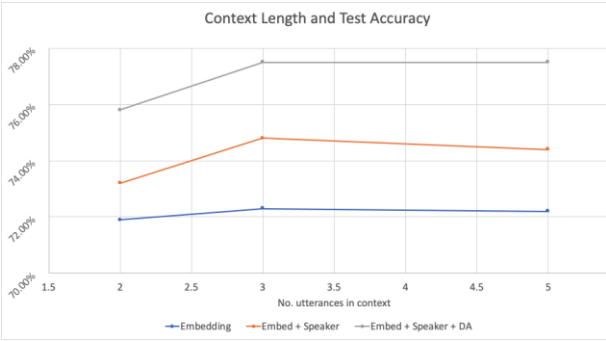

Figure 5: Influence of The Number of Utterances in the Context on the Dialog Act Accuracy.

The results observed are roughly equal to the state-of-the-art results realized by Bothe et al. for backward looking models. Moreover, the human inter-annotated agreement is only 84% suggesting that for certain utterances, the appropriate label isn't completely understood. An example of where the model experienced this was in the simple utterance of "um," in the test set where the model predicted a label of Abandoned/turn-exit while the true label is backchannel. It seems both labels could be appropriate and, as a human reader, it is hard to distinguish between what is correct just from the text and previous context. This highlights another interesting artifact about the dataset in that it is an audio dataset that has been transcribed as text. As noted by Miltenburg et al. (2018), spoken language is different to textual language and varies by context. While not explored for this dataset, it is perhaps suggested that there is information loss (i.e. loss of intonations and cadence) in the transcribing process, which could limit the upper bound of performance.

Further areas the model had issues were with rhetorical phrases such as "You know what I mean?" where a response of "yeah" was labeled a Backchannel while the model predicted Yes Answers. Again, distinguishing what is the correct label is a hazy task that is perhaps even suggestive that there is overlap between the labels.

Finally, the model struggled a little bit with non-language features that appeared in the utterances from limited preprocessing. A parser available online was used and certain disfluency markers present in the text still appeared in the inputs to the model. One such example where the model misclassified the utterance was the utterance "(()) not be…". It may be that the USE embeddings did not represent such an utterance appropriately and perhaps better preprocessing or inclusion of additional domain specific embeddings would help improve performance here.

5 Conclusion

This research paper presents a model for DAR that could be employed in live dialog systems. The research illustrates how DAR benefits from a previous context, highlighting the temporal and contextual nature of dialogs. That contextual importance does, however, seem to be short ranged. Although not numerically visible in this paper, there was some indication in the experiments carried out that with a greater amount of information (e.g., including previous Dialog Acts as an input), a greater amount of context can be beneficial. The relevance of the context is still believed to be fairly short ranged compared to other dialog tasks that include memory recall.

One possible extension of this research is to apply the proposed DAR framework to other datasets and evaluate its performance. A dynamic-length context may also be useful, where the lengths of the context for utterances, speakers and Dialog Acts can be varied independently or in a correlated manner.

Acknowledgments

The authors thank Prof. Mark Butler of University of California at Berkeley for useful suggestions throughout this research project.

References

J. Allen and M. Core. 1997. *Draft of DAMSL: Dialogue Act Markup in Several Layers*.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

C. Bothe, C. Weber, S. Magg, and S. Wermter. 2018. *A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks*. https://arxiv.org/pdf/1805.06280v1.pdf .

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, St. John R., N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil. 2018. Universal Sentence Encoder.

Z. Chen, R. Yang, Z. Zhao, D. Cai, X. He. 2017. *Dialogue Act Recognition via CRF-Attentive Structured Network*. https://arxiv.org/pdf/1711.05568v1.pdf .

B. J. Grosz. 1982. *Discourse Analysis. Sublanguage. Studies of Language in Restricted Semantic Domains*, pages 138–174.

P. Král and C. Cerisara. 2010. *Dialogue Act Recognition Approaches.* Computing and Informatics. 29. 227-250.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. *Dialogue Act Sequence Labeling using Hierarchical encoder with CRF*. https://arxiv.org/pdf/1709.04250v2.pdf

Norbert Reithinger and Martin Klesen. 1997. *Dialogue act classification using language models*. In EuroSpeech.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. *GloVe: global vectors for word representation*. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 12:1532–1543.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech*. Computational Linguistics, 26(3):339–373.

M. Tavafi, Y. Mehdad, S. R. Joty, G. Carenini, and R. T. Ng. 2013. *Dialogue act recognition in synchronous and asynchronous conversations*. In SIGDIAL.