

# Neural-based Context Representation Learning for Dialog Act Classification

Daniel Ortega      Ngoc Thang Vu

Institute for Natural Language Processing (IMS)

University of Stuttgart

Pfaffenwaldring 5b, 70569 Stuttgart, Germany

{daniel.ortega, thang.vu}@ims.uni-stuttgart.de

## Abstract

We explore context representation learning methods in neural-based models for dialog act classification. We propose and compare extensively different methods which combine recurrent neural network architectures and attention mechanisms (AMs) at different context levels. Our experimental results on two benchmark datasets show consistent improvements compared to the models without contextual information and reveal that the most suitable AM in the architecture depends on the nature of the dataset.

## 1 Introduction

The study of spoken dialogs between two or more speakers can be approached by analyzing the dialog acts (DAs), which is the intention of the speaker at every utterance during a conversation. Table 1 shows a fragment of a conversation from the Switchboard (SwDA) dataset with DA annotation. Automatic DA classification is an important pre-processing step in natural language understanding tasks and spoken dialog systems. This classification task has been approached using traditional statistical methods such as hidden Markov models (HMMs) (Stolcke et al., 2000), conditional random fields (CRF) (Zimmermann, 2009) and support vector machines (SVMs) (Henderson et al., 2012). However, recent works with deep learning (DL) techniques have brought state-of-the-art models in DA classification, such as convolutional neural networks (CNNs) (Kalchbrenner and Blunsom, 2013; Lee and Dornoncourt, 2016), recurrent neural networks (RNNs) (Lee and Dornoncourt, 2016; Ji et al., 2016) and long short-term memory (LSTM) models (Shen and Lee, 2016).

Utterance	Dialog act
A: <i>Are you a musician yourself?</i>	Yes-no-question
B: <i>Uh, well, I sing.</i>	Affirmative non-yes answer
A: <i>Uh-huh.</i>	Acknowledge (Backchannel)
B: <i>I don't play an instrument.</i>	Statement-non-opinion

Table 1: Examples from the SwDA dataset.

Given an utterance in a dialog without any previous context, it is not always obvious even for human beings to find the corresponding dialog act. In many cases, the utterances are too short so that is hard to classify them, for example the utterance '*Right*' can be either an *Agreement* or a *Backchannel* indicating the interlocutor to go on talking, in this case the context plays a key role at disambiguating. Therefore, using context information from the previous utterances in a dialog flow is a crucial step for improving DA classification. Few papers in the literature have suggested to utilize context as a potential knowledge source for DA classification (Lee and Dornoncourt, 2016; Shen and Lee, 2016). Recently, Ribeiro et al. (2015) presented an extensive analysis of the influence of context on DA recognition concluding that contextual information from preceding utterances helps to improve the classification performance. Nonetheless, such information should be differentiable from the current utterance information, otherwise, the contextual information could have a negative impact.

Attention mechanisms (AMs) introduced by Bahdanau et al. (2014) have contributed to significant improvements in many natural language processing tasks, for instance machine translation (Bahdanau et al., 2014), sentence classification (Shen and Lee, 2016) and summarization (Rush et al., 2015), uncertainty detection (Adel and Schütze, 2017), speech recognition (Chorowski et al., 2015), sentence pair modeling (Yin et al., 2015), question-answering (Golub and He, 2016),

document classification (Yang et al., 2016) and entailment (Rocktäschel et al., 2015). AMs let the model decide what parts of the input to pay attention to according to the relevance for the task.

In this paper, we explore the use of AMs to learn the context representation, as a manner to differentiate the current utterance from its context as well as a mechanism to highlight the most relevant information, while ignoring unimportant parts for DA classification. We propose and compare extensively different neural-based methods for context representation learning by leveraging a recurrent neural network architecture with LSTM (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRUs) (Cho et al., 2014; Chung et al., 2014) in combination with AMs.

## 2 Model

The model architecture, shown on the left side of Figure 1, contains two main parts: the CNN-based utterance representation and the attention mechanism for context representation learning. Finally, the context representation is fed into a softmax layer which outputs the posterior of each predefined DA given the current dialog utterance.

### 2.1 CNN-based Dialog Utterance Representation

We used CNNs for the representation of each utterance. CNNs perform a discrete convolution on an input matrix with a set of different filters. For the DA classification task, the input matrix represents a dialog utterance and its context, this is  $n$  previous utterances: each column of the matrix stores the word embedding of the corresponding word. We use 2D filters  $f$  (with width  $|f|$ ) spanning all embedding dimensions  $d$ . This is described by the following equation:

$$(w * f)(x, y) = \sum_{i=1}^d \sum_{j=-|f|/2}^{|f|/2} w(i, j) \cdot f(x-i, y-j) \quad (1)$$

After convolution, a max pooling operation is applied that stores only the highest activation of each filter. Furthermore, we apply filters with different window sizes 3-5 (multi-windows), i.e. spanning a different number of input words. Then, all feature maps are concatenated to one vector which represents the current utterance and its context.

### 2.2 Internal Attention Mechanism

Attention mechanisms can be applied in different sequences of input vectors, e.g. representations of consecutive dialog utterances. For each of the input vectors  $u(t-i)$  at time step  $t-i$  in a dialog and  $t$  is the current time step, the attention weights  $\alpha_i$  are computed as follows

$$\alpha_i = \frac{\exp(f(u(t-i)))}{\sum_{0 < j < m} \exp(f(u(t-j)))} \quad (2)$$

where  $f$  is the scoring function. In this work,  $f$  is the linear function of the input  $u(t-i)$

$$f(u(t-i)) = W^T u(t-i) \quad (3)$$

where  $W$  is a trainable parameter. The output *attentive\_u* after the attention layer is the weighted sum of the input sequence.

$$\textit{attentive\_u} = \sum_i \alpha_i u(t-i) \quad (4)$$

Another option (*order-preserved attention* as proposed in Adel and Schütze (2017)) is to store the weighted inputs into a vector sequence *attentive\_v* which preserves the order information.

$$\textit{attentive\_v} = [\alpha_0 u(t), \alpha_1 u(t-1), \dots] \quad (5)$$

### 2.3 Neural-based Context Modeling

In this subsection, we present different methods, depicted on the right side of Figure 1, to learn the context representation.

**(a) Max** We apply max-pooling on top of the dialog utterance representations which spans all the contexts and the vector dimension.

**(b) Attention** We apply directly attention mechanism on the dialog utterance representations. The weighted sum of all the dialog utterances represents the context information.

**(c) RNN** We introduce a recurrent architecture with LSTM or GRU cells on top of the dialog utterance representations to model the relation between the context and the current utterance over time. The output of the hidden layer of the last state is the context representation.

**(d) RNN-Output-Attention** Based on the previous option, we apply the attention mechanisms on the output sequence of the RNN. The context representation is the weighted sum of all the output vectors.

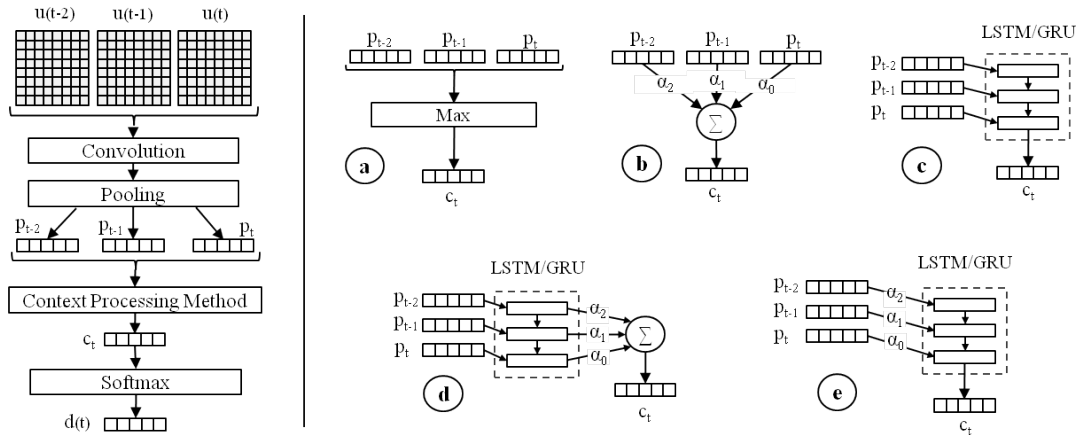


Figure 1: Model architecture for DA classification. On the left side is the overview of the model. The right site contains six neural-based methods for context representation learning.

Ⓔ **RNN-Input-Attention** We first apply the order-preserved attention mechanism on the dialog utterance representations to obtain a sequence of weighted inputs. Afterwards, an RNN with LSTM or GRU cells is introduced to model the relation of the weighted context.

### 3 Experimental Setup

#### 3.1 Data

We test our model on two DA datasets:

- **MRDA**: ICSI Meeting Recorder Dialog Act Corpus (Janin et al., 2003; Shriberg et al., 2004; Dhillon et al., 2004), a dialog corpus of *multiparty meetings*. The 5-tag-set used in this work was introduced by Ang et al. (2005).
- **SwDA**: Switchboard Dialog Act Corpus (Godfrey et al., 1992; Jurafsky et al., 1997), a dialog corpus of *2-speaker conversations*.

Train, validation and test splits on both datasets were taken as defined in Lee and Derroncourt (2016)<sup>1</sup>, summary statistics are shown in Table 2. In both datasets the classes are highly unbalanced, the majority class is 59.1% on MRDA and 33.7% on SwDA.

#### 3.2 Hyperparameters and Training

The hyperparameters for both datasets are summarized in Table 3, they were selected by vary-

<sup>1</sup>Concerning SwDA, the data setup in Lee and Derroncourt (2016) was preferred over Stolcke et al. (2000)’s, because it was not clearly found in the latter which conversations belong to each split.

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

Table 2: Data statistics: **C** is the number of classes, **|V|** is the vocabulary size and **Train/Validation/Test** are the no. of utterances.

ing one hyperparameter at a time while keeping the others fixed. The filter widths and feature maps were taken from the CNN architecture for sentence classification in Kim (2014). Dropout rate of 0.5 was found to be the most effective in the range of [0-0.9]. The rectified linear unit (ReLU) was used as non-linear activation function, 1-max as pooling operation at utterance level as suggested in Zhang and Wallace (2015). The only dataset specific hyperparameter is the mini-batch size: 150 and 50 for SwDA and MRDA, respectively. Word2vec (Mikolov et al., 2013) was used for word vector representation. Training was done for 30 epochs with averaged stochastic gradient descent (Polyak and Juditsky, 1992) over mini-batches. The learning rate was initialized at 0.1 and reduced 10% every 2000 parameter updates. We kept the word vector unchanged during training. The context length was optimized on the development set, ranging from 1-5. Our best results were obtained with three context utterances for MRDA and two for SwDA.

## 4 Experimental Results

### 4.1 Baseline Models

We define two models as baseline, both are a one-layer CNN for sentence classification based on

Hyperparameter	Value
Filter width	3, 4, 5
Feature maps per filter	100
Dropout rate	0.5
Activation function	ReLU
Pooling	1-max pooling per utterance
Mini-batch size	50 (MRDA) – 150 (SwDA)
Word embeddings	word2vec (dim. 300)

Table 3: Hyperparameters.

Kim (2014) but with an input variation: a) Baseline I: The input is a single utterance a time without any contextual information and b) Baseline II: The input is the concatenation of the current utterance and previous utterances.

## 4.2 Results

Table 4 summarizes the results of all the models. Results on the Baseline I and the Baseline II on both datasets show that a simple context concatenation is not enough to model the context information for this task. While on SwDA the accuracy improves by 1.3%, it slightly drops on MRDA. Other simple methods such as *Max* and *Attention* do not improve the results over the baseline either.

Our results are consistently improved on both datasets after introducing RNN architecture to model the relation between the contexts. It indicates that hierarchical structure is crucial to learn the context representation. Attention mechanisms contribute to the overall improvements. On MRDA, the AM was more useful when it was applied to the inputs of the RNN, whereas on SwDA when it was applied to the outputs. Our intuition is that in multiparty dialogs the dependency between the utterances should be weighted before being processed by the RNN.

Model	MRDA	SwDA
Baseline I	83.6	71.3
Baseline II	83.5	72.6
Max	58.5	48.0
Attention	83.5	72.4
RNN (LSTM)	83.8	73.1
RNN (GRU)	83.8	72.8
RNN-Output-Attention (LSTM)	84.1	<b>73.8</b>
RNN-Output-Attention (GRU)	84.0	73.1
RNN-Input-Attention (LSTM)	<b>84.3</b>	73.3
RNN-Input-Attention (GRU)	83.6	73.1

Table 4: Accuracy (%) of baselines and models with different context processing methods.

## 4.3 Impact of Context Length

Our experiments revealed that context length plays an important role for DA classification and the best length is corpus dependent. By experimenting in the context range of 0-5 utterances, we found that the best context length for MRDA is three utterances and two for SwDA. Table 5 shows the results at different context lengths.

<i>n</i> -context	MRDA	SwDA
1	83.8	73.1
2	83.9	<b>73.8</b>
3	<b>84.3</b>	73.5
4	84.0	73.1
5	84.0	72.9

Table 5: Comparison of accuracy (%) on different context lengths (*n*-context, where *n* is the number of sentences as context).

## 5 Comparison with Other Works

Table 6 compares our results with other works. To the best of our knowledge, Lee and Deroncourt (2016) is the newest research in DA classification, which published train/validation splits and claimed to be the state-of-the-art on that setup. Therefore, an accurate comparison of our results can be only done with this work. Our model yields comparable results to the state-of-the-art on both datasets, 84.3% against 84.6% on MRDA and 73.8% against 73.1% on SwDA. Ji et al. (2016) and Kalchbrenner and Blunsom (2013) obtained higher accuracy on SwDA but with different setup.

Model	MRDA	SwDA
Our best model	84.3	73.8
CNN-FF	84.6	73.1
LSTM-FF	84.3	69.6
HBM	81.3	—
LV-RNN	—	77.0
HCNN	—	73.9
CA-LSTM	—	72.6
HMM	—	71.0
Majority class	59.1	33.7

Table 6: Comparison of accuracy (%). *CNN-FF* and *LSTM-FF*: proposed in Lee and Deroncourt (2016), *HBM*: hidden backoff model (Ji and Bilmes, 2006). *LV-RNN*: latent variable RNN with conditional training (Ji et al., 2016). *HCNN*: hierarchical CNN (Kalchbrenner and Blunsom, 2013). *CA-LSTM*: contextual attentive LSTM (Shen and Lee, 2016). *HMM* Stolcke et al. (2000).

## 6 Conclusions

We explored different neural-based context representation learning methods for dialog act classification which combine RNN architectures with attention mechanisms at different context levels. Our results on two benchmark datasets reveal that using RNN architecture is important to learn the context representation. Moreover, attention mechanisms contribute to the overall improvements, however, the place where AM should be applied depends on the nature of the dataset.

## References

- Heike Adel and Hinrich Schütze. 2017. [Exploring different dimensions of attention for uncertainty detection](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 22–34. <http://www.aclweb.org/anthology/E17-1003>.
- Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1061–1064.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. [Neural machine translation by jointly learning to align and translate](#). *CoRR* abs/1409.0473. <http://arxiv.org/abs/1409.0473>.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. [On the properties of neural machine translation: Encoder-decoder approaches](#). *CoRR* abs/1409.1259. <http://arxiv.org/abs/1409.1259>.
- Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. [Attention-based models for speech recognition](#). *CoRR* abs/1506.07503. <http://arxiv.org/abs/1506.07503>.
- Junyoung Chung, Çağlar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. [Empirical evaluation of gated recurrent neural networks on sequence modeling](#). *CoRR* abs/1412.3555. <http://arxiv.org/abs/1412.3555>.
- Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. [Meeting Recorder Project: Dialog Act Labeling Guide](#). Technical report, ICSI Tech. Report. <https://goo.gl/TtLJIE>.
- John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. [Switchboard: Telephone speech corpus for research and development](#). In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*. IEEE Computer Society, Washington, DC, USA, ICASSP'92, pages 517–520. <http://dl.acm.org/citation.cfm?id=1895550.1895693>.
- David Golub and Xiaodong He. 2016. [Character-level question answering with attention](#). *CoRR* abs/1604.00727. <http://arxiv.org/abs/1604.00727>.
- M. Henderson, M. Gai, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young. 2012. [Discriminative spoken language understanding using word confusion networks](#). In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 176–181. <https://doi.org/10.1109/SLT.2012.6424218>.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Comput.* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. [The icisi meeting corpus](#), pages 364–367.
- Gang Ji and Jeff Bilmes. 2006. [Backoff model training using partially observed data: Application to dialog act tagging](#). In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL '06, pages 280–287. <https://doi.org/10.3115/1220835.1220871>.
- Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. [A latent variable recurrent neural network for discourse relation language models](#). *CoRR* abs/1603.01913. <http://arxiv.org/abs/1603.01913>.
- D. Jurafsky, E. Shriberg, and D. Biasca. 1997. [Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual](#). Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.
- Nal Kalchbrenner and Phil Blunsom. 2013. [Recurrent convolutional neural networks for discourse compositionality](#). *CoRR* abs/1306.3584. <http://arxiv.org/abs/1306.3584>.
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). *CoRR* abs/1408.5882. <http://arxiv.org/abs/1408.5882>.
- Ji Young Lee and Franck Dernoncourt. 2016. [Sequential short-text classification with recurrent and convolutional neural networks](#). *CoRR* abs/1603.03827. <http://arxiv.org/abs/1603.03827>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). *CoRR* abs/1310.4546. <http://arxiv.org/abs/1310.4546>.

- B. T. Polyak and A. B. Juditsky. 1992. [Acceleration of stochastic approximation by averaging](#). *SIAM J. Control Optim.* 30(4):838–855. <https://doi.org/10.1137/0330046>.
- Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2015. [The influence of context on dialogue act recognition](#). *CoRR* abs/1506.00839. <http://arxiv.org/abs/1506.00839>.
- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kociský, and Phil Blunsom. 2015. [Reasoning about entailment with neural attention](#). *CoRR* abs/1509.06664. <http://arxiv.org/abs/1509.06664>.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. The Association for Computational Linguistics, pages 379–389. <http://aclweb.org/anthology/D/D15/D15-1044.pdf>.
- 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](#). *CoRR* abs/1604.00077. <http://arxiv.org/abs/1604.00077>.
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. [The icsi meeting recorder dialog act \(mrda\) corpus](#). In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Cambridge, Massachusetts, USA, pages 97–100. <http://www.aclweb.org/anthology/W04-2319>.
- 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](#). *Comput. Linguist.* 26(3):339–373. <https://doi.org/10.1162/089120100561737>.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. [Hierarchical attention networks for document classification](#). In *HLT-NAACL*.
- Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. [ABCNN: attention-based convolutional neural network for modeling sentence pairs](#). *CoRR* abs/1512.05193. <http://arxiv.org/abs/1512.05193>.
- Ye Zhang and Byron C. Wallace. 2015. [A sensitivity analysis of \(and practitioners’ guide to\) convolutional neural networks for sentence classification](#). *CoRR* abs/1510.03820. <http://arxiv.org/abs/1510.03820>.
- Matthias Zimmermann. 2009. [Joint segmentation and classification of dialog acts using conditional random fields](#). In *INTERSPEECH*, pages 864–867.