Contextual Dialogue Act Classification for Open-Domain Conversational Agents

Ali Ahmadvand
Computer Science Department,
Emory University, Atlanta, GA
ali.ahmadvand@emory.edu

Jason Ingyu Choi
Computer Science Department,
Emory University, Atlanta, GA
in.gyu.choi@emory.edu

Eugene Agichtein
Computer Science Department,
Emory University, Atlanta, GA
eugene.agichtein@emory.edu

arXiv:2005.13804v1 [cs.CL] 28 May 2020

ABSTRACT

Classifying the general intent of the user utterance in a conversation, also known as Dialogue Act (DA), e.g., open-ended question, statement of opinion, or request for an opinion, is a key step in Natural Language Understanding (NLU) for conversational agents. While DA classification has been extensively studied in human-human conversations, it has not been sufficiently explored for the emerging open-domain automated conversational agents. Moreover, despite significant advances in utterance-level DA classification, full understanding of dialogue utterances requires conversational context. Another challenge is the lack of available labeled data for open-domain human-machine conversations. To address these problems, we propose a novel method, CDAC (Contextual Dialogue Act Classifier), a simple yet effective deep learning approach for contextual dialogue act classification. Specifically, we use transfer learning to adapt models trained on human-human conversations to predict dialogue acts in human-machine dialogues. To investigate the effectiveness of our method, we train our model on the well-known Switchboard human-human dialogue dataset, and fine-tune it for predicting dialogue acts in human-machine conversation data, collected as part of the Amazon Alexa Prize 2018 competition. The results show that the CDAC model outperforms an utterance-level state of the art baseline by 8.0% on the Switchboard dataset, and is comparable to the latest reported state-of-the-art contextual DA classification results. Furthermore, our results show that fine-tuning the CDAC model on a small sample of manually labeled human-machine conversations allows CDAC to more accurately predict dialogue acts in real users' conversations, suggesting a promising direction for future improvements.

1 INTRODUCTION AND RELATED WORK

In order for a conversational system to respond properly, it must first understand the intent of the user utterance. Dialogue Act (DA)

identification is a traditional approach to intent classification in dialogue system research, which aims to predict the goal of each utterance, such as information request, statement of opinion, greeting, opinion request, and others. [11] categorized DAs as having two main categories: 1) primary intents; 2) secondary intents. Each of these intents can further be divided into implicit or explicit. Since identifying these intents correctly is crucial for dialogue systems, this problem has been studied extensively for decades, for human-human conversations. Recently, this idea was also extended to human-machine conversations [7, 9, 13].

Utterances in natural conversations are contextually dependent in nature, which makes the DA prediction challenging. For example, the utterance like "Oh, yeah" can be interpreted as "Yes-Answer", "Accept-Agree", or "Backchannel", which requires the previous context to disambiguate [2, 9]. Therefore, we propose a context-aware model for this task.

For human-machine conversations, DA classification is more challenging due to three additional factors: 1) Often, human utterances are short (only 2.8 words on average in our data); 2) Automatic Speech Recognition (ASR) is still not quite a human level performance; 3) Lack of available open-domain labeled human-machine conversation data. To address these challenges, we propose a deep learning based model, which utilizes contextual evidence, such as preceding utterances alongside the system state information, for more accurate predictions. To reduce requirements for labeled training data, we follow the approach of pre-training the model followed by fine-tuning, which has proven its effectiveness on various natural language processing tasks such as question answering [3, 10].

To represent the utterance and system state for each conversation turn, we built on previous studies which identified effective features for human-human DA classification including syntax, prosody, and lexical cues (e.g., [4, 13]). We integrate many of these ideas into the proposed lexical and syntactic features in our CDAC system, and augment these with representation learning.

Recently, deep learning and representation learning approaches shown promising results on many tasks including text classification and DA classification (e.g., [1]). For instance, reference [1] proposed a context-based RNN model for dialogue act classification for the human-human Switchboard dataset, while reference [9] proposed a hierarchical CNN and RNN model for this task.

Reference [6] demonstrated the benefits of accurate DA classification for topic classification in open-domain dialogue systems. Inspired by the promising results of [6], [14], and [12], we propose a novel, yet relatively simple Contextual Dialogue Act Classifier (CDAC) model, which incorporates lexical, syntactic, semantic, and contextual evidence into DA classification. To our knowledge, CDAC is the first to extend and adapt the ideas [6] for contextual
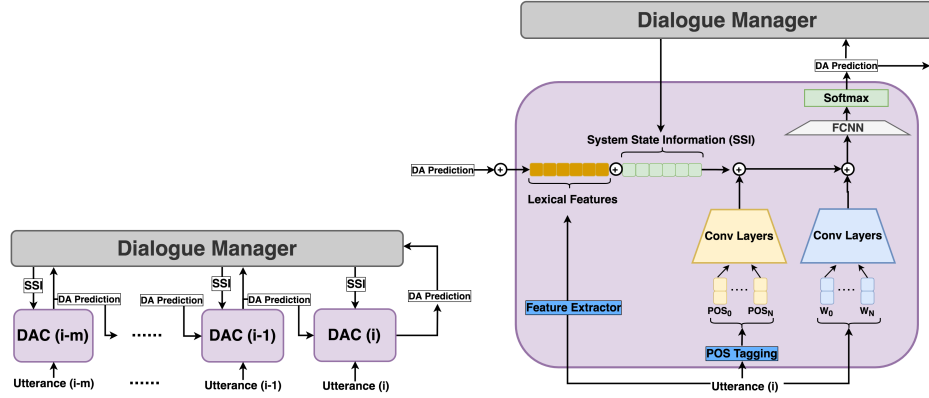
Figure 1: Contextual Dialogue Act Classifier (CDAC) architecture overview (left), where "DAC" is the dialogue act classifier for individual utterances (shown in detail on right). The extracted features, system state features, and details of the convolution (Conv) layers are described in detail in (Section 2.)

DA classification in open-domain conversational systems, such as those fielded in the Amazon Alexa 2018 Challenge. Interestingly, previous state-of-the-art DA models (e.g., References [9] and [7]), rely on complete conversations for context, including future utterances, while CDAC uses only the utterances from earlier in the conversation, which makes CDAC feasible for online (real-time) conversational DA classification. As far as we know, CDAC is also the first successful attempt to fine-tune a DA classification model trained on a dataset of *human-human conversations*, to predict DAs in open domain *human-machine* conversations.

In summary, our contributions are twofold: (1) development of a novel context-aware Dialogue Classification model, CDAC, for open domain human-machine conversations; (2) demonstrating promising results after fine-tuning CDAC trained on human-human conversations to human-machine conversations, which is a necessary step for intelligent open-domain conversational agents.

2 CONTEXTUAL DIALOGUE ACT CLASSIFIER (CDAC) MODEL

In this section, we describe our proposed method, CDAC, for contextual dialogue act classification. First, we describe the features used to represent user utterances, and the conversation and system context. Then, we present the CDAC model architecture, implementation, and training details.

Content and Context Representation. For individual utterance representation, we use both word embedding, and surface (lexical and syntactic) features, described next. The word representation weights are initialized using pre-trained Word2Vec embeddings (300 dimensions), and updated during training. The embedding-based utterance representation is augmented using three types of features: 1) lexical; 2) syntactic; 3) system state information (SSI), summarized below.

The CDAC model combines individual Dialogue Act Classifiers (DACs) for each utterance. Each DAC uses textual features as well as syntactic features, i.e., part of speech tags, are modeled using a convolutional layer pipeline, as shown in Figure 1 (right). Each "Conv Layers" in Figure 1 is a 3-layer CNN with kernels of size 1, 2, and 3, with 100 filter maps for each kernel. Each layer is implemented using a 2D-convolution followed by a max-pooling, batch

Lexical Features	Short Description
F_1 - Word Count	Word_{count} in utterance
F_2 - Char Count	Char_{count} in utterance
F_3 - Sentence Count	Sentence_{count} in utterance
F_4 - Average Word Count	Average Word_{count} in utterances
F_5 - Average Char Count	Average Char_{count} in utterances
F_6 - IsQuestion	Binary feature to check for "?"

Syntactic & SSI Features	Short Description
F_7 - POS Tagging	Part of speech tags
F_8 - Topic Distribution	Topic distribution vector
F_9 - Suggested Topic	Suggested topic to user
F_{10} - Suggested Item	Suggested entity to user
F_{11} - Speaker Id	ID assigned to each user

normalization, and relu activation function. To implement the batch normalization, we used a momentum of $M = 0.997$ and an epsilon of $\epsilon = 1e - 5$. Then, the output of both pipelines (syntactic and word embeddings) for each utterance representation, and the lexical features, as well as SSI features, are combined through a Fully Connected Neural Network layer (FCNN) with the size of 100. A dropout rate of 0.5 is applied at the FCNN layer to prevent the model from overfitting to the limited training data. Finally, softmax activation is used to obtain the final multi-class DA distribution. For training, categorical cross-entropy loss is minimized using Adam optimizer, with a $\alpha = 1e - 3$ learning rate and mini-batch size of 64.

The full Contextual DA Classifier model, CDAC, uses the DAC predictions from up to m previous turns in the conversation, where m effectively controls the scope of the conversation context. These DA predictions for the previous turns are appended to the current utterance representation vector through the FCNN layer, as shown in Figure 1 on right.

The window size m and other hyper-parameters above, largely follow the previous literature. Overall, our CDAC architecture and implementation choices builds on the ideas from the recent state of the art models for DA classification on the Switchboard dataset, while keeping the model simple and inexpensive to train.

System State Information (SSI) Features. We hypothesized that modeling context in human-machine conversations is similar to human-human conversations, but with a major difference that the system state information (SSI), unlike the state of a human, can be directly captured and represented as features. We incorporate SSI features alongside the utterance and context representation features

Utterance	Response	Topic	PS	ST	DA
Alexa, let's chat	Nice to meet you! Do you want to talk about music?	*Music*	*NONE*	*Music*	**fp:** Opening
I want some recent music from Drake	How about God's plan by Drake?	*Music*	*Open*	*NONE*	**sv:** Statement-opinion
I really like that song!	You're welcome. Do you want some more information about Drake?	*Music*	*Music*	*NONE*	**sv:** Statement-opinion
Sure, sounds good	Drake is a Canadian rapper with 30 million followers on Spotify. Do you want some news about him?	*Music*	*Music*	*News*	**aa:** Accept
I'm good, tell me some fun dog facts	Sure, do you want to know why dogs have wet noses?	*Animal*	*Music*	*NONE*	**qo:** Open-question
Yes please	Dog's wet noses are for absorbing scent chemicals! Do you want another one?	*Animal*	*Animal*	*NONE*	**aa:** Accept
Do you have a dog?	Unfortunately, I do not have one. What do you want to chat next?	*Phatic*	*Animal*	*NONE*	**qy:** Yes-No-Question
Can you tell me about penguins?	Sure, do you want to hear about the fastest penguin species?	*Animal*	*Animal*	*NONE*	**qo:** Open-Question
Okay, that sounds interesting	Gentoo penguins can reach swimming speed up to 22 mph. Do you want more about penguins?	*Animal*	*Animal*	*NONE*	**aa:** Accept
No, I'm done chatting for today	Nice to talk to you, good bye!	*Phatic*	*Animal*	*NONE*	**fc:** Closing

Table 1: An example human-machine conversation, where "Topic" is current system state (topic), "PS" is the previous system topic, "ST" is the system-initiated topic suggestion, if any, and "DA" is the manually chosen Dialogue Act label.

for each *turn* of the conversation. The SSI features include system topic distribution, the suggested topics (e.g., "Music"), and suggested items (e.g., specific artists). To encode topic distribution features, we used one-hot encoding, while specific items are represented using Word2Vec word embeddings of the words in the item names. Note that the SSI features were not used or available for human-human conversations.

Transfer Learning from Human-Human Conversations. To fine-tune the CDAC model from human-human to human-machine conversations, all the weights in the CDAC model are first trained on the human-human Switchboard dataset. Then, all of the network weights are tuned using the Alexa Prize data, but with a smaller learning rate of $\alpha = 1e - 4$.

3 EXPERIMENTAL SETUP

We now introduce the human-human (Switchboard) and human-machine (Alexa Prize) conversation datasets used for training and evaluating CDAC. Then, we explain the human annotation procedure to obtain ground truth labels for the human-machine conversations, followed by the experimental design and metrics.

Switchboard Dataset (Human-Human Conversations). Switchboard DA corpus [5] is a well-known telephone speech corpus, which contains 42 main DA labels[1]. The Switchboard corpus contains two official splits: the training split with 1,115 conversations and 196,258 utterances, and the test split, with 40 conversations and 7535 utterances.

Alexa Prize 2018 Dataset (Human-Machine Conversations). We collected the human-machine conversation data during the Amazon Alexa Prize 2018. 200 conversations of real users with our open-domain conversational agent, containing more than 3,000 utterances were randomly selected. Table 1 shows an example conversation[2] of a hypothetical (not real) user with the actual system responses. Two different human annotators were asked to manually label two hundred conversations in the human-machine Alexa prize data. The inter-annotator agreement was 0.790, and Kappa was 0.755, indicating strong agreement between the annotators. For the final ground truth label values, in case of disagreements, the label was randomly chosen between the two annotator labels. The distribution of annotated DAs[3] is reported in Table 2. The top four most frequent dialogue acts observed are Agree/Accept (aa), Conventional Opening (fp), Reject (ar), and Statement Opinion (sv), accounting for over 68.2% of the user utterances.

DA	Frequency	DA	Frequency
aa	655 (21.7%)	*fp*	501 (16.6%)
ar	478 (15.8%)	*sv*	425 (14.1%)
qo	227 (7.5%)	*fc*	198 (6.6%)
sd	154 (5.1%)	*b^m*	114 (3.8%)
no	107 (3.5%)	*qw*	80 (2.7%)
qy	48 (1.6%)	*%*	24 (0.8%)
ft	7 (0.2%)		

Table 2: Dialogue Acts (DA) frequency distribution in user utterances in the Alexa Prize dataset.

Switchboard Experimental Design. For CDAC model training, the training conversations were split into 1,000 for training and 115 for validation, leaving the test split untouched during training. For testing, some of the previous studies [1, 8, 13] used only 19 conversations of the available 40 test conversations. Instead, we follow the convention of Liu et al. [9] and use all 40 test conversations for evaluation. The main baseline model for this experiment is [13], based on a hidden Markov model, as it remained a state-of-the-art method for more than 10 years. Other state-of-the-art reported results are from the three recent DA classifiers described in references [1, 2, 8] respectively.

Alexa Prize Experimental Design. For the Alexa prize dataset experiment, Support Vector Machines (SVM) and Multinomial Bayes models are selected as baseline models, using lexical features (words) with tf-idf term weights due to simplicity and low requirements for labeled training data. Contextual features such as previous utterances and system state features are appended to the bag of words feature vector for each utterance. 5-fold cross-validation was used, where 4 folds were used for tuning the weights, and the last fold for the prediction. Finally, following the conventions of the DA classification literature, the main evaluation metric was **overall (micro-averaged) Accuracy**.

4 RESULTS AND DISCUSSION

In this section, we report the overall accuracy of CDAC in comparison to previous state-of-the-art baselines on Switchboard and Alexa data. Feature ablation and error analysis are also reported to provide insights into CDAC system performance.

DA Prediction Results on Switchboard Data. Our main results on the Switchboard dataset are summarized in Table 3. CDAC improves the baseline model [13] by 8.0%. Moreover, compared to the best known contextual model, we reach comparable results with a more general and simple model. Context window of size 3 yields the strongest performance. However, our results are on all 40 conversations in Switchboard dataset, while [2] used only 19 of the text conversations. We were unable to replicate the model and results reported in reference [2], due to the required model complexity and

---

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

[2] This is a representative conversation between the authors and the real system, since user utterances from live system deployment cannot be reported to protect user privacy.

[3] See *http://compprag.christopherpotts.net/swda.html for full description of DA labels.*

not having access to a working implementation or sufficient details to reproduce their exact system. In contrast, our proposed model is simpler, while producing state-of-the-art DA classification accuracy on the standard benchmark of human-human conversations.

Methods	Accuracy
Baseline Stolcke et al. [13]	71.00
Previous state of the art methods	
Kalchbrenner et al. [1]	73.90
Young et al. [8]	73.10
Bothe et al. [2]*	**77.34 (+8.9%)**
CDAC-2	76.40
CDAC-3	**76.70 (+8.0%)**
CDAC-4	76.51
Annotator Agreement	84.00

Table 3: DA classification Accuracy (micro-averaged) on Switchboard dataset, where (*) represents the latest reported state-of-the-art contextual DA classification[2]. CDAC-2,-3, and -4 stands for the model using on contexts of size 2, 3 and 4 turns, respectively.

DA Prediction Results on Alexa Prize dataset. Table 4 summarizes the baseline and CDAC performances on the human-machine Alexa Prize conversation data. CDAC outperforms traditional classification models such as SVM and Multinomial Bayes (without context) by about 22.7%. Furthermore, by adding context to SVM and Multinomial Bayes, there are 1.8% and 4.5% improvements over the respective baselines. Encouragingly, by pre-training the CDAC model on human-human Switchboard data, and then fine-tuning on the (limited) labeled human-machine data, CDAC achieves an additional 2.7% improvement. It is important to note that despite annotating for only 13 most common DA classes observed in the human-machine conversation data, compared to the 42 classes in Switchboard human-human data, DA performance degrades on human-machine conversations. This confirms our observation that human-machine DA classification is a more challenging task than for human-human conversations.

Methods	Accuracy
Without context	
Multinomial Bayes	58.21
SVM	65.73
With context	
Multinomial Bayes	59.26
SVM	68.70
CDAC	**73.25*(+6.6%)**
CDAC + Transfer Learning	**75.34*(+9.6%)**

Table 4: DA prediction micro-averaged Accuracy on Alexa dataset with and without context information (size 3 turns), where (*) represents significance levels of p < 0.05.

Feature Ablation and Error Analysis. Table 5 summarizes the change in accuracy by systematically removing feature sets on the Alexa prize human-machine conversation data. Both lexical and syntactic features are important for DA classification since removing either group decreased the accuracy. Interestingly, the most common error is distinguishing between statement-opinion and open-question labels. For instance, the utterance "I like to talk about animals" is challenging to classify, since without context, it is difficult to determine whether a user expressed an opinion, or

requested information from the system. Knowing the context and the system state can enable such disambiguation.

Syntactic Features	Lexical Features	Accuracy
-	-	73.94 (-1.64%)
-	✓	74.80 (-0.50%)
✓	-	74.91 (-0.35%)
✓	✓	**75.18**

Table 5: Feature ablation on CDAC with context window size 1. - and ✓ indicates features removed and added respectively.

In summary, we proposed a contextual Dialogue Act classification model, CDAC, which incorporates lexical, syntactic, and semantic information, in context. Additionally, we introduced a new group of context features to capture the internal system information. Finally, we demonstrated a promising use of fine-tuning on a limited set of labeled human-machine conversations, to decrease manual annotation requirements, and to utilize the existing human-human labeled conversation data. As a result, CDAC was able to outperform state-of-the-art DA classification baselines: by 8.0% on Switchboard data, and by 9.6% on the Alexa Data, and performed comparably to the latest reported and more complex state-of-the-art contextual DA classification model. CDAC was also shown to be general enough to be easily fine-tuned for DA classification in human-machine conversations. The implementation is released to the research community[4]. We believe CDAC represents a promising advance in general user intent classification for intelligent conversational agents.

REFERENCES

[1] P. Blunsom and N. Kalchbrenner. Recurrent convolutional neural networks for discourse compositionality. In *Proc. of the Workshop on Continuous Vector Space Models and their Compositionality*, 2013.

[2] C. Bothe, C. Weber, S. Magg, and S. Wermter. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *arXiv preprint arXiv:1805.06280*, 2018.

[3] Y.-A. Chung, H.-Y. Lee, and J. Glass. Supervised and unsupervised transfer learning for question answering. In *Proc. of NAACL-HLT*, 2017.

[4] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar. Dialogue act classification using a bayesian approach. In *Proc. of Conference on Speech and Computers.*, 2004.

[5] D. Jurafsky. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. In *Technical Report.* University of Colorado, 1997.

[6] C. Khatri, R. Goel, B. Hedayatnia, A. Metanillou, A. Venkatesh, R. Gabriel, and A. Mandal. Contextual topic modeling for dialog systems. In *Proc. of SLT*, 2018.

[7] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proc. of AAAI*, 2018.

[8] J. Y. Lee and F. Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. In *arXiv preprint arXiv:1603.03827*, 2016.

[9] Y. Liu, K. Han, Z. Tan, and Y. Lei. Using context information for dialog act classification in dnn framework. In *Proc. of EMNLP*, pages 2170–2178, 2017.

[10] L. Mou, Z. Meng, R. Yan, G. Li, Y. Xu, L. Zhang, and Z. Jin. How transferable are neural networks in nlp applications? In *In Proc. of EMNLP*, 2016.

[11] S. Pareti and T. Lando. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proc. of LREC*, 2018.

[12] C. Ruey-Cheng, E. Yulianti, M. Sanderson, and W. B. Croft. On the benefit of incorporating external features in a neural architecture for answer sentence selection. pages 1017–1020. In *Proc.* of SIGIR, 2017.

[13] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Computational linguistics*, pages 339–373, 2000.

[14] J. Wang, Z. Wang, D. Zhang, and J. Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *Proc* of IJCAI, 2017.

---

[4]Available at *https://github.com/emory-irlab/CDAC*