Using Reinforcement Learning for Dialogue Act Classification in Task-oriented Conversation Systems

Qingyang Xia

ABSTRACT

With rapid development and comprehensive application of artificial intelligence (AI), dialogue system, question-answering system and chatting robot go into people's daily lives. Therefore, DA classification plays an important role because it helps the system handle different types of user motivations. For this reason, most of dialogue systems, especially task-oriented dialogue systems, create a DA classification module in order to make the whole framework more vivid and simpler to develop and usually it does determine the overall quality of the output result. Previous work on DA classification focused on how to improve the performance of neural network models or deep learning models by modifying or mixing the model framework, most commonly the inner layers. However, little attention was paid on investigating the effect of reinforcement learning (RL) on DA classification. We simply use convolutional neural networks (CNN) as our baseline system and conduct our training progress using reinforcement learning. Compared to traditional methods and machine learning methods, our method firstly exploits reinforcement learning for DA classification and we conduct certain experiments to demonstrate the significant improvement of classification results in task-oriented dialogue systems. We create two Chinese-language industrial datasets by collecting data from task-oriented conversation systems in restaurant-ordering and cooking assistant and set up live study experiments to prove that our method not only improves the performance of DA classification, but also makes the overall quality of dialogue system better.

Xia Qingyang, Tsinghua University, Chengfu Street, Haidian District, Beijing 100084, China
INTRODUCTION

All the sentences in conversations represent some kind of intention of the speaker, understanding the speaker's purpose leads to high-quality and effective conversations. For the above reasons, DA becomes one of the most important and intrinsic part of a sentence. So it is worth taking great effort for utterance dialogue act classification and recognition when developing human-to-computer dialog systems. In recent years, lots of internet companies have been developing their chatting robots, which are usually open-domain and based on massive data. For instance, Microsoft launches their first version of chatting robot names "Xiao ice" in 2014 and continually upgrades its performance and internal models until now. They recently announce that "Xiao ice" has evolved to its fifth version and its technological focus moves from fundamental artificial intelligence to "senior sense" generation. This interactive mode combines original emotion quotient (EQ) kernel conversation engine with other input sources like images, textual documents, sound utterances and even movies so that this pattern can obviously strengthen the interactive experience to the level of human natural reaction. Besides these famous applications, enterprises still want to make specific-domain dialog systems which not only follows the trending of recent information industry but also helps companies effectively reduce their human resource cost. Effective dialogue act (DA) classification also helps get rid of the constraint that conversation robot can only give simple response to user utterances. Clear detection of sentences' aim provides conversation robots the ability of leading the theme of the conversation instead of rigidly supporting with an answer or response.

Siri and Cortana, shown in figure 1, are also typical applications of chatting robot. Siri is a dialogue system product developed by Apple Company and it is widely implemented in current iPhone operating system (iOS). Cortana is the first intelligent assistant product of Microsoft to help users manage their individual schedule. Both of them adopt methods like cloud computing, search engines and unstructured data analyzing to understand the textual and semantic target. The appearance of these products symbolizes the rapid development of personal computing. It is clear that DA recognition technique plays an important role on the quality of chatting systems and the performance of DA recognition module can be optimized from different perspectives. Our work only pay attention on the textual and literal factors and ignore others like data collecting for this specific research.
In this article, we will begin with a brief introduction to the comprehensive background of dialogue act classification and conversation system, followed by the summary of existing research work and efforts on this task. Then we propose our method in the next part and present our experiments and results with our objective analyzing. The paper ends with some further discussion and expected future research directions. The contributions of our paper can be concluded as follows: 1) We propose a new way that we can conduct DA classification using RL. 2) Our model gets excellent result in Chinese datasets. 3) We improve the live performance of conversation systems though live experiments.

RELATED WORK

Methods discussed in previous research can be divided into two categories, traditional methods and methods based on deep learning. Traditional methods include support vector machines(SVM), hidden Markov models(HMM), graphical models, conditional random fields(CRF), maximum entropy. They mostly do much work on dialog context information factors. These factors can be used as indicators of DA analysis because we can intuitively find out some intuitive relations between different utterances as long as their DA labels change. For example, when speaker A asks a yes-no question like "Have you ever gone to the United States?", the speaker B will certainly give an answer to this yes-no question, such as "Yeah" which should have a type of "yes-answer" or "not yet" as a type of "no-answer". The significance of previous observation has been verified by methods including hidden Markov model(HMM) and conditional random fields(CRF). As a different trend of DA classification methods, methods based on deep learning have completely different focuses, such as using hierarchical neural networks to exploit and capture different features like syntactic, literal and lexical features. These kind of context information is publicly regarded as powerful characteristic in DA analysis, and it has been adequately explored by models like DNN[1] (Rojas-Barahona et al., 2016), recurrent CNN[2] (Kalchbrenner and Blunsom, 2013) and hierarchical LSTM and CNN[3] (Yang Liu et al., 2017). However, previous research did not pay much attention on the potential and feasibility of involving reinforcement learning(RL) in resolving...
DA classification problem, which is the most important part in a task-oriented conversation systems[4][5][6][7][8][9][10].

**Reinforcement Learning(RL)** is a framework of machine learning which teaches the model how to make decisions in specific circumstances through rewards from iterative trials and errors. It has been proved that RL has excellent performance on learning optimal dialog management policy for specific-domain(task-oriented) conversation systems [11](Georgila and Traum, 2011). Besides improvement on system performance, RL also helps reduce the amount of data for training[12] (Gasic 2010). Furthermore, many of RL approaches have been adopted for other tasks[13] like incremental dialogue processing[14] (Hatim Khouzaimi et al., 2016) and dialog state management[15] (Tiancheng Zhao and Maxine Eskenazi, 2016).

**METHODS**

**Baseline CNN model**

We build up a simple CNN model as our baseline model. It only classifies the utterances by learning its textual information without any context information. As a fact, CNN model is widely used as baseline model in many classification and generation tasks[16] (Collobert et al., 2011). The context independent CNN model is shown in Figure 2.

Utterances with n words are firstly represented with vectors w[1..n], where wi is d- dimensional embedding vector of the ith word. After zero padding, we input the sequence w'[1..n] to a convolution layer and a filter map f is used for the convolution operation. Then we apply a max pooling layer to get fixed-length vector as our result feature vector representation of the utterance. At last, we use a multi-layer perceptron(MLP) to do the sentence DA classification.

**Reinforcement Learning method**

In task-oriented systems, we call conversations which are used to accomplish certain target, episodes. For example, a restaurant ordering system should help customers order their dishes and pay bills. We find that dynamic changes and special turns play an important role in success of independent tasks. We take these factors into consideration and feed them back as reward. Table I shows an example episode.
TABLE I. EXAMPLE CONVERSATION OF DIALOG SYSTEM.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Sentence</th>
<th>DA type</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>What can I help you?</td>
<td></td>
</tr>
<tr>
<td>User</td>
<td>I would like to have a spicy dish.</td>
<td>ASK_RECOMMEND</td>
</tr>
<tr>
<td>System</td>
<td>I recommend Shredded pork and green pepper.</td>
<td></td>
</tr>
<tr>
<td>User</td>
<td>Anything else?</td>
<td>CHANGE_RECOMMEND</td>
</tr>
<tr>
<td>System</td>
<td>I recommend Mapo Tofu.</td>
<td></td>
</tr>
<tr>
<td>User</td>
<td>This is fine, take one please.</td>
<td>ORDER</td>
</tr>
<tr>
<td>System</td>
<td>Okay, one Mapo Tofu.</td>
<td></td>
</tr>
<tr>
<td>User</td>
<td>That’s it, thank you.</td>
<td>FINISH</td>
</tr>
</tbody>
</table>

For each episode, when the model receives a new utterance, it decides a DA label from the DA label corpus \( \text{LABEL} = \{\text{label}^[1..m]\} \) to match the input sentence, where \( m \) is the total number of DA label types. Therefore, the action space is exactly the DA label corpus as shown below:

\[
A = \text{LABEL} = \{\text{label}_1, \text{label}_2 \ldots \text{label}_m\} \quad (1)
\]

The reward function of each utterance consists of the following elements:

\[
r = \lambda_1 r_1 + \lambda_2 r_2 \quad (2)
\]

We assume that earlier the misunderstanding of dialog act appears in an episode, more difficult the conversation theme can be adjusted back to the original trend. In order to give expression to this phenomenon, we set the reward as below:

\[
r_1 = \gamma^i \cdot \text{LC}_i \quad (3)
\]

We set \( \gamma = 0.9 \) where \( i \in \{1, 2 \ldots m\} \) denotes the position of sentence in the episode and \( \text{LC}_i \) denotes the correctness of DA label classification. \( \text{LC}_i \) equals to 1
when the label is correct and equals to 0 in contrast. This part of reward makes sure that earlier DA recognition brings more terrible impact.

On the other hand, we need to measure the conversational fluency and user interaction experience.

\[ r_2 = \frac{N_w}{m} \]  

(4)

Since the average number of DA labels of all episodes is 6.5, we set \( N_w = -10 \) if the number of wrong labels exceeds 3, set \( N_w = -2 \) if the number is 1 or 2 and set \( N_w = 5 \) if no mislabeling occurs. We think that the conversation mission is extremely difficult to complete if the system fails to understand the dialog act, which happens when there are over 3 wrong labels in a single round. It also destroys the user experience in a severe way. Therefore, we set \( \lambda_f = 0.4 \) and \( \lambda_2 = 0.6 \) for the final reward. A reward will be observed after each end of utterance input.

**Policy Gradient**

The expected reward will be as below:

\[ J(\theta) = \sum_{x \in X} \pi(s, a)r \]  

(5)

Where \( X \) is the corpus of user utterances, \( r \) refers to the reward function of each utterance as previously mentioned. According to our experimental situation, state \( s \) can be represented as input \( x \) because we get \( \pi(s, a) \) from a pre-trained baseline network so \( r \) can be directly substituted as equation (2):

\[ \pi(s, a) = P(a|s) \]  

(6)

So on the basis of policy gradient theorem, we can derive the gradient of function \( J(\theta) \) with respect to the model’s parameters \( \theta \), indicated as \( \nabla_\theta J(\theta) \), as following:

\[ \nabla_\theta J(\theta) = \sum_{x \in X} \sum_{a \in A} \pi(s, a)r \cdot \nabla_\theta \log \pi_\theta(s, a) \]  

\[ \nabla_\theta J(\theta) = \sum_{x \in X} E_{a \sim \pi(s, \cdot)} [r \cdot \nabla_\theta \log \pi_\theta(s, a)] \]  

(7)

As the form of equation (7), we define the part in the bracket as the loss of our RL framework.
EXPERIMENTS AND RESULT

Datasets

SIMORDER

This corpus contains 228 episodes in the restaurant ordering scenario. The utterances are made by a customer and the developing ordering robot. Each sentence is manually labelled with 10 different types of DA labels. We randomly pick 18 conversations as our testing data and the remaining 210 conversations are used as training set.

MEIFRIDGE

Similar to the former, this corpus contains 326 episodes between user and intelligent cooking assistant. 8 different types of DA labels are used for each user utterance. In the same way, we choose 20 episodes to be the testing data and training data consists of the other 306 episodes.

Live Experiment Set up

We invite 52 volunteers to participate in our live experiment, each of them is asked to make 5 dialogues by using the ordering system and the same amount of dialogues with cooking assistant. When all experiments are completed, 520 dialogues and 3214 user utterances are gathered. Half of the data are processed by the baseline model and the other half is handled by RL. Volunteers need to label the correctness for each DA classification result and mark whether the dialogue successfully reach the target. In addition, volunteers are asked to finish a questionnaire after each conversation finishes. The questionnaire is listed as below. The questions are measured with a score from 1 to 5. 1 and 5 represent the worst performance and best performance respectively:

EFFICIENCY
Whether the system gets your point and returns information or result you want?

COMPREHENSION
How effective does the system understand your sentence when you use flexible sentence form like short sentence?

HELPFULNESS
Whether the system make you feel easier to finish your target during the conversation?

OVERALL PERFORMANCE
What do you think about the over quality of the system?
Results

Table II provides the results of both datasets. The values seem to be satisfactory because there are only a few DA labels in the label corpus for such specific domain and task. Compare to those academic datasets with over 40 labels, usually task-oriented dialogue systems have an average number of 7 different types of labels. So the overall results are around 90% and can be clearly explained. Task completion is judged by the number of wrong DA classification results with a critical value of 3. This draws lessons from the reward function in RL framework. The task will be marked successfully completed if the wrong DA labels does not reach 3 and vice versa. We observe that RL obviously performs better than the baseline model and makes an improvement in DA classification result for both datasets. Furthermore, it is intuitive that better DA classification results lead to better task completion ratio. So RL takes all-round advantages.

Table III shows the mean values of live study experiments. In the same way, RL takes the leader position in both task completion ratio and DA classification performance. But in subjective scoring criterion, the results also give part of the factors at which baseline model beats RL. When we take a closer look at the two questions on which baseline model has higher score than our RL model, it can be found that the difference between the two scores of two models are significantly smaller than the difference of scores on other subjective questions. The most possible explanation is that it is hard for user to notice the difference between subjective questions when the difference of feelings is not a wide gap. However, the

<table>
<thead>
<tr>
<th>TABLE II. RESULTS OF DATASETS.</th>
<th>SimOrder</th>
<th></th>
<th>MeiFridge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>RL</td>
<td>Baseline</td>
<td>RL</td>
</tr>
<tr>
<td>Task completion</td>
<td>0.89</td>
<td>0.95</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>DA classification</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.88</td>
</tr>
</tbody>
</table>

| TABLE III. LIVE STUDY RESULTS. |  |  |
|-------------------------------|  |  |
|                               | Baseline | RL |
| Task completion               | 0.82     | 0.86 |
| DA classification             | 0.78     | 0.83 |
| efficiency                    | 3.83     | 3.82 |
| comprehension                 | 3.65     | 3.56 |
| helpfulness                   | 4.12     | 4.54 |
| overall performance           | 4.05     | 4.32 |
overall performance of RL is preferred by most of users, which implies a general feeling that RL is better than the baseline.

CONCLUSION

We propose a new approach that utilizing RL framework for dialogue act classification in task-oriented dialogue systems. In order to demonstrate the effect of adopting RL, we use simple CNN as our baseline model and conduct several experiments to prove that using reinforcement learning (RL) for this DA classification task not only helps get good results in DA classification task, but also improves the overall performance of the dialogue system. Our results represent all-round victory of adopting new methods in task-oriented systems compared to traditional methods. We believe that there is still much work to be done for improvement of DA classification or dialogue system by implementing RL. We plan to work on the application of user simulation techniques to help improve such tasks. It is possibly beneficial for further improvement of DA classification performance.

ACKNOWLEDGEMENTS

The author is grateful for his tutors, Zhu Xiaoyan and Huang Minlie for their bovine, rigorous guidance during his investigation and essay-writing. Also, the author also thanks Zhang Zheng and Liu Biao for lots of meaningful discussions and various experience sharing. Lastly, the author thanks Dai Wentao for the supplying of the conversation system data. All above appreciation is in great sincere.

REFERENCES