Dialog Act Modeling for Virtual Personal Assistant Applications Using a Small Volume of Labeled Data and Domain Knowledge

Donghyeon Lee, Jinsik Lee, Eun-Kyoung Kim, Jaewon Lee
DMC R&D Center, Samsung Electronics, Suwon, South Korea
{dh.semko.lee, jinsik12.lee, ekkim.kim, jwonlee}@samsung.com

Abstract

Recently, virtual personal assistant (VPA) applications have been employed in mobile devices, which provide a natural and convenient interface between human and machines. As the VPA services become popular, consumers demand for a wider service than their scope, so the rapid development becomes more important. This paper introduces a dialog act modeling approach for VPA applications, which is an extension of a Latent Dirichlet Allocation model. This approach enables the rapid and cost-effective development by reducing human efforts for manual labeling and the development of a fail-safe product by incorporating domain knowledge such as dictionaries, cross-lingual data, and logic rules. The experimental results showed that a reliable and high-performance dialog act model was built only with a small volume of labeled data and domain knowledge.

Index Terms: dialog act modeling, spoken language understanding, virtual personal assistant

1. Introduction

A virtual personal assistant (VPA) application is able to perform various tasks such as answering questions, making recommendations, and performing actions. For example, people can make queries like “How’s the weather today?”, “Show me the restaurants nearby,” and “Turn off the wi-fi.”

In the VPA, dialog act (DA) modeling serves a very important role such as the identification of the user’s intention and the determination of the system action. The dialog act is a domain dependent and function-level concept, while the speech act is a domain independent and surface-level concept [1, 2, 3]. For example, in the case of sentence “How is the temperature today?”; the dialog act is GetTemperature, but the speech act is WhQuestion. Hence, a dialog act tag set is differently defined according to domains and applications.

There are two main approaches to a dialog act modeling: a grammar-based approach [4, 5, 6] and a statistical approach [7, 8, 9]. According to the recent studies, the statistical approach requires less human efforts than the grammar-based approach. However, to achieve the better performance in the statistical approach, a large amount of labeled data is generally required.

This paper proposes a dialog act model to alleviate human efforts for manual labeling, which is based on a Latent Dirichlet Allocation (LDA) [10] based approach for clustering utterances. In addition, for the fail-safe and rapid development, this model was extended to use various domain knowledge: dictionaries, cross-lingual data, and logic rules. The use of domain knowledge significantly reduces the manual labeling cost and improves the performance compared to the fully supervised methods.

Table 1: The example utterances with the corresponding dialog act tags.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dialog Act</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music</td>
<td>PlayMusic</td>
<td>Play me a song Avril Lavigne.</td>
</tr>
<tr>
<td></td>
<td>AddToList</td>
<td>Put the song on the playlist.</td>
</tr>
<tr>
<td>Navigation</td>
<td>Move to</td>
<td>Navigate to my apartment.</td>
</tr>
<tr>
<td></td>
<td>GetDistance</td>
<td>How far from here?</td>
</tr>
<tr>
<td>Schedule</td>
<td>SetAlarm</td>
<td>Wake me up at 7:00 AM.</td>
</tr>
<tr>
<td></td>
<td>GetTime</td>
<td>What time is it in Cleveland?</td>
</tr>
<tr>
<td>Weather</td>
<td>GetTemperature</td>
<td>What is the degrees today?</td>
</tr>
<tr>
<td></td>
<td>IsRain</td>
<td>Should I bring an umbrella?</td>
</tr>
</tbody>
</table>

The organization of this paper is as follows. Section 2 introduces the related work to reduce human efforts in the dialog act modeling. Section 3 proposes the LDA-based dialog act modeling approach. Section 4 presents the experimental results and discussions. Finally, Section 5 concludes this paper.

2. Related work

Many researches have been conducted to reduce the human labor required in dialog act modeling. To tackle this problem, combining an active learning approach and a semi-supervised learning approach was introduced [11]. A statistical model was built from a small volume of labeled data by applying the supervised learning methods, and then the statistical model was applied to the labeling of a large volume of unlabeled data. The data with a low confidence score were selectively chosen for the human labeling.

To extend domain and language, the dialog act modeling requires additional labeled data. In [12], a transfer learning based domain adaptation method was proposed. In [13, 14, 15, 16], machine translation technologies were applied to the cross-lingual spoken language understanding problem. The translated results were used to utilize existing resources from another language.

The use of unlabeled or automatically labeled data in the model training may degrade the reliability of the model. This can be resolved by reflecting domain knowledge. In [17], external knowledge such as web search logs were incorporated into a joint model for domains, dialog acts, and named entities. In [18], logic rules were combined with LDA for topic modeling.

Compared with the previous researches, the main contributions of this work are 1) the design of an LDA-based dialog act model which utilizes various domain knowledge to achieve the higher reliability and higher performance and 2) the application of a semi-supervised learning approach to reduce the human efforts for manual labeling.
abilities are stored in the dictionary $D$.

The proposed model uses seed labeled data and logic rules to represent the word types and the corresponding probabilities. This dictionary mostly includes named entities and functional words. For the out-of-dictionary words, the word type is set as it is and the probability is defined as 1.

### 3.3. Cross-lingual data

The development of VPAs with a multi-language environment requires the data preparation for each language. However, if labeled data are already prepared for one language, this can be used for another language. This work employs cross-lingual data as the previous work [13, 14, 15, 16].

There are two strategies on how to use the cross-lingual data: TrainOnTarget and TestOnSource. The TrainOnTarget approach trains a model using the translated dataset to a target language, and the prediction of a dialog act is performed on an input utterance written in the target language. On the contrary, the TestOnSource approach trains a model using the dataset written in a source language, and the prediction of a dialog act is performed on an input utterance which is translated from the target language to the source language.

This work adopts the TestOnSource approach. The labeled data written in a source language are used to train a dialog act model. Unlabeled data written in a target language are translated to the source language using a machine translation system.

### 3.4. Logic rule

The use of the logic rules contributes to the improvement of the clustering performance because it affects the hidden topic $z$ and its multinomials. This work uses the First-Order Logic (FOL) convention for the representation of logic rules. The followings are the variables and predicates defined.

- $Z(u, k)$ is true if $z_u = k$, and false otherwise.
- $T(u, i, c)$ is true if $t_{u,i} = c$, and false otherwise.

For example, the following logic rule represents that, for an utterance $u$, the dialog act should be set to SetAlarm when the word types time and set co-occur.

$$
\forall_{i,j} : T(u, i, \text{time}) \land T(u, j, \text{set}) \implies Z(u, \text{SetAlarm}) \tag{1}
$$

The rules written by experts are stored in FOL knowledge base $KB = \{(\lambda_1, \psi_1), \ldots, (\lambda_N, \psi_N)\}$. The $KB$ consists of $N$ pairs; $\psi_n$ is an FOL rule, and $\lambda_n$ is a weight of the rule which is set by the experts according to the significance of the rule. The inference method using the knowledge base $KB$ will be explained in Section 3.6.

### 3.5. Generative Story

The full generative story of the dialog act model proposed in this paper is presented in Algorithm 1.

### 3.6. Inference

The dialog act for a given utterance is sampled from the probability given by the following equation. For sampling, the Gibbs sampling method was used [20]. In our model, all hyper-parameters were set as 0.1.
Algorithm 1 The full generative story of the dialog act model.

Choose a distribution $\pi \sim \text{Dir}(\alpha)$.

For each topic $k = 1, \ldots, K$:
- choose a distribution $\phi_k \sim \text{Dir}(\phi_0)$.
- choose a distribution $\theta_k \sim \text{Dir}(\theta_0)$.

For each utterance $u = 1, \ldots, U$:
- If $L_u = k$, then $z_u = k$.
- else select a topic $z_u \sim \text{Multi}(\pi)$.
- Select a pDA $p_u \sim \text{Multi}(\theta_{z_u})$.
- For word types $t_{u,i}$ in utterance $u$:
  - select a word type $u_{i,j} \sim \text{Multi}(\phi_{z_u})$.
  - select a word from $p(w_{u,i}|t_{u,i}, D)$.

\[
p(z_u|\mathbf{z}_{-u}, \mathbf{t}, \mathbf{w}, \mathbf{p}, \mathbf{D}, \alpha, \psi_0, \theta_0, KB, \mathbf{o}) \propto \left( \sum_n n(z_u, t_{u,i}) + \alpha_{z_u,t_{u,i}} \right) \times \prod_{u,i} \frac{n(z_u, t_{u,i})}{\sum_n n(z_u, t_{u,i})} \times p(w_{u,i}|t_{u,i}, D) \times \exp \left( \sum_{g \in G(\psi_0)} \lambda_g \mathbf{I}_g(z_u \cup \mathbf{z}_{-u}) \right)
\]

The notation is as follows: $n(z_u)$ is the number of occurrences of the dialog act; $n(z_u, t_{u,i})$ is the number of occurrences of the dialog act-pDA pairs; $n(z_u, t_{u,i}, D)$ is the number of occurrences of the dialog act-word type pairs; $\mathbf{V}_u$ is the number of words in an utterance $u$; $KB$ is the number of unique dialog acts; and $|C|$ is the number of unique word types.

The knowledge base $KB$ mentioned in Section 3.4 is incorporated into the proposed model through groundings. As shown in the exp$(\cdot)$ term above, for each FOL rule $\psi_0$, $G(\psi_0)$ is defined as a set of mapped groundings with the variables defined in $\psi_0$. The $\mathbf{I}_g(\cdot)$ is defined for each grounding; the $\mathbf{I}_g(\cdot)$ has value 1 when $g$ is true and 0 when $g$ is false.

4. Results and discussions

4.1. Corpus

The user utterances were collected for four domains: music, navigation, schedule, and weather. The dialog act tag set was defined based on the collected corpus (Table 2), and all utterances were manually annotated. Table 3 shows the statistics of labeled corpora. In the experiments, we used 80% of the Korean corpus as the training set and 20% as the test set. The English corpus was used for the cross-lingual data.

4.2. Domain knowledge

The model proposed in this paper incorporates domain knowledge: dictionaries, cross-lingual data, and logic rules. The dictionaries were constructed from the gazetteers and databases for VPA services. The cross-lingual data, the utterances which belong to the different languages but the same domain and dialog act, was used to train a dialog act model for the prediction of pDA. The logic rules were written based on the following information: high-frequency unigrams, bigrams, word pairs, and keywords extracted by the LDA-based clustering.

4.3. Experimental setup

The experiments were performed under various conditions according to the combinations of domain knowledge and the proportions of labeled and unlabeled data (Table 4). The conditions of domain knowledge combinations are represented by attaching suffixes: D for the use of dictionary, R for the use of logic rules, and T for the use of cross-lingual data. The annotated data (Table 3) were divided into the ‘labeled data’ and the ‘unlabeled data’; the ‘labeled data’ set includes the annotated dialog act tag information, while the ‘unlabeled data’ set includes the predicted dialog act tag information. Different from the LDA approaches, the traditional supervised model (SUP) uses labeled data only; the Maximum Entropy model was adopted with unigrams, bigrams, and trigrams word features in the experiments.

4.4. Experimental results

The accuracies of dialog act predictions were increased by using domain knowledge and more labeled data (Table 5). In particular, the effect of the use of domain knowledge was significant when the small-sized labeled data were used.

The logic rules significantly contributed to the performance improvement: LDA_DRT, LDA_DR, and LDA_R achieved much higher accuracies than LDA_DT, LDA_D, and LDA, respectively. For example, the dialog acts AlarmOn and AlarmOff consist of the utterances having similar patterns, so they were often grouped into the same cluster. By using the logic rules, these utterances can be clustered separately. The use of the dictionary was also effective because it prevents the distribution distortion. However, the use of cross-lingual data did not show a significant improvement.

Compared with the traditional supervised approach (SUP), the proposed model (LDA_DRT) showed the higher accuracies (Figure 2). The difference of the accuracies between these two...
models became larger when the small-sized labeled data were used. This shows the proposed model is more effective when the size of labeled data is small and the domain knowledge can be easily built.

4.5. Discussions

4.5.1. Writing rules

Some unintended clusters may be created by an LDA-based approach because it only depends on the word distribution. The experimental results showed that the use of logic rules alleviated this problem. To build a fail-safe product, it is mandatory to write sophisticated rules by experts to secure controllability.

In this work, the experts who have more than 10 years of related work experience wrote the rules by referring the word frequency information.

This paper introduced a method to reduce the human efforts to write rules. Instead of starting from the scratch, the rules can be written by referring to the seed labeled data. In addition, the rules can be refined by comparing the clustered results of unlabeled data and the seed labeled data. The rule refinement procedure should be performed iteratively to write high-quality rules, so the active learning method and the use of annotation tools would be beneficial to increase the efficiency.

4.5.2. Translation performance

The experimental results showed that the use of the cross-lingual data did not significantly contribute to the performance improvement. In this work, the Korean to English translator was used to obtain pDA. However, because of the limitation of the technology (Korean and English have very different grammar structures), the translation results contained many errors which acted as the noise. In the other cross-lingual approaches using machine translation systems, the noise reduction techniques were conducted to prevent the propagation of errors. If this technique was adopted to this work, the use of the cross-lingual data could contribute to the performance improvement.

5. Conclusions

This paper introduced a dialog act model for VPAs. The proposed model is an extension version of the LDA-based model to incorporate domain knowledge such as dictionaries, cross-lingual data, and logic rules. This enables to reduce the human efforts, rapidly build a fail-safe product, and readily extend scenarios, domains, and languages. The experimental results showed that the proposed model outperformed the traditional supervised approach especially when only small-sized labeled data were available.

6. References


