Speaker Role Contextual Modeling for Language Understanding and Dialogue Policy Learning

Ta-Chung Chi[⋆] **Po-Chun Chen**[⋆] **Shang-Yu Su**[†] **Yun-Nung Chen**[⋆]
[⋆]Department of Computer Science and Information Engineering
[†]Graduate Institute of Electrical Engineering
National Taiwan University
{b02902019,r06922028,r05921117}@ntu.edu.tw y.v.chen@ieee.org

Abstract

Language understanding (LU) and dialogue policy learning are two essential components in conversational systems. Human-human dialogues are not well-controlled and often random and unpredictable due to their own goals and speaking habits. This paper proposes a role-based contextual model to consider different speaker roles independently based on the various speaking patterns in the multi-turn dialogues. The experiments on the benchmark dataset show that the proposed role-based model successfully learns role-specific behavioral patterns for contextual encoding and then significantly improves language understanding and dialogue policy learning tasks[1].

1 Introduction

Spoken dialogue systems that can help users to solve complex tasks such as booking a movie ticket become an emerging research topic in the artificial intelligence and natural language processing area. With a well-designed dialogue system as an intelligent personal assistant, people can accomplish certain tasks more easily via natural language interactions. Today, there are several virtual intelligent assistants, such as Apple's Siri, Google's Home, Microsoft's Cortana, and Amazon's Echo. Recent advance of deep learning has inspired many applications of neural models to dialogue systems. Wen et al. (2017), Bordes et al. (2017), and Li et al. (2017) introduced network-based end-to-end trainable task-oriented dialogue systems.

A key component of the understanding system is a language understanding (LU) module—it parses user utterances into semantic frames that capture the core meaning, where three main tasks of LU are domain classification, intent determination, and slot filling (Tur and De Mori, 2011). A typical pipeline of LU is to first decide the domain given the input utterance, and based on the domain, to predict the intent and to fill associated slots corresponding to a domain-specific semantic template. Recent advance of deep learning has inspired many applications of neural models to natural language processing tasks. With the power of deep learning, there are emerging better approaches of LU (Hakkani-Tür et al., 2016; Chen et al., 2016b,a; Wang et al., 2016). However, most of above work focused on single-turn interactions, where each utterance is treated independently.

The contextual information has been shown useful for LU (Bhargava et al., 2013; Xu and Sarikaya, 2014; Chen et al., 2015; Sun et al., 2016). For example, the Figure 1 shows conversational utterances, where the intent of the highlighted tourist utterance is to ask about location information, but it is difficult to understand without contexts. Hence, it is more likely to estimate the location-related intent given the contextual utterance about location recommendation. Contextual information has been incorporated into the recurrent neural network (RNN) for improved domain classification, intent prediction, and slot filling (Xu and Sarikaya, 2014; Shi et al., 2015; Weston et al., 2015; Chen et al., 2016c). The LU output is semantic representations of users' behaviors, and then flows to the downstream dialogue management component in order to decide which action the system should take next, as called *dialogue policy*. It is intuitive that better understanding could improve the dialogue policy learning, so that the dialogue management can be further

[1]The source code is available at: https://github.com/MiuLab/Spk-Dialogue.

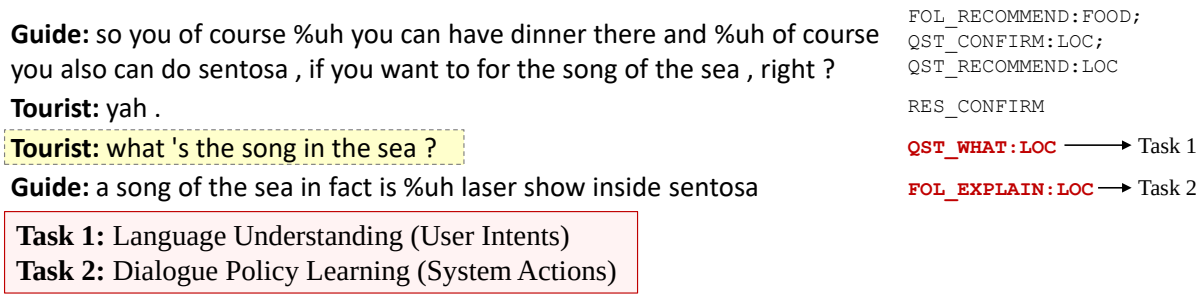

Figure 1: The human-human conversational utterances and their associated semantics from DSTC4.

boosted through interactions (Li et al., 2017).

Most of previous dialogue systems did not take speaker roles into consideration. However, we discover that different speaker roles can cause notable variance in speaking habits and later affect the system performance differently (Chen et al., 2017). From Figure 1, the benchmark dialogue dataset, Dialogue State Tracking Challenge 4 (DSTC4) (Kim et al., 2016)[2], contains two specific roles, a tourist and a guide. Under the scenario of dialogue systems and the communication patterns, we take the tourist as a user and the guide as the dialogue agent (system). During conversations, the user may focus on not only *reasoning (user history)* but also *listening (agent history)*, so different speaker roles could provide various cues for better understanding and policy learning.

This paper focuses on LU and dialogue policy learning, which targets the understanding of tourist's natural language (LU; language understanding) and the prediction of how the system should respond (SAP; system action prediction) respectively. In order to comprehend what the tourist is talking about and predict how the guide reacts to the user, this work proposes a role-based contextual model by modeling role-specific contexts differently for improving system performance.

2 Proposed Approach

The model architecture is illustrated in Figure 2. First, the previous utterances are fed into the contextual model to encode into the history summary, and then the summary vector and the current utterance are integrated for helping LU and dialogue policy learning. The whole model is trained in an end-to-end fashion, where the history summary vector is automatically learned based on two downstream tasks. The objective of the proposed model is to optimize the conditional probability $p(\hat{\mathbf{y}} \mid \mathbf{x})$, so that the difference between the predicted distribution $q(\hat{y_k} = z \mid \mathbf{x})$ and the target distribution $q(y_k = z \mid \mathbf{x})$ can be minimized:

$$\mathcal{L} = -\sum_{k=1}^{K}\sum_{z=1}^{N} q(y_k = z \mid \mathbf{x}) \log p(\hat{y_k} = z \mid \mathbf{x}), \quad (1)$$

where the labels $\mathbf{y}$ can be either intent tags for understanding or system actions for dialogue policy learning.

Language Understanding (LU) Given the current utterance $\mathbf{x} = \{w_t\}_1^T$, the goal is to predict the user intents of $\mathbf{x}$, which includes the speech acts and associated attributes shown in Figure 1; for example, `QST_WHAT` is composed of the speech act `QST` and the associated attribute `WHAT`. Note that we do not process the slot filling task for extracting `LOC`. We apply a bidirectional long short-term memory (BLSTM) model (Schuster and Paliwal, 1997) to integrate preceding and following words to learn the probability distribution of the user intents.

$$\mathbf{v}_{\text{cur}} = \text{BLSTM}(\mathbf{x}, W_{\text{his}} \cdot \mathbf{v}_{\text{his}}), \quad (2)$$

$$\mathbf{o} = \texttt{sigmoid}(W_{\text{LU}} \cdot \mathbf{v}_{\text{cur}}), \quad (3)$$

where W_{his} is a dense matrix and $\mathbf{v}_{\text{his}}$ is the history summary vector, $\mathbf{v}_{\text{cur}}$ is the context-aware vector of the current utterance encoded by the BLSTM, and $\mathbf{o}$ is the intent distribution. Note that this is a multi-label and multi-class classification, so the `sigmoid` function is employed for modeling the distribution after a dense layer. The user intent labels $\mathbf{y}$ are decided based on whether the value is higher than a threshold θ.

Dialogue Policy Learning For system action prediction, we also perform similar multi-label

[2] http://www.colips.org/workshop/dstc4/

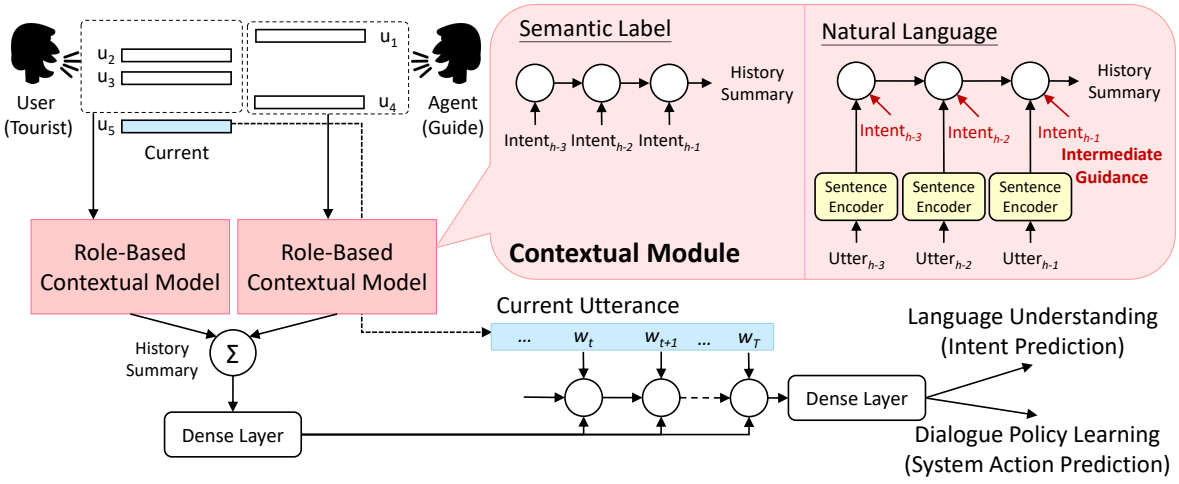

Figure 2: Illustration of the proposed role-based contextual model.

multi-class classification on the context-aware vector $\mathbf{v}_{\text{cur}}$ from (2) using `sigmoid`:

$$\mathbf{o} = \texttt{sigmoid}(W_\pi \cdot \mathbf{v}_{\text{cur}}), \tag{4}$$

and then the system actions can be decided based on a threshold θ.

2.1 Contextual Module

In order to leverage the contextual information, we utilize two types of contexts: 1) semantic labels and 2) natural language, to learn history summary representations, $\mathbf{v}_{\text{his}}$ in (2). The illustration is shown in the top-right part of Figure 2.

Semantic Label Given a sequence of annotated intent tags and associated attributes for each history utterance, we employ a BLSTM to model the explicit semantics:

$$\mathbf{v}_{\text{his}} = \text{BLSTM}(\text{intent}_t), \tag{5}$$

where intent_t is the vector after one-hot encoding for representing the annotated intent and the attribute features. Note that this model requires the ground truth annotations of history utterances for training and testing.

Natural Language (NL) Given the natural language history, a sentence encoder is applied to learn a vector representation for each prior utterance. After encoding, the feature vectors are fed into a BLSTM to capture temporal information:

$$\mathbf{v}_{\text{his}} = \text{BLSTM}(\text{CNN}(\text{utt}_t)), \tag{6}$$

where the CNN is good at extracting the most salient features that can represent the given natural language utterances illustrated in Figure 3. Here the sentence encoder can be replaced into different encoders[3], and the weights of all encoders are tied together.

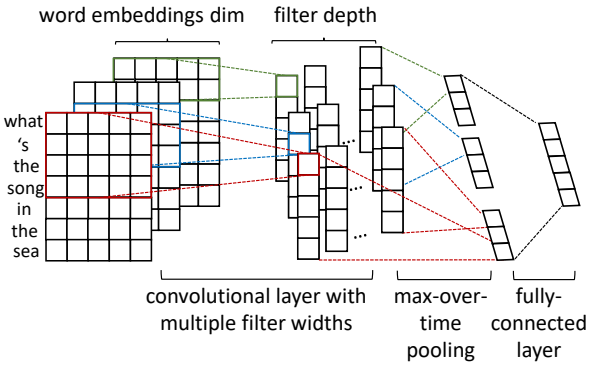

Figure 3: Illustration of the CNN sentence encoder for the example sentence "*what's the song in the sea*".

NL with Intermediate Guidance Considering that the semantic labels may provide rich cues, the middle supervision signal is utilized as intermediate guidance for the sentence encoding module in order to guide them to project from input utterances to a more meaningful feature space. Specifically, for each utterance, we compute the cross entropy loss between the encoder outputs and corresponding intent-attributes shown in Figure 2. Assuming that l_t is the encoding loss for utt_t in the history, the final objective is to minimize $(\mathcal{L} + \sum_t l_t)$. This model does not require the

[3] In the experiments, CNN achieved slightly better performance with fewer parameters compared with BLSTM.

ground truth semantics for history when testing, so that it is more practical compared to the above model using semantic labels.

2.2 Speaker Role Modeling

In a dialogue, there are at least two roles communicating with each other, each individual has his/her own goal and speaking habit. For example, the tourists have their own desired touring goals and the guides are try to provide the sufficient touring information for suggestions and assistance. Prior work usually ignored the speaker role information or only modeled a single speaker's history for various tasks (Chen et al., 2016c; Yang et al., 2017). The performance may be degraded due to the possibly unstable and noisy input feature space. To address this issue, this work proposes the role-based contextual model: instead of using only a single BLSTM model for the history, we construct one individual contextual module for each speaker role. Each role-dependent recurrent unit $\text{BLSTM}_{\text{role}_x}$ receives corresponding inputs $\text{x}_{i,\text{role}_x}$ $(i = [1, ..., N])$, which have been processed by an encoder model, we can rewrite (5) and (6) into (7) and (8) respectively:

$$\begin{aligned} \mathbf{v}_{\text{his}} &= \text{BLSTM}_{\text{role}_a}(\text{intent}_{t,\text{role}_a}) \quad (7) \\ &+ \text{BLSTM}_{\text{role}_b}(\text{intent}_{t,\text{role}_b}). \\ \mathbf{v}_{\text{his}} &= \text{BLSTM}_{\text{role}_a}(\text{CNN}(\text{utt}_{t,\text{role}_a})) \quad (8) \\ &+ \text{BLSTM}_{\text{role}_b}(\text{CNN}(\text{utt}_{t,\text{role}_b})) \end{aligned}$$

Therefore, each role-based contextual module focuses on modeling the role-dependent goal and speaking style, and $\mathbf{v}_{\text{cur}}$ from (2) is able to carry role-based contextual information.

3 Experiments

To evaluate the effectiveness of the proposed model, we conduct the LU and dialogue policy learning experiments on human-human conversational data.

3.1 Setup

The experiments are conducted on DSTC4, which consists of 35 dialogue sessions on touristic information for Singapore collected from Skype calls between 3 tour guides and 35 tourists (Kim et al., 2016). All recorded dialogues with the total length of 21 hours have been manually transcribed and annotated with speech acts and semantic labels at each turn level. The speaker labels are also annotated. Human-human dialogues contain rich and complex human behaviors and bring much difficulty to all dialogue-related tasks. Given the fact that different speaker roles behave differently, DSTC4 is a suitable benchmark dataset for evaluation.

We choose a mini-batch `adam` as the optimizer with the batch size of 128 examples (Kingma and Ba, 2014). The size of each hidden recurrent layer is 128. We use pre-trained 200-dimensional word embeddings $GloVe$ (Pennington et al., 2014). We only apply 30 training epochs without any early stop approach. The sentence encoder is implemented using a CNN with the filters of size $[2, 3, 4]$, 128 filters each size, and max pooling over time. The idea is to capture the most important feature (the highest value) for each feature map. This pooling scheme naturally deals with variable sentence lengths. Please refer to Kim (2014) for more details.

For both tasks, we focus on predicting multiple labels including speech acts and attributes, so the evaluation metric is average F1 score for balancing recall and precision in each utterance. Note that the final prediction may contain multiple labels.

3.2 Results

The experiments are shown in Table 1, where we report the average number over five runs. The first baseline (row (a)) is the best participant of DSTC4 in IWSDS 2016 (Kim et al., 2016), the poor performance is probably because tourist intents are much more difficult than guide intents (most systems achieved higher than 60% of F1 for guide intents but lower than 50% for tourist intents). The second baseline (row (b)) models the current utterance without contexts, performing 62.6% for understanding and 63.4% for policy learning.

3.2.1 Language Understanding Results

With contextual history, using ground truth semantic labels for learning history summary vectors greatly improves the performance to 68.2% (row (c)), while using natural language slightly improves the performance to 64.2% (row (e)). The reason may be that NL utterances contain more noises and the contextual vectors are more difficult to model for LU. The proposed role-based contextual models applying on semantic labels and NL achieve 69.2% (row (d)) and 65.1% (row (f)) on F1 respectively, showing the significant improvement all model without role modeling. Fur-

Model			Language Understanding	Policy Learning
Baseline	(a)	*DSTC4-Best*	52.1	-
	(b)	BLSTM	62.6	63.4
Contextual-Sem	(c)	BLSTM	68.2	66.8
	(d)	+ Role-Based	**69.2**†	**70.1**†
Contextual-NL	(e)	BLSTM	64.2	66.3
	(f)	+ Role-Based	65.1†	66.9†
	(g)	+ Role-Based w/ Intermediate Guidance	**65.8**†	**67.4**†

Table 1: Language understanding and dialogue policy learning performance of F-measure on DSTC4 (%). † indicates the significant improvement compared to all methods without speaker role modeling.

thermore, adding the intermediate guidance acquires additional improvement (65.8% from the row (g)). It is shown that the semantic labels successfully guide the sentence encoder to obtain better sentence-level representations, and then the history summary vector carrying more accurate semantics gives better performance for understanding.

3.2.2 Dialogue Policy Learning Results

To predict the guide's next actions, the baseline utilizes intent tags of the current utterance without contexts (row (b)). Table 1 shows the similar trend as LU results, where applying either role-based contextual models or intermediate guidance brings advantages for both semantics-encoded and NL-encoded history.

3.3 Discussion

In contrast to NL, semantic labels (intent-attribute pairs) can be seen as more explicit and concise information for modeling the history, which indeed gains more in our experiments for both LU and dialogue policy learning. The results of Contextual-Sem can be treated as the upper bound performance, because they utilizes the ground truth semantics of contexts. Among the experiments of Contextual-NL, which are more practical because the annotated semantics are not required during testing, the proposed approaches achieve 5.1% and 6.3% relative improvement compared to the baseline for LU and dialogue policy learning respectively.

Between LU and dialogue policy learning tasks, most LU results are worse than dialogue policy learning results. The reason probably is that the guide has similar behavior patterns such as providing information and confirming questions etc., while the user can have more diverse interactions. Therefore, understanding the user intents is slightly harder than predicting the guide policy in the DSTC4 dataset.

With the promising improvement for both LU and dialogue policy learning, the idea about modeling speaker role information can be further extended to various research topics in the future.

4 Conclusion

This paper proposes an end-to-end role-based contextual model that automatically learns speaker-specific contextual encoding. Experiments on a benchmark multi-domain human-human dialogue dataset show that our role-based model achieves impressive improvement in language understanding and dialogue policy learning, demonstrating that different speaker roles behave differently and focus on different goals.

Acknowledgements

We would like to thank reviewers for their insightful comments on the paper. The authors are supported by the Ministry of Science and Technology of Taiwan, Google Research, Microsoft Research, and MediaTek Inc..

References

Anshuman Bhargava, Asli Celikyilmaz, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 8337–8341.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *ICLR*.

Po-Chun Chen, Ta-Chung Chi, Shang-Yu Su, and Yun-Nung Chen. 2017. Dynamic time-aware attention

to speaker roles and contexts for spoken language understanding. In *Proceedings of 2017 IEEE Workshop on Automatic Speech Recognition and Understanding*.

Yun-Nung Chen, Dilek Hakanni-Tür, Gokhan Tur, Asli Celikyilmaz, Jianfeng Guo, and Li Deng. 2016a. Syntax or semantics? knowledge-guided joint semantic frame parsing. In *Proceedings of 2016 IEEE Spoken Language Technology Workshop*. IEEE, pages 348–355.

Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng. 2016b. Knowledge as a teacher: Knowledge-guided structural attention networks. *arXiv preprint arXiv:1609.03286* .

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016c. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*. pages 3245–3249.

Yun-Nung Chen, Ming Sun, Alexander I. Rudnicky, and Anatole Gershman. 2015. Leveraging behavioral patterns of mobile applications for personalized spoken language understanding. In *Proceedings of 17th ACM International Conference on Multimodal Interaction*. ACM, pages 83–86.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*. pages 715–719.

Seokhwan Kim, Luis Fernando DHaro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. 2016. The fourth dialog state tracking challenge. In *Proceedings of the 7th International Workshop on Spoken Dialogue Systems*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Xuijun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of The 8th International Joint Conference on Natural Language Processing*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. volume 14, pages 1532–1543.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Yangyang Shi, Kaisheng Yao, Hu Chen, Yi-Cheng Pan, Mei-Yuh Hwang, and Baolin Peng. 2015. Contextual spoken language understanding using recurrent neural networks. In *Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 5271–5275.

Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky. 2016. An intelligent assistant for high-level task understanding. In *Proceedings of The 21st Annual Meeting of the Intelligent Interfaces Community*. pages 169–174.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Zhangyang Wang, Yingzhen Yang, Shiyu Chang, Qing Ling, and Thomas S Huang. 2016. Learning a deep l encoder for hashing. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*. pages 2174–2180.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*.

Jason Weston, Sumit Chopra, and Antoine Bordesa. 2015. Memory networks. In *Proceedings of International Conference on Learning Representations*.

Puyang Xu and Ruhi Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 136–140.

Xuesong Yang, Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng. 2017. End-to-end joint learning of natural language understanding and dialogue manager. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 5690–5694.