Contextual and Structural Language Understanding in Dialogues

Yun-Nung Chen Dilek Hakkani-Tür Gokhan Tur
Jianfeng Gao Asli Celikyilmaz Li Deng
National Taiwan University, Taiwan
Google Research, Mountain View, CA
Microsoft Research, Redmond, WA
yvchen@ieee.org

Abstract

Natural language understanding (NLU) is a core component of a dialogue system. Recently recurrent neural networks (RNN) obtained strong results on NLU due to their superior ability of preserving sequential information over time. Traditionally, the NLU module ignores the contexts of utterances and tags semantic slots for utterances considering their flat structures, as the underlying RNN structure is a linear chain. However, contexts and linguistic properties provide informative cues for better understanding. This paper introduces a novel model, a generalization of RNN to additionally incorporate 1) contextual utterances and 2) non-flat network topologies guided by prior knowledge. The model automatically figures out the salient and accurate contexts and substructures that are essential to predict the semantic tags of the given sentences, resulting in better understanding. The experiments showed the significant improvement on contextual and structural language understanding scenarios.

1 Introduction

Goal-oriented spoken dialogue systems are being incorporated in various devices and allow users to speak to systems freely in order to finish tasks more efficiently. A key component of these conversational systems is the natural language understanding (NLU) module-it refers to the targeted understanding of human speech directed at machines (Tur and De Mori, 2011). A typical NLU first decides the domain of user's request given the input utterance, and based on the domain, predicts the intent and fills associated slots corresponding to a domain-specific semantic template (Tur and De Mori, 2011). For example, a user utterance, "*show me the flights from seattle to san francisco*" can be formulated as a semantic frame, find_flight(origin="seattle", dest="san francisco"). Traditionally, slot filling is framed as a word sequence tagging task, where the IOB (in-out-begin) format is applied for representing slot tags (Pieraccini et al., 1992; Wang et al., 2005). With the advances on deep learning, Yao et al. (2013) and Mesnil et al. (2015) employed RNNs for sequence labeling in order to perform slot filling. However, the above studies focused on single-turn understanding and ignored the linguistic properties when tagging sequences.

In order to address the issues and better learn the sequence tagging models, this paper proposes a generalization of RNNs that automatically learn how to incorporate contextual and structural attention for generating sentence-based representations specifically for modeling sequence tagging.

2 Proposed Model

Given an utterance with a sequence of words/tokens $\vec{s} = w_1, ..., w_T$, our NLU model is to predict corresponding semantic tags $\vec{y} = y_1, ..., y_T$ for each word/token by incorporating 1) history sentences and 2) knowledge-guided structures for contextual and structural language understanding respectively. The proposed model is illustrated in Figure 1. The knowledge encoding module first leverages contextual or structural knowledge $\{x_i\}$ to generate a set of encoded knowledge representations. The model learns the representation for the whole sentence by paying different attention on the encoded knowledge stored in the memory. Then the learned vector encoding the contexts or structures is used for improving the semantic tagger.

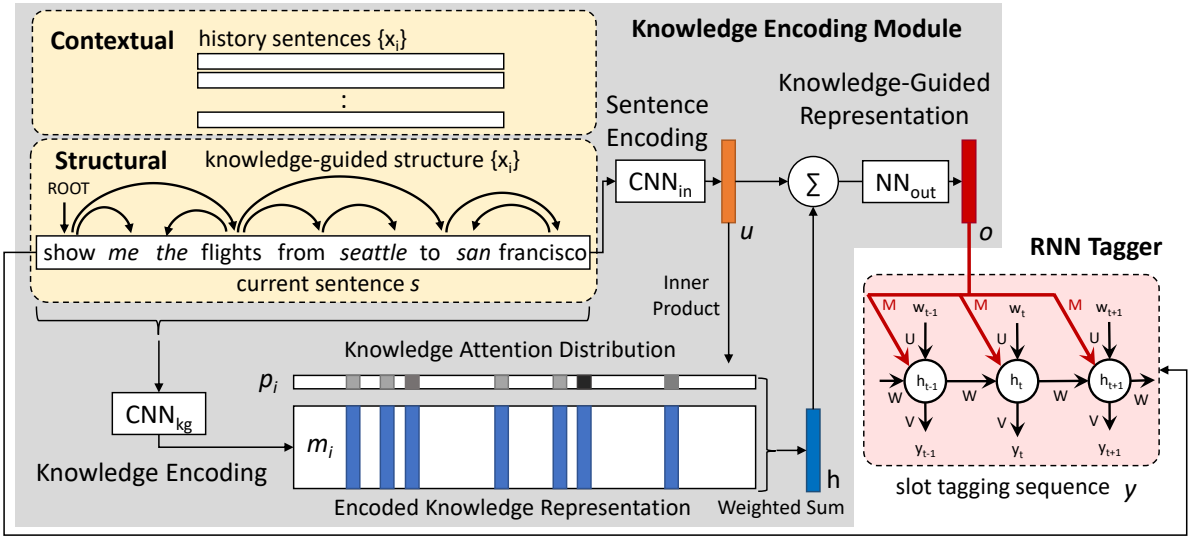

Figure 1: The illustration of contextual and structural attention networks for NLU.

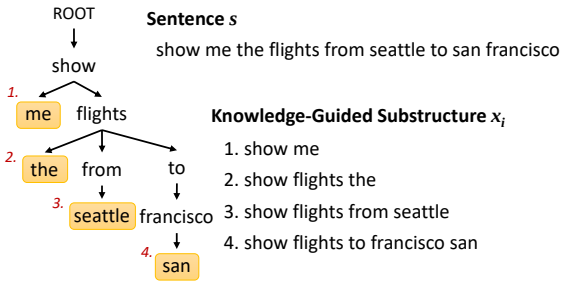

Figure 2: The illustration of substructures obtained from the utterance "*show me the flights from seattle to san francisco*".

2.1 Knowledge Encoding Module

The top-left component of Figure 1 illustrates the module for modeling contexts and structures, contextual language understanding and structural language understanding described below.

Contextual LU To store the knowledge in the previous turns, we convert each utterance from previous turns, x_i, into a memory vector m_i by embedding the utterances in a continuous space through a CNN (Chen et al., 2016b), which is motivated by the memory networks (Weston et al., 2015). Therefore, the contextual LU module would capture the most salient parts in the history through an attention mechanism to better understand the current utterance.

Structural LU The prior knowledge obtained from external resources, such as dependency relations, knowledge bases, etc., provides richer information to help decide the semantic tags given an input utterance. This paper takes dependency relations for knowledge encoding, and other structured relations can be applied in the same way. The input utterance is parsed by a dependency parser, and the substructures are built according to the paths from the root to all leaves (Chen and Manning, 2014). For example, the dependency parsing of the utterance "*show me the flights from seattle to san francisco*" is shown in Figure 1, where the associated substructures are obtained from the parsing tree for knowledge encoding (Chen et al., 2016a) illustrated in Figure 2. Each substructure is also encoded into a knowledge vector via a CNN in order to help understanding.

2.2 End-to-End Training

The model embeds all contextual and structural knowledge into a continuous space and stores embeddings of all x's in the knowledge memory. The representation of the input utterance is then compared with encoded knowledge representations to integrate the salient history and carried structures guided by knowledge via an attention mechanism. Then the knowledge-guided representation of the sentence is taken together with the word sequence for estimating the semantic tags. The whole model can be trained in an end-to-end manner by maximizing the probability of output slots $\vec{y}$ given the sentence $\vec{s}$ and automatically learned knowledge-guided representation o: $\vec{y} = \text{RNN}(o, \vec{s})$. Specifically, given only the current utterance and associated slot tags along with their prior sentences or

Model (Klg.)	Result		
Contextual	**First-Turn**	**Other**	**Overall**
- RNN (✗)	57.6	56.0	56.3
- RNN (✓)	69.9	60.8	62.5
- Proposed (✓)	**73.8**†	**66.5**†	**68.0**†
Structural	**Small**	**Medium**	**Large**
- RNN (✗)	68.58	84.55	92.97
- CNN (✗)	73.57	85.52	93.88
- DCNN (✓)	70.24	83.80	93.25
- Tree-RNN (✓)	73.50	83.92	92.28
- Proposed (✓)	**74.60**†	**87.99**†	**94.86**†

Table 1: The F1 scores of predicted slots. Small: 1/40 set; Medium: 1/10 set; Large: original set. († indicates significant improvement over baselines with $p < 0.05$ in the t-test.)

the dependency relations, the model can automatically decide the knowledge encoding network, CNN_{kg}, the sentence encoding network, CNN_{in}, the attention weights, the knowledge-guided representation network, NN_{out}, and the final RNN tagger.

3 Experiments

The proposed model is evaluated in the 1) contextual understanding scenario using multi-turn data from the Microsoft Cortana and the 2) structural language understanding scenario using the benchmark ATIS dataset. Table 1 shows the experimental results, where the proposed model outperforms all baselines for both contextual and structural language understanding.

4 Conclusion

This paper proposes a novel model that leverages contexts and structures to improve natural language understanding, which can automatically figure out the salient and accurate contexts and substructures that are useful to predict the semantic tags of the current sentence. The experiments show benefits and effectiveness of the proposed model on both contextual and structural language understanding tasks, where the salient history sentences and substructures result in promising results on the Microsoft Cortana data and the benchmark ATIS dataset.

References

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*. pages 740–750.

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

Yun-Nung Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016b. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of Interspeech*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(3):530–539.

Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, Jean-Luc Gauvain, Esther Levin, Chin-Hui Lee, and Jay G Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, volume 1, pages 193–196.

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

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine* 22(5):16–31.

Jason Weston, Sumit Chopra, and Antoine Bordesa. 2015. Memory networks. In *International Conference on Learning Representations (ICLR)*.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding. In *INTERSPEECH*. pages 2524–2528.