CONTEXT MEMORY NETWORKS FOR MULTI-OBJECTIVE SEMANTIC PARSING IN CONVERSATIONAL UNDERSTANDING

Asli Celikyilmaz, Dilek Hakkani-Tur, Gokhan Tur, Yun-Nung Chen, Bin Cao, Ye-Yi Wang

Microsoft

ABSTRACT

The end-to-end multi-domain and multi-task learning of the full semantic frame of user utterances (i.e., domain and intent classes and slots in utterances) have recently emerged as a new paradigm in spoken language understanding. An advantage of the joint optimization of these semantic frames is that the data and feature representations learnt by the model are shared across different tasks (e.g., domain/intent classification and slot filling tasks use the same feature sets). It's important that the model should learn to pay attention to global and local aspects of the utterances while learning to map the entire utterance to an intent class and tag each word with a slot tag. We introduce the Context Memory Network (CMN), a neural network architecture which specifically focuses on learning better representations as attention vectors from past memory to be reasoned with for the end task of jointly learning the intent class and slot tags. The utterances trigger a dynamic memory network, which learns attention based representation for each word by allowing the model to condition on the list of related phrases in the form of memory networks. These representations are then provided to a new multi-objective long short term memory network (LSTM) to infer the intent class and slot tags. Our empirical investigations on CMN show impressive gains over the end-to-end LSTM baselines on ATIS dataset as well as two other human-to-machine conversational datasets.

Index Terms— recurrent neural networks, long-short term memory networks (LSTM), attention, embedding, memory networks, spoken language understanding.

1. INTRODUCTION

The Spoken Language Understanding (SLU) in conversational dialog systems parses user utterances into corresponding semantic concepts. These concepts include the domain/intent classes of the utterances as well as slot tags of words/phrases. Lets consider the following sentence as our running sample:

$$\textit{how many metals did } \overbrace{\textit{us}}^{\texttt{country}} \textit{ win in the } \overbrace{\textit{2012}}^{\texttt{year}} \overbrace{\textit{London olympics}}^{\texttt{event}} \textit{?}$$

Common approaches to SLU [1] usually build two utterance classification models, a domain classifier to map this utterance to "*sports-olympics*" domain and an intent classifier to map the utterance to "*find-score*" intent, while a separate sequence learning model builds a slot tagger to tag the words to corresponding slots as shown in the running example. It is only natural to learn to learn to infer these semantic concepts jointly as the data and features of the classifiers and sequence taggers are in the same nature. Multi-task learning approaches [] can learn to infer about different aspects of the utterances. This is what the recent approaches to SLU have proposed. Common approaches use recurrent neural network (RNN) architectures to integrate the three SLU tasks for domain detection, intent detection and slot filling in a single SLU model (REF). They either use a standard RNN architecture, e.g, (REF uses LSTM) or sequence to sequence learning methods (REF bowen) to learn the full semantic frame. Although these end-to-end SLU systems have been quite effective, in this paper, we test the hypothesis that better representations can be obtained by incorporating the contextual knowledge about words of an utterance in the model architecture. The intuition underlying our approach is that while intent and domain tags learn to optimize with features that indicate the global aspects of the conversation, the slot tagging models focus on local aspects and mainly discover contextual features. A network that can learn to attend these aspects would yield better representations that will eventually improve the performance of the semantic parsing in SLU. This has not been thoroughly investigated in these recent end-to-end learning approaches.

In this paper, we focus on learning better representations for words, the basic units in representing the utterances, using contextual information about the words that the model should learn the attend to while reasoning about the slots and intents separately. We propose the Context Memory Network (CMN), a neural network based framework for semantic frame parsing task for SLU that is trained using the user utterances, the input word-context as inputs and semantic frame tags (e.g, intent and slot tags) as outputs. We extend the recent Memory Networks (REF) to learn representations for words through an attention process that itself learn to pay more attention to specific context given the utterance. The CMN memory module then retrieves facts and provides vector representation of all the relevant information to map the utterance to the correct intent as well as each word to a slot tag.

In the next section, we will provide background to end-to-end SLU models with RNN architectures as well as memory networks. In Section XX, we will extend these architectures to present the details of CMN for the SLU task. In the experiment section, we will investigate the performance of the CMN in comparison the baseline RNN architectures.

2. END-TO-END DEEP LEARNING FOR SLU

The end-to-end learning (or joint learning) of the SLU semantic frames taking into account the dependencies of domain and intent classification and slot tagging tasks has been investigated in the past. One of the first to these approaches are the triangular chain conditional random fields (Tri-CRF), which was introduced by [2]. It can jointly learn two of the SLU's components (intents and slots) in a single pass, where their dependencies are exploited. Later, [3] used a hierarchical end-to-end Bayesian learning approach to jointly learn the domain, intent and slot models with only relying on the n-gram dependencies of the domain and intent classes as well as slot tags as priors. Extending the earlier Tri-CRF work of [2], rather than providing manual features, [4] proposed to automatically learn the features through a convolutional neural networks (CNN).

With the advances of deep learning, more recent approaches has shifted the focus for SLU semantic frame tagging to more sophisticated methods such as RNNs. Among several research, for instance [5, 6] employ RNNs with different architectures for slot filling tasks. A comprehensive review of the RNN based SLU semantic frame parsing is presented in [7]. Only recently [8] propose a holistic end-to-end joint modeling of the domain, intent and slot tags for the SLU semantic frame identification in a single LSTM architecture as shown in Figure 3. The input layer is represented as one-hot or uses pre-trained embeddings. To consider contextual features, the left and right context of the current word is combined to form the current input representation (as shown in dashed lines between input-embedding layer). The LSTM layer is comprised of LSTM cells forming the hidden units. After the activation layer is applied, the output layer predicts a distribution over the entire semantic frame $|S+I+D|$ where S, I and D indicate number of slot, intent and domains found in the corpus which forms the schema. the model can predict any of these tags as possible output tags as follows:

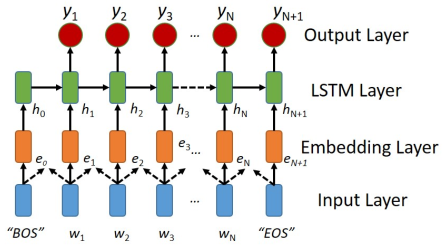

Fig. 1. RNN-LSTM architecture [8] for end-to-end SLU semantic frame learning task.

$$y_i = argmax_{k \in Z^{|S+I+D|}} p(y_i = k | f(h_i)) \qquad (1)$$

where $f(h_i)$ is the activation function which is normalized to obtain posterior probability. The last word (i.e., ”EOS”) can encode domain or intent tag. If the training data is small, a viterbi layer on top of these posteriors can handle unexpected slot tag sequences. They proposed different architectures that can generalize well to the complete semantic frame tagging. [9] use a similar approach but focus on learning representations from previous turn in a dialog via Memory Network architecture and jointly learn to predict the intent and slot tags of the current turn utterance.

Extending these latest approaches, in this paper, we propose a modular approach to end-to-end learning and present two LSTM architectures to enable joint learning as a multi-task learning architecture. The context memory network, which learns the representation of each word in an utterance, memorizes the relationship of each word to its context in the entire corpus and learns an attention vector indicating which parts of the context the model should attend to while learning to predict the slot tags. The CMN then combines the sentence representation with the each of its word's memory attention vector as input representation to a multi-objective LSTM layer. We introduce two separate objective terms for intent and slots. Decoupling enables to learn different aspects jointly and the model expert can learn to balance the objective functions. By context-dependent memory we refer to improved recall of the information specific to certain slots and intents by way of presenting the context at encoding and decoding time. Because all modules in the CMN communicate over vector representations and various types of differentiable and deep neural networks with gates, the entire CMN model can be trained via backpropagation and gradient descent.

3. CONTEXT MEMORY NETWORKS

Attention mechanisms are a quite new phenomena and we are going to provide some background on them in this section.

Attention mechanisms in NNs allow the network to focus only on a certain subset of the data provided for a given task. Being able to distinguish between the necessary information at a specific step of a task further reduces the amount of information that has to be processed. The idea behind attention mechanisms is motivated by observing the visual attention of humans. Despite processing the visual input all at the same time, humans rather pay attention to small regions one after the other of for example a picture. This allows to keep the amount of information to be manageable.

For utterance understanding tasks, e.g. the intent detection and slot tagging, when RNNs are used, the problem arises that we usually rely on one final embedding of an utterance in order predict its class, e.g, its intent. Similarly, we rely on the current word's embedding (along with its surrounding words in the utterance) to predict the correct slot tag. It would be more informative for the model if it could learn to attend to certain specific words in the input sentence to represent the sentence before the activation function of the word or the sentence is applied.

These correspondences are learned by keeping the embeddings of individual words and attend to them through learning the appropriate weights. Here, an attention mechanism allows to explicitly see where the algorithm is looking at before predicting the distribution of the tags of a word or a sentence.

4. MULTI-OBJECTIVE SEMANTIC PARSING FOR SLU

4.1. Single-Layer Multi-Objective LSTM for SLU

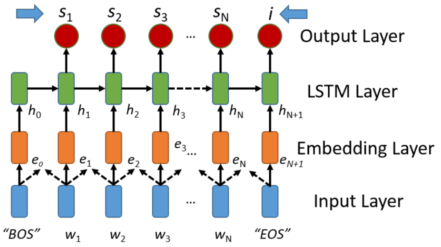

Fig. 2. RNN-LSTM architecture with multi-objective function for end-to-end SLU semantic frame learning task.

4.2. Multi-Layer Multi-Objective LSTM for SLU

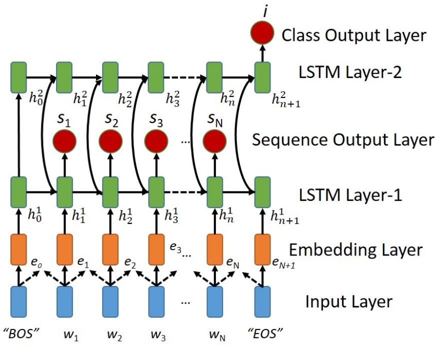

Fig. 3. RNN-LSTM architecture with multi-objective function for end-to-end SLU semantic frame learning task.

5. EXPERIMENTS

6. DISCUSSIONS

7. CONCLUSION AND FUTURE WORK

8. REFERENCES

[1] G. Tur and R. D. Mori, "Spoken language understanding: Systems for extracting semantic information from speech," New York, NY: John Wiley and Sons, 2011.

[2] M. Jeong and G.G. Lee, "Triangular-chain conditional random fields," .

[3] A. Celikyilmaz and D. Hakkani-Tur, "A joint model of discovery of aspects in utterances," .

[4] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling," .

[5] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning method for spoken language understanding," .

[6] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural network for language understanding," .

[7] K. Yao Y. Bengio L. Deng D. Hakkani-Tur X. He L. Heck G. Tur D. Yu G. Mesnil, Y. Dauphin and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding," .

[8] D. Hakkani-Tur, G. Tur, A. Celikyilmaz, Y-N Chen, J. Gao, L. Deng, and Y-Y-Y Wang, "Multi-domain joint semantic frame parsing using bi-directional rnn-lstm," .

[9] Y-N Chen, D. Hakkani-Tur, G. Tur, J. Gao, and L. Deng, "End-to-end memory networks with knowledge carryover for multi=turn spoken language understanding," .