Adaptive Global-Local Context Fusion for Multi-Turn Spoken Language Understanding

Thanh Tran[*], **Kai Wei**[*], **Weitong Ruan, Ross McGowan, Nathan Susanj, Grant P. Strimel**
Amazon Alexa
{tdt, kaiwe, weiton, rosmcgow, nsusanj, gsstrime}@amazon.com

Abstract

Recent years have seen significant advances in multi-turn Spoken Language Understanding (SLU), where dialogue contexts are used to guide intent classification and slot filling. However, how to selectively incorporate dialogue contexts, such as previous utterances and dialogue acts, into multi-turn SLU still remains a substantial challenge. In this work, we propose a novel contextual SLU model for multi-turn intent classification and slot filling tasks. We introduce an adaptive global-local context fusion mechanism to selectively integrate dialogue contexts into our model. The local context fusion aligns each dialogue context using multi-head attention, while the global context fusion measures overall context contribution to intent classification and slot filling tasks. Experiments show that on two benchmark datasets, our model achieves absolute F1 score improvements of 2.73% and 2.57% for the slot filling task on Sim-R and Sim-M datasets, respectively. Additional experiments on a large-scale, de-identified, in-house dataset further verify the measurable accuracy gains of our proposed model.

Introduction

The last few years have seen an increasing application of Spoken Language Understanding (SLU) systems, such as Google Assistant, Amazon Alexa, etc. One of the fundamental tasks of these systems is to map the meaning of spoken utterances expressed in natural language to machine comprehensible language (Allen 1995; Tur and De Mori 2011). For example, the machine learns to map *find a restaurant in Richmond* to an intent for finding restaurants (intent classification) and to slots such as Richmond: Location (slot filling).

One important topic in the SLU research is effectively interpreting a user's intents in multi-turn dialogues, where the user and the system have multiple turns of back-and-forth conversations to achieve the user's goal. Historically, this line of work has focused on using traditional machine learning methods (Miller et al. 1996; Bhargava et al. 2013). For example, Bhargava et. al. (2013) found that using previous utterances as contexts in an SVM-HMM SLU system could help resolve ambiguities. Recently, deep learning approaches have become increasingly popular to incorporate contextual information (Qin et al. 2021; Su, Yuan, and Chen 2019; Abro et al. 2019; Chen et al. 2019; Su, Yuan, and Chen 2018; Gupta, Rastogi, and Hakkani-Tur 2018; Chen et al. 2016; Wei et al. 2021). Chen et. al. (2016) proposed to use end-to-end memory networks to model previous utterance transcripts in multi-turn dialogues. Gupta et al. (2018) proposed an efficient method to encode dialogue acts with a feedforward network from prior dialogue history with limited degradation in accuracy compared to the end-to-end memory network approach (Gupta, Rastogi, and Hakkani-Tur 2018). Gupta et. al. (2019) fuses signals like previous intents via a self-attention mechanism with a variable context window. Wang et. al. (2019) encodes historical utterances using the Bidirectional Long Short Term Memory (BiLSTM) networks and ConceptNet to encode external knowledge, and construct knowledge attention over these contexts. Qin et. al. (2021) proposed a context-aware graph convolutional network for contextual SLU. Yet, how to *selectively* incorporate both dialogue acts and previous utterance context to multi-turn intent detection and slot filling still remains under-explored.

In this paper, we propose a contextual SLU model for intent classification and slot filling in multi-turn dialogues, where both dialogue acts and previous utterance contexts are exploited. To selectively incorporate dialogue contexts into the model, we propose an adaptive global-local context fusion mechanism. The local context fusion aligns each contextual source information with the utterance transcript signals using the multi-head attention (Vaswani et al. 2017), while the global context measures contribution of all contextual information. The closest work to ours are (Gupta, Rastogi, and Hakkani-Tur 2018) and (Qin et al. 2021). However, these works use BiLSTM to encode previous utterances, whereas our work uses BERT to enrich their contextually semantic representations. Moreover, Qin et. al. (2021) focuses on graph-based methods to filter out irrelevant information only for slot filling, whereas our work uses the global-local multi-head attention for both slot filling and intent detection. Our model achieves the SOTA results on intent classification and outperforms previous methods for slot filling by a large margin on two benchmark datasets. We further experiment with our model on a large scale, in-house, de-identified datatset. In addition, we study the effects of contexts by conducting ablation studies and visualizing the

[*]Equal contribution

global attention weights to demonstrate the effectiveness of our proposed design.

Problem Statement

Our contextual SLU model takes a current utterance u_t, and a list of previous dialogue acts $\mathcal{D}^t$=$\{(a_1, s_1)$, ..., $(a_{|\mathcal{D}_t|}, s_{|\mathcal{D}_t|})\}$, and previous utterance transcripts $\mathcal{U}^t$=$\{u_1, u_2, ..., u_{t-1}\}$. Each (a_i, s_i) pair indicates a dialogue action a_i and a dialogue slot s_i. Given the ground truth intent y_t^{int} of u_t and ground truth slot $y_{t,i}^{slot}$ per each word token $p_{t,i} \in u_t$, our contextual SLU model aims to maximize the intent probability $P(y_t^{int}|u_t, \mathcal{U}^t, \mathcal{D}^t)$ for u_t, and the slot probability $P(y_t^{slot_i}|\, u_t, \mathcal{U}^t, \mathcal{D}^t)$ for each $p_{t,i}$.

Proposed Model

Figure 1 shows our proposed contextual SLU model architecture. At a high level, we input (i) the wordpiece embeddings of current utterance transcripts and (ii) the context encoder of dialogue acts and previous utterance transcripts into our adaptive global-local context fusion mechanism.

The local context fusion aligns each contextual source information with the utterance transcript signals using the multi-head attention (Vaswani et al. 2017), while the global context measures the contribution of all contextual information. The local context fusion considers each context encoding type as a separated key and value and the wordpiece embeddings as the query. Then, it assigns attention scores to the context encoding. Intuitively, the global attention serves as a gating layer to produce how much all contexts can contribute to the input query. Without the global attention, the local attention always give an accumulated attention score of 1.0, as a result of performing a softmax function. This is not optimal. In many cases, the contexts contribute insignificantly to the SLU tasks. For example, a user asks a voice assistant system to *call uncle sam*, and the system confirms back to see if the user wants to *call a nearby uncle sam's sandwich bar*. The user then says in a second turn *call my uncle who's first name is sam*. In this case, dialogue contexts coming from the first turn are not helpful for the interpretation of the second turn. Therefore, we propose this global-local fusion mechanism to allow the model to selectively pay attention to previous dialogue contexts in multi-turn dialogues.

After selectively fusing contextual information with the wordpiece embeddings, we use a BiLSTM encoder to learn context-aware embeddings, the output of which are used for the intent classification and slot filling tasks, simultaneously. We detail our architecture as below.

Embedding Layer

We pre-train a SentencePiece (SP)[1] model on the training data with 4,500 wordpieces to avoid the explosion of vocabulary size and the *out-of-vocabulary* problems when using a word-level representation. We denote $E \in \mathcal{R}^{4500 \times d}$ as an embedding layer with d as the embedding size. Given the input utterance transcript u_t, we use the pretrained SP tokenizer to tokenize the input transcript and project the tokenized wordpieces into E to obtain the wordpiece embeddings $\mathbf{P^t}$ = $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$, where n is the number of tokenized wordpieces.

Context Encoder

Figure 1 shows our methods for encoding dialogue acts (in light orange color) and previous utterance transcripts (in dark orange color). We describe the details of our context encoder methods below:

Encoding Dialogue Acts: Its input contains a list of dialogue action and slot pairs $\mathcal{D}^t$=$\{(a_1, s_1)$, ..., $(a_{|\mathcal{D}_t|}, s_{|\mathcal{D}_t|})\}$. Given that ℓ_D is the maximum number of dialogue action-slot pairs in all training input data instances, if $|\mathcal{D}^t| < \ell_D$, we pad $\mathcal{D}^t$ with default action-slot pairs until reaching ℓ_D. During inference, if an utterance in the test set has more than ℓ_D action-slot pairs, we only take the latest ℓ_D pairs. We use *zero-embeddings* for the padding actions and slots so that they have no effect on our model.

Embedding Layer: This layer maintains two embedding matrices: a dialogue action embedding matrix $\mathcal{A} \in \mathcal{R}^{|A| \times d}$ and a dialogue slot embedding matrix $\mathcal{S} \in \mathcal{R}^{|S| \times d}$, where $|\mathcal{A}|$ and $|\mathcal{S}|$ refer to the number of dialogue actions and slots in the model, respectively. By projecting each action a_i and slot s_i in the action-slot pair $(a_i, s_i) \in \mathcal{D}^t$ via $\mathcal{A}$ and $\mathcal{S}$, we obtain their corresponding embeddings $\mathbf{a_i}$ and $\mathbf{s_i}$.

Processing Layer: With each $(a_i, s_i) \in \mathcal{D}$, we obtain output action embedding $\mathbf{a_i}$ and slot embedding $\mathbf{s_i}$ from the embedding layer. We then perform an element-wise addition to fuse $\mathbf{a_i}$ and $\mathbf{s_i}$. Next, we transform the fused embedding of $\mathbf{a_i}$ and $\mathbf{s_i}$ by a linear transformation with a *ReLu* activation to obtain $\mathbf{g}_i$ as follows:

$$\mathbf{g}_i = ReLU\big(W_{\mathbf{g}}(\mathbf{a_i} + \mathbf{s_i}) + b_{\mathbf{g}}\big) \tag{1}$$

For all $|\mathcal{D}^t|$ action-slot pairs in $\mathcal{D}^t$, we obtain the corresponding fused embeddings $\{\mathbf{g}_1, \mathbf{g}_2 ..., \mathbf{g}_{l_D}\}$ by following the same process that produces $\mathbf{g}_i$ in Eq.(1). To obtain the output embeddings, we perform a row-wise concatenation across the fused embeddings.

$$\mathbf{G}^t = \mathbf{g}_1 \oplus \mathbf{g_2} \oplus ... \oplus \mathbf{g}_{l_D} \tag{2}$$

Encoding Previous Utterance Transcripts: Its input is a list of previous utterance transcripts $\mathcal{U}^t$=$\{u_1, u_2, ..., u_{t-1}\}$. To learn the contextual embeddings of an utterance transcript $u_j \in \mathcal{U}^t$, we use the pre-trained uncased BERT-based language model (Devlin et al. 2019). Specifically, we first tokenize each u_j with the BERT-based tokenizer. Next, we prepend a *[CLS]* token and append a *[SEP]* token to the tokenized transcript. Since utterances at different turns have a different number of previous utterance transcripts, we use ℓ_U as the maximum number of turns in all the training examples. At turn t-th ($t < \ell_U$), we pad $\ell_U - t$ *empty transcripts* to obtain a length of ℓ_U. During inference, if an utterance has more than ℓ_U turns, we only take its latest ℓ_U previous utterance transcripts.

Processing Layer: We input each $u_j \in \mathcal{U}^t$ into the pre-trained BERT-based model and extract the embeddings from

[1] https://github.com/google/sentencepiece

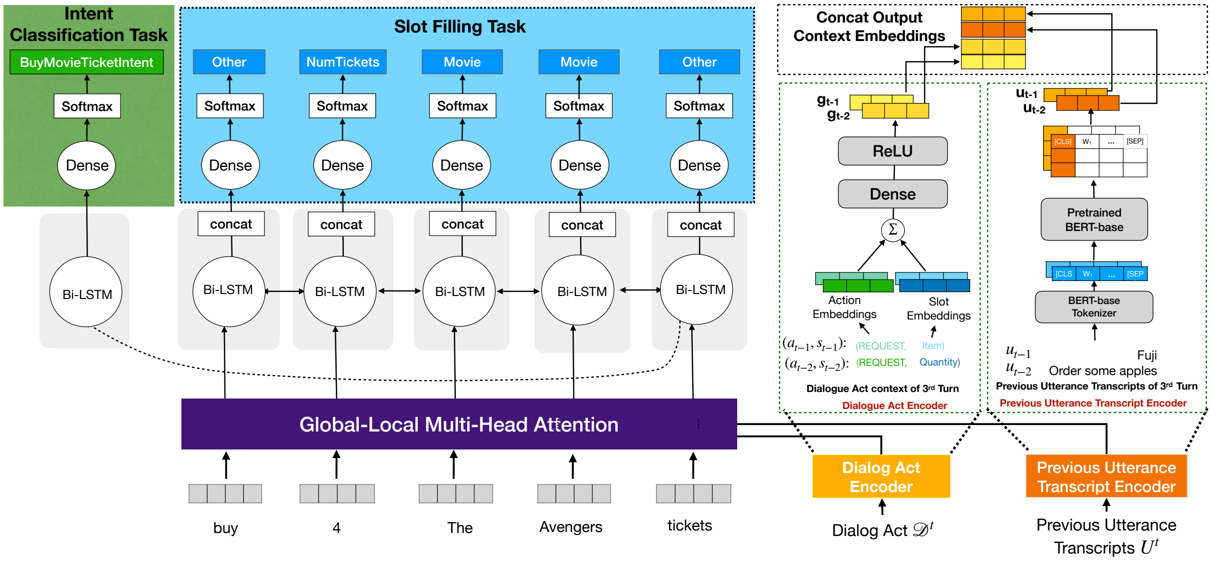


Figure 1: Architecture of our proposed contextual SLU model.

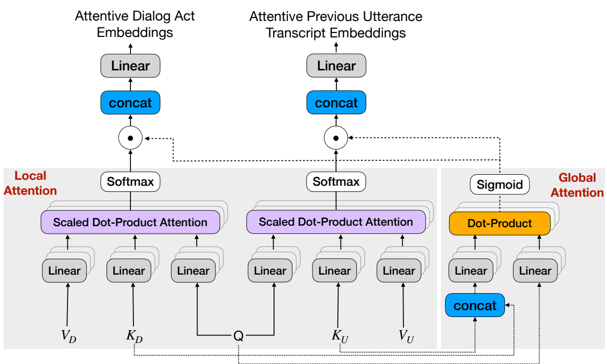


Figure 2: Adaptive attention via global-local context fusion.

the *[CLS]* token as the summarized embeddings for u_j. For all previous utterance transcripts in $\mathcal{U}^t$, we obtain the corresponding output embeddings $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{\ell_U}\}$. We mask the padded empty transcripts as *zero* embeddings so that they have no effect on our model performance.

Output: We perform a row-wise concatenation for all previous utterance transcripts' embeddings $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{\ell_U}\}$ as follows:

$$\mathbf{U}^t = \mathbf{u}_1 \oplus \mathbf{u}_2 \oplus ... \oplus \mathbf{u}_{\ell_U} \tag{3}$$

Adaptive Global-Local Context Fusion

Figure 2 shows our proposed adaptive global-local context fusion mechanism. We use the multi-head attention to compute local attention scores and design a global attention mechanism to measure the contribution of all contexts. Then, we fuse the global and local attention scores into one. Details of this architecture are described below:

Global-Local Multi-Head Attention Layer: Recall $\mathbf{G}^t$ is a row-wise concatenation of all dialogue act embeddings (Equation 2), $\mathbf{U}^t$ is a row-wise concatenation of all previous utterance transcript embeddings (Equation 3), and $\mathbf{P}^t = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ is the wordpiece embeddings. Considering $\mathbf{P}^t$ as the query, we apply the scaled dot attention (Vaswani et al. 2017) to measure the local attention scores α_G between $\mathbf{G}^t$ and $\mathbf{P}^t$, and the local attention scores α_U between $\mathbf{U}^t$ and $\mathbf{P}^t$ as follows:

$$\alpha_G = \text{softmax}\left(\frac{Q_G K_G^T}{\sqrt{d}}\right); \quad \alpha_U = \text{softmax}\left(\frac{Q_U K_U^T}{\sqrt{d}}\right) \tag{4}$$

where Q_G, K_G and V_G are learned by linearly transforming the corresponding $\mathbf{P}^t$ and $\mathbf{G}^t$. Q_U, K_U and V_U are learned by linearly transforming the corresponding $\mathbf{P}^t$ and $\mathbf{U}^t$.

$$\begin{aligned} Q_G &= W_G^{(q)}\mathbf{P}^t + b_G^{(q)};\ K_G = W_G^{(k)}\mathbf{G}^t + b_G^{(k)};\ V_G = W_G^{(v)}\mathbf{G}^t + b_G^{(v)} \\ Q_U &= W_U^{(q)}\mathbf{P}^t + b_U^{(q)};\ K_U = W_U^{(k)}\mathbf{U}^t + b_U^{(k)};\ V_U = W_U^{(v)}\mathbf{U}^t + b_U^{(v)} \end{aligned} \tag{5}$$

To measure global attention scores, we first perform a column-wise concatenation between $\mathbf{G}^t$ and $\mathbf{U}^t$, resulting in a long context vector $\mathbf{C}^t \in \mathcal{R}^{1\times(\ell_D \times d\ +\ \ell_U \times 768)}$ (d is the dialogue act embedding size and 768 is the BERT-based embedding size). Then, we measure the global attention scores as following:

$$\beta = \text{sigmoid}\Big(Q_\beta K_\beta^T\Big) \tag{6}$$

where Q_β, K_β are learned by linearly transforming $\mathbf{P}^t$ and $\mathbf{C}^t$ as follows:

$$Q_\beta = W_\beta^{(q)}\mathbf{P}^t + b_\beta^{(q)};\ \ K_\beta = W_\beta^{(k)}\mathbf{C}^t + b_\beta^{(k)}$$

Note that β is an $n \times 1$ matrix, where each entry $\beta_i \in \beta$ shows how much all the contextual information contributes to each subquery $\mathbf{p}_i \in \mathbf{P}^t$. Thus, we replicate β to have a similar dimension size with α_G and α_U, resulting in β_G and β_U respectively. Then, we perform an element-wise product between α_G and β_G, as well as α_U versus β_U:

$$\gamma_G = \alpha_G \odot \beta_G; \quad \gamma_U = \alpha_U \odot \beta_U$$

Lastly, we perform matrix multiplication between V_G and γ_G to obtain adaptive dialogue act embeddings $\mathbf{C}^t_{G,att} \in$

$\mathcal{R}^{n\times d}$, and between V_U and γ_U to obtain adaptive previous utterance transcript embeddings $C^t_{U,att} \in \mathcal{R}^{n\times 768}$. Finally, we column-wise concatenate $\mathbf{C}^t_{G,att}$ and $\mathbf{C}^t_{U,att}$ with wordpiece embeddings $\mathbf{P}^t$.

$$\begin{aligned} \mathbf{C}^t_{G,att} &= \gamma_G V_G; \quad \mathbf{C}^t_{U,att} = \gamma_U V_U \\ \mathbf{P}^t_{context} &= [\mathbf{P}^t,\ \mathbf{C}^t_{G,att},\ \mathbf{C}^t_{U,att}] \end{aligned} \tag{7}$$

Processing Layer: With $\mathbf{P}^t_{context}$ established, we pass $\mathbf{P}^t_{context}$ through a *m*-layer Bi-LSTM encoder to produce a series of context-aware hidden states $H^t_{slot} = \{h_1^{(slot)}, h_2^{(slot)}, ..., h_n^{(slot)}\}$ and a summarized bidirectional embedding vector $h^{(int)}$. Here, we use the BiLSTM encoder to have a fair comparison against previous works such as (Qin et al. 2021; Gupta, Rastogi, and Hakkani-Tur 2018). In addition, it has also been shown that the BiLSTM encoder outperforms the transformer-based models on the public benchmark datasets used in this study (Qin et al. 2021). Of note, our adaptive global-local context fusion design can be also integrated with the transformer-based models.

$$\begin{aligned} &\overrightarrow{h_i}^{(k)} = \overrightarrow{LSTM}(h_i^{(k-1)}, \overrightarrow{h_{i-1}}) \\ &\overleftarrow{h_i}^{(k)} = \overleftarrow{LSTM}(h_i^{(k-1)}, \overleftarrow{h_{i+1}}) \\ &\text{with } i \in [1,n], k \in [1,m], h_i^{(0)} = \mathbf{p}_i^{(context)} \\ &h_i^{(slot)} = [\overrightarrow{h_i}^{(m)}, \overleftarrow{h_i}^{(m)}],\ h^{(int)} = [\overrightarrow{h_n^{(m)}}, \overleftarrow{h_1}^{(m)}] \end{aligned} \tag{8}$$

Intent Classification and Slot Filling

Intent Classification: It is a multi-class classification problem. We use $h^{(int)}$ in Eq. (8) to produce an intent distribution over all $|I|$ intents at each input utterance u_t. We define the cross entropy loss for u_t as follows:

$$\begin{aligned} \widehat{y}^{(int)}_{t,I} &= \text{softmax}\big(W^{(int)} h^{(int)} + b^{(int)}\big) \\ L_{int} &= -\sum_{j=1}^{|I|} y^{(int)}_{t,j} \log(\widehat{y}^{(int)}_{t,j,I}) \end{aligned} \tag{9}$$

Slot Filling: Similar to the intent classification, we use H^t_{slot} for the slot filling task for u_t with $|S|$ slots over each of n tokens at each input utterance using the following cross entropy loss:

$$\begin{aligned} \widehat{y}^{(slot)}_{t,i,S} &= \text{softmax}\big(W^{(slot)} h_i^{(slot)} + b^{(slot)}\big) \\ L_{slot} &= -\sum_{i=1}^{n}\sum_{k=1}^{|S|} y^{(slot)}_{t,i,k,S} \log(\widehat{y}^{(slot)}_{t,i,k,S}) \end{aligned} \tag{10}$$

Multi-Task Learning: We use a multi-task learning strategy to train our model. The joint cost function is defined as:

$$L = L_{ic} + \lambda L_{sf} \tag{11}$$

where λ are hyper-parameters to control loss contribution.

Experimental Settings

Dataset: We conduct experiments on the bechmark Simulated Dialogue dataset (Sim) (Liu et al. 2018; Shah et al. 2018), which consists of two datasets: (i) Simulated Restaurant Dialogue (Sim-R) and (ii) Simulated Movie Dialogue (Sim-M). Sim-R contains dialogues for booking a restaurant table, whereas Sim-M contains dialogues for buying a movie ticket. Specifically, Sim-R has 11k turns in 1,116 training dialogues, 349 development dialogues, and 775 testing dialogues. Sim-M has nearly 4k turns in 384 training dialogues, 120 development dialogues, and 264 testing dialogues. In total, the Sim dataset has 3 intents, 12 slot types, and 21 user dialogue act types. A key challenge of this dataset is the presence of unseen entities in testing sets. For example, only 13% of movie names in the validation and test sets are in the training set.

In addition, we conduct experiments on our large-scale, de-identified, multi-domain in-house dataset. This dataset has both single turn utterances and multi-turn utterances.

Baseline Models: We compare our proposed model with the following baseline models:

- NoContext: It is a two-layer stacked Bidirectional RNN using GRU and LSTM cells respectively, where no context information is incorporated. We report its best architecture's results.
- PrevTurn: It is similar to the *NoContext* baseline, but encodes only the utterances in previous turns.
- MemNet (Chen et al. 2016): An end-to-end memory network that dynamically exploits the contextual knowledge.
- SDEN (Bapna et al. 2017): It uses a sequential dialogue encoder to encode contexts from the dialogue history in chronological order with recurrent neural networks.
- EfficientNet (Gupta, Rastogi, and Hakkani-Tur 2018): It is a hierarchical recurrent neural network that efficiently encodes dialogue act context.
- GraphNet (Qin et al. 2021): It uses a Graph Convolutional Network for integrating dialogue act contexts.

Note that we do not compare with BERT because Qin at. al., (2021) showed that GraphNet outperformed BERT on the Sim datasets.

Implementation Details: Our experiments are implemented in Tensorflow 2.3 (Abadi et al. 2016). The hyperparameters are selected based on the best performance on the validation set. During training, we minimize the sum of intent and slot losses using Adam optimizer (Kingma and Ba 2015) for 100 training steps with a batch size of 32. We use two BiLSTM layers with each having a wordpiece embedding size of 256. The decoders for slot filling and intent classification both are two-layer dense networks with 256 and 512 units, respectively. For the context encoder, the dialogue act embedding size is set to 256; and the embedding size for BERT encoding of previous utterances is set to 768. The attention size for global-local multi-head attention is set to 256 with 1 head. We set $\lambda = 1$ in Eq. (11) to give equal contribution for L_{int} and L_{slot} losses.

Evaluation Metrics: To benchmark against SOTA approaches, we report the performance of the intent classification task using intent accuracy and the performance of

Model	Sim-R Results		Sim-M Results	
	Intent Acc.	Slot F1	Intent Acc.	Slot F1
NoContext	83.61%	94.24%	88.51%	86.91%
PrevTurn	99.37 %	94.96%	99.12%	88.63%
MemNet-6	99.75%	94.42%	99.12%	89.76%
MemNet-20	99.67%	94.28%	98.76%	90.70%
SDEN-20	99.84%	94.81%	99.60%	90.93%
EfficientNet	99.65%	94.70%	99.27 %	93.73%
GraphNet	99.97%	95.37%	99.93%	94.41%
Our model	**99.97%**	**98.10%**	**100%**	**96.98%**

Table 1: Overall Performance on Sim-R and Sim-M datasets.

Model	Sim-R Results	Sim-M Results
	Slot F1	Slot F1
Our model	**98.10%**	**96.98%**
w/o G-CF	97.81 %	96.54 %
w/o GL-CF	97.34 %	96.35 %

Table 2: Effectiveness of our global-local context fusion. w/o GL-CF means that we remove the global and local context fusion from the proposed model; w/o G-CF means that we remove the global context fusion from the proposed model.

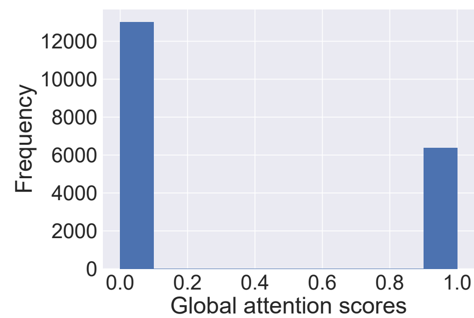
the slot filling task using the slot chunk F1 score (Tjong Kim Sang and Buchholz 2000) for the Sim-R and Sim-M dataset. For the in-house dataset, we report model performance on intent classification error rate (ICER), and semantic error rate (SemER). ICER measures the proportion of utterances with a misclassified intent, i.e. ICER = 1.0 - *intent accuracy*. IRER and SemER measures the utterance-level error rate that considers both intent and slot errors. SemER (Makhoul et al. 1999) combines intent and slot accuracy into a single metric, i.e. SemER = # (slot errors + intent errors) / # (slots + intents in reference). For the in-house dataset, we report relative improvements with respect to the baseline .

Results

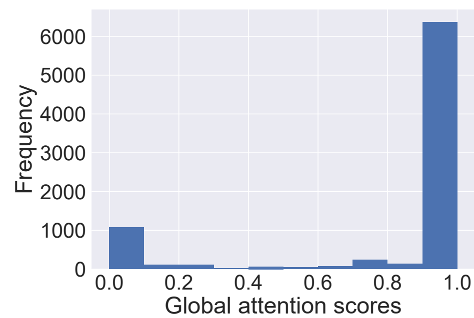
Overall Performance: Table 1 shows the performance of our proposed model and SOTA baselines for intent accuracy and slot F1 on Sim-R and Sim-M datasets. Overall, our proposed model achieves the best performance for intent accuracy and slot F1 on the two benchmarking datasets. On the Sim-R dataset, our model achieves 99.97% intent accuracy and 98.10% slot F1. Compared to the previously best performing model, our model achieves the same intent accuracy and an absolute slot F1 improvement of 2.73% (*p-value* < 0.001 under the non-directional Mann-Whitney U test). On Sim-M dataset, our model achieves 100% intent accuracy and 96.98% slot F1. On averge, our model achieves better results than the previously best performed model, with an absolute intent accuracy improvement of 0.07% and an absolute slot F1 improvement of 2.57% (*p-value* < 0.001).

Next, we present ablation studies to understand the effectiveness of our global-local context fusion design and different contexts. As obtaining a high intent accuracy is trivial on Sim-R and Sim-M datasets, we report only Slot F1 results.

Effectiveness of Global-Local Fusion: Table 2 shows that removing global-local fusion has a negative impact on our model performance, with an average drop of 0.08% in the absolute intent accuracy and 0.3% in the absolute slot F1. To further verify the effectiveness of the global attention, we plot the histogram of global attentive scores in Sim-R and Sim-M datasets. Figure 3 shows global attention scores are mostly distributed in ranges of [0.9, 1.0] and [0.0, 0.1] in the Sim-R and Sim-M datasets. This suggests that our global-local fusion can reduce the context contributions when all global attention scores are close to [0.0, 0.1], which is an enhancement over the traditional multi-head attention approach (Vaswani et al. 2017).



(a) Sim-R



(b) Sim-M

Figure 3: Distribution of global attention scores in Sim-R and Sim-M datasets. The global attention scores (blue bars) are mostly distributed in ranges of [0.9, 1.0] and [0.0, 0.1].

Model	Sim-R Results	Sim-M Results
	Slot F1	Slot F1
Our model	**98.10%**	**96.98%**
w/o DialogAct	97.09%	95.91 %
w/o PrevUtt	96.85 %	96.12 %

Table 3: Effectiveness of different contexts. w/o DialogAct means that we remove the dialogue act encoder from the proposed model. w/o PrevUtt means that we remove the previous utterance encoder from the proposed model.

Effectiveness of Different Contexts: Table 3 shows the ablation study results when removing different contexts and attention mechanisms. A first observation is that removing dialogue acts leads to the lowest performance on the Sim-M dataset, whereas removing previous utterances leads to the lowest performance on the Sim-R dataset. This indicates that dialogue history contexts play a crucial role in improving SLU task in the multi-turn dialogue setting.

		ICER	SEMER	IRER
Overall	NoContext	baseline	baseline	baseline
	Our model	↓ 2.86%	↓ 3.35%	↓ 1.03%
2-turn	NoContext	baseline	baseline	baseline
	Our model	**↓ 48.21%**	**↓ 42.15%**	**↓ 37.90%**
3-turn	NoContext	baseline	baseline	baseline
	Our model	↓ 47.77%	↓ 37.29%	↓ 31.66%
4-turn+	NoContext	baseline	baseline	baseline
	Our model	↓ 44.77%	↓ 13.80%	↓ 17.25%

Table 4: Relative error deduction of our proposed model on in-house datasets. ↓ shows the decrease in error rate with respect to the proposed model. Higher numbers suggest more error deduction and better model performance.

Results on In-House Dataset

Table 4 presents the overall performance of our proposed model and the baseline model without contexts. Compared to baseline model without contexts, our proposed model relatively reduces 2.86% of ICER, 3.35% of SEMER, and 1.03% of IRER on the overall dataset that includes both single turn and multi-turn utterances. We further examined the model performance on multi-turn utterances and evaluated utterances with only two turns (2-turn), with only three turns (3-turn), and with four turns or more (4-turn+). For 2-turn datasets, the proposed model relatively reduces 48.21% of ICER, 42.15% of SEMER, and 37.90% of IRER. For 3-turn dataset, the proposed model relatively reduces 47.77% of ICER, 37.29% of SEMER, and 31.66% of IRER. For 4-turn datasets, the proposed model relatively reduces 44.77% of ICER, 13.81% of SEMER, and 17.25% of IRER. These results present a large improvement margin of our model on multi-turn datasets. In addition, these results are consistent with our model performance on the public benchmark datasets (Sim-R and Sim-M), suggesting a critical role of dialogue contexts in improving intent and slot prediction.

We also replicate the ablation study when removing different contexts and global and local context fusion mechanisms on the in-house dataset. Table 5 and 6 show the model performance on multi-turn datasets. An interesting observation is that removing dialogue acts leads to the biggest performance degradation for our proposed model on all the examined multi-turn utterances. For example, after removing dialogue acts, the ICER relatively drops by 22.89% for 2-turn utterances, 100.95% for 3-turn utterances, and 53.65% for 4-turn+ utterances. Similarly, the SEMER relatively drops by 24.03% for 2-turn utterances, 58.67% for 3-turn utterances, and 13.30% for 4-turn+ utterances. These findings are also consistent with what we observe on the public datasets (Sim-R and Sim-M), suggesting the importance of dialogue act history in the intent and slot prediction in multi-turn dialogues. Across all the turns, 3-turn utterances suffer the most performance degradation when removing dialogue acts, whereas 2-turn utterances suffer the most performance degradation when removing previous utterances.

We also observe that the proposed model has the lowest performance for slot prediction when we remove global and local context fusion. Across all the turns, the 4-turn+ datasets suffer the most for intent prediction, whereas 2-turn datasets suffer the most for slot prediction. These findings suggest the importance of global and local context fusion in predicting intents and slots in multi-turn dialogues.

		ICER	SEMER	IRER
2-turn	Our model	proposed	proposed	proposed
	w/o DialogAct	↑ **22.89** %	↑ **24.03** %	↑ **23.78** %
	w/o PrevUtt	↑ 4.18%	↑ 4.06%	↑ 5.71%
3-turn	Our model	proposed	proposed	proposed
	w/o DialogAct	↑ **100.95** %	↑ **58.67** %	↑ **53.79**%
	w/o PrevUtt	↑ **2.64**%	↑ **1.35**%	↑ **1.19**%
4-turn +	Our model	proposed	proposed	proposed
	w/o DialogAct	↑ **53.65**%	↑ **13.30**%	↑ **12.38**%
	w/o PrevUtt	↑ 1.10%	↑ 0.31%	↑ 0.71%

Table 5: Relative performance degradation when removing dialogue contexts on in-house multi-turn datasets. ↑ shows the increase in error rate with respect to the proposed model. Higher numbers suggest more degradation on the model performance. w/o DialogAct means that we remove the dialogue act encoder from the proposed model. w/o PrevUtt means that we remove the previous utterance encoder from the proposed model.

		ICER	SEMER	IRER
2-turn	Our model	proposed	proposed	proposed
	w/o GL-CF	↑ **17.12**%	↑ **23.31**%	↑ **23.35**%
	w/o G-CF	↑ 3.01%	↑ 2.35%	↑ 2.52%
3-turn	Our model	proposed	proposed	proposed
	w/o GL-CF	↑ **30.67**%	↑ **11.95**%	↑ **14.84**%
	w/o G-CF	↑ 4.11%	↑ 2.79%	↑ 1.30%
4-turn+	Our model	proposed	proposed	proposed
	w/o GL-CF	↑ **34.47**%	↑ **10.92**%	↑ **8.00**%
	w/o G-CF	↑ 4.50%	↑ 0.70%	↑ 0.92%

Table 6: Effectiveness of global and local context fusion on the in-house multi-turn datasets. ↑ shows the increase in error rate with respect to the proposed model. Higher numbers suggest more degradation on the model performance. w/o GL-CF means that we remove the global and local context fusion from the proposed model; w/o G-CF means that we remove the global context fusion from the proposed model.

Conclusion

We propose a novel E2E SLU model designed for multi-turn dialogues where dialogue acts and previous utterance transcripts are utilized as contexts to improve the performance of intent prediction and slot filling tasks. We introduce a global-local multi-head attention mechanism to effectively incorporate contextual signals into our model. We demonstrate that our proposed approach improves intent accuracy and slot F1 – two well known SLU metrics over six state-of-the-art baselines on two publicly available datasets. Extensive experiments on an in-house dataset further verify the effectiveness of our proposed model.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 265–283. USENIX Association.

Abro, W. A.; Qi, G.; Gao, H.; Khan, M. A.; and Ali, Z. 2019. Multi-turn intent determination for goal-oriented dialogue systems. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Allen, J. 1995. *Natural Language Understanding (2nd Ed.)*. USA: Benjamin-Cummings Publishing Co., Inc. ISBN 0805303340.

Bapna, A.; Tür, G.; Hakkani-Tür, D.; and Heck, L. 2017. Sequential Dialogue Context Modeling for Spoken Language Understanding. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 103–114. Association for Computational Linguistics.

Bhargava, A.; Celikyilmaz, A.; Hakkani-Tür, D.; and Sarikaya, R. 2013. Easy contextual intent prediction and slot detection. In *2013 ieee international conference on acoustics, speech and signal processing*, 8337–8341. IEEE.

Chen, Q.; Zhuo, Z.; Wang, W.; and Xu, Q. 2019. Transfer learning for context-aware spoken language understanding. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 779–786. IEEE.

Chen, Y.-N.; Hakkani-Tür, D.; Tür, G.; Gao, J.; and Deng, L. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, 3245–3249.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Gupta, A.; Zhang, P.; Lalwani, G.; and Diab, M. 2019. CASA-NLU: Context-Aware Self-Attentive Natural Language Understanding for Task-Oriented Chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1285–1290. Association for Computational Linguistics.

Gupta, R.; Rastogi, A.; and Hakkani-Tur, D. 2018. An efficient approach to encoding context for spoken language understanding. *arXiv preprint arXiv:1807.00267*.

Kingma, D. P.; and Ba, J. L. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Liu, B.; Tür, G.; Hakkani-Tür, D.; Shah, P.; and Heck, L. 2018. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2060–2069.

Makhoul, J.; Kubala, F.; Schwartz, R.; Weischedel, R.; et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, 249–252. Herndon, VA.

Miller, S.; Stallard, D.; Bobrow, R.; and Schwartz, R. 1996. A fully statistical approach to natural language interfaces. In *34th Annual Meeting of the Association for Computational Linguistics*, 55–61.

Qin, L.; Che, W.; Ni, M.; Li, Y.; and Liu, T. 2021. Knowing Where to Leverage: Context-Aware Graph Convolutional Network With an Adaptive Fusion Layer for Contextual Spoken Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1280–1289.

Shah, P.; Hakkani-Tür, D.; Tür, G.; Rastogi, A.; Bapna, A.; Nayak, N.; and Heck, L. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Su, S.-Y.; Yuan, P.-C.; and Chen, Y.-N. 2018. How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2133–2142.

Su, S.-Y.; Yuan, P.-C.; and Chen, Y.-N. 2019. Dynamically context-sensitive time-decay attention for dialogue modeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7200–7204. IEEE.

Tjong Kim Sang, E. F.; and Buchholz, S. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, 127–132. Association for Computational Linguistics.

Tur, G.; and De Mori, R. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Y.; He, T.; Fan, R.; Zhou, W.; and Tu, X. 2019. Effective Utilization of External Knowledge and History Context in Multi-turn Spoken Language Understanding Model. In *2019 IEEE International Conference on Big Data (Big Data)*, 960–967. IEEE.

Wei, K.; Tran, T.; Chang, F.-J.; Sathyendra, K. M.; Muniyappa, T.; Liu, J.; Raju, A.; McGowan, R.; Susanj, N.; Rastrow, A.; et al. 2021. Attentive Contextual Carryover for Multi-Turn End-to-End Spoken Language Understanding. *arXiv preprint arXiv:2112.06743*.