Inducing Global and Local Knowledge Attention in Multi-turn Dialog Understanding

Anonymous ACL submission

Abstract

In multi-turn dialog understanding, semantic frames are constructed by detecting intents and slots within each user utterance. However, recent works lack the capability of modeling multi-turn dynamics within a dialog where the contexts are mostly adopted for updating dialog states instead of capturing overall intent semantic flows in spoken language understanding (SLU). Moreover, external knowledge related to dialogs may be beneficial in exploring deep semantic information across dialog turns, which many works only considered for end-to-end response generation. In this paper, we propose to equip a BERT-based joint framework with a context attention module and a knowledge attention module to introduce knowledge attention with contexts between two SLU tasks. We propose three attention mechanisms to induce both global and local attention on knowledge triples. Experimental results in two complicated multi-turn dialog datasets have demonstrated significant improvements of our proposed framework by mutually modeling two SLU tasks with filtered knowledge and dialog contexts. Attention visualization also provides nice interpretability of how our modules leverage knowledge across the utterance.

1 Introduction

In conventional task oriented dialog systems, spoken language understanding (SLU) modules aim to transform utterances into meaningful semantic representations for dialog management (Weld et al., 2021; Zhang et al., 2020). It mainly detects associated dialog acts or intents and extracts key slot information as so-called *'semantic frames'* (Abbeduto, 1983), shown in Table 1. An external knowledge base may also proffer some knowledge triples when predicting the overall intent semantics and slot values of individual words.

In early attempts of SLU tasks, utterances were isolated and analyzed separately for user intents and semantic slots (Raymond and Riccardi, 2007; Liu et al., 2017). However, such ambivalent treatment hinders the transitions of shared knowledge for each supervised signal. Models that maximize the joint distribution likelihood were then proposed to amend the gap (Liu and Lane, 2016; Wang et al., 2018; Wu et al., 2021a; Li et al., 2018a). Some works also tackled utterances with multiple intents (Qin et al., 2019; Rashmi Gangadharaiah, 2019; Qin et al., 2020). While driven by large pretrained corpus, these methods still fall short of employing complete dynamic interactions within dialogs. In contrast, humans can naturally adopt history contexts to identify intentions with their background knowledge. Some works have then integrated previous dialog contexts for more robust SLU (Wang et al., 2019; Gupta et al., 2019; Su et al., 2021; Wu et al., 2021c).

Speaker	**Utterance**
1. User	Is there something that's maybe a good **intelligent** **comedy**?
Act & Slots:	*Request (genre:* ***comedy****)*
Knowledge:	*(**intelligent**; related to; well_informed)* *(**comedy**; related to; comic)* *(**comedy**; is a; drama)*
2. System	Whiskey Tango **Foxtrot** is the only **Adult** comedy I see playing in your **area**. Would you like to try that?
Act & Slots:	*Inform (movie: Whiskey Tango* ***Foxtrot****)* *Inform (genre:* ***Adult*** *comedy)* *Inform (distance limits: in your* ***area****)* *Confirm_question*
Knowledge:	*(**foxtrot**; related to; dance)* *(**foxtrot**; related to; rhythm)* *(**adult**; capable of; work)* *(**area**; is a; region)*

Table 1: Excerpt of a single turn within a dialog with corresponding dialog acts, slots and knowledge samples that are related to **keywords** in the utterance.

Nevertheless, inadequacy of considering external knowledge may limit the machine to fully digest contexts and set constraints of comprehension boundaries. Much efforts have pushed forward the progress in knowledge grounded dialog generation

(Wang et al., 2021b; Zhao et al., 2020; Zheng et al., 2021), where relevant documents or a knowledge base auxiliarily guide the language autoregressive progress. Term-level denoising (Zheng et al., 2021) or filtering techniques (Wang et al., 2021b) refine the adopted knowledge for better semantic considerations. Therefore, utilizing the correlation between language and knowledge is also imperative to some extent diminish ambiguity in dialog context understanding by extending their external semantics, which recent SLU works often neglect. Wang et al. (2019) has proposed to adopt knowledge attention for joint tasks. However, it adopts a single LSTM layer to couple all knowledge without filtering and contexts for two tasks, which cannot model complex interactions well and is ambiguous in how these two components affect each other.

To solve above concerns, we propose a **G**lobal and **L**ocal **K**nowledge **A**ttention Framework (GLKA) to effectively incorporate dialog history and external knowledge in joint SLU tasks. We propose three different attention modules that consider local and global awareness of knowledge at token and utterance levels respectively. After obtaining knowledge-enriched vectors, we predict intents and slots coherently with two LSTM decoders with different fused inputs. Experiment results have shown superior performances of our methods in manipulating contexts and knowledge and beat all competitive baselines. Our contributions are as follows:

1. We propose our SLU frameworks: LKA, GKA and GLKA to dynamically select external knowledge for current utterance and previous dialog history for joint multiple dialog act and slot filling detection, where previous SLU works are not grounded with knowledge and contexts.
2. We explore the mechanisms of how SLU models should consider commonsense knowledge locally or globally and demonstrate the effectiveness of different attention in each subtask.
3. Experimental and attention visualization results show that our model achieves superior performances over several competitive baselines and provide good interpretability of how our model utilizes the knowledge.

2 Problem Formulation

For each utterance $x_n = \{w_1^n, w_2^n, \dots, w_T^n\}$ in a task-oriented dialog $\mathbf{X}$ with N utterances, given the domain ontology of a dialog act set $\mathbf{A}$ and a slot set $\mathbf{S}$, we aim to find one or more acts $\{a_i^n\}$ [1] and a sequence of slot tags $\{s_1^n, s_2^n, \dots, s_T^n\}$ to construct a semantic frame. Namely, we hope to maximize the joint log likelihood of $\mathbf{A}$ and $\mathbf{S}$ in Eq. 1 given a parametrized model θ, its context $\mathbf{C_n} = \{x_1, \dots, x_{n-1}\}$ and associated knowledge $\mathbf{K_n} = \phi(K_G, x_n)$ for the current utterance x_n. We deem K_G as an external large knowledge base with knowledge triples and $\phi(\cdot)$ helps to extract related knowledge pairs for x_n. It will be critical to match correct knowledge based on current dialog history and the utterance for better dialog understanding.

$$\mathcal{L}(\mathbf{A}, \mathbf{S}) \triangleq \sum_n log\ P(A_n, S_n \mid x_n, \mathbf{C_n}, \mathbf{K_n}; \theta) \tag{1}$$

3 Methodology

3.1 Context Attention

Our overall framework is illustrated in Figure. 1. To fully leverage the dialog context information, we propose to first encode the dialog at token and turn levels respectively. At token level, we adopt BERT (Devlin et al., 2019), a powerful contexualized representation model in NLP, to extract semantic representations. For each utterance x_n in a dialog X, we encode it with BERT and obtain the token-level representations $\mathbf{H} = \{h_1, h_2, \dots, h_N\}$ for N utterances.

At turn level, to better capture semantic flows within a dialog, we first take the hidden vectors from the [CLS] token of each utterance's representations $\mathbf{H}$ to form $\mathbf{H'}$ as unified sentence representations. Then, by denoting H_b as the BERT hidden size, we further encode $\mathbf{H'} \in \mathbb{R}^{N \times H_b}$ with a context-aware unidirectional transformer encoder with the hidden size H_a, which contains a stack of L layers with each layer of a masked multi-head self-attention sublayer (MHA) and a point-wise feed forward network (FFN) with residual mechanism and layer normalization. We will send $\mathbf{H'}$ as the first layer input $\mathbf{C^1}$ and iteratively encode with two sublayers in Eq. 2. For each layer, it will first project the input $\mathbf{C}$ with weight matrices: $\{\mathbf{W^Q}, \mathbf{W^K}, \mathbf{W^V}\} \in \mathbb{R}^{H_b \times H_a}$ to be $\mathbf{C^Q} = \mathbf{CW^Q}$, $\mathbf{C^K} = \mathbf{CW^K}$, $\mathbf{C^V} = \mathbf{CW^V}$. Then each of them will be separated into h heads, with each head i to be $\mathbf{C_i} \in \mathbb{R}^{N \times (H_a/h)}$. These $\mathbf{C_i}$ will be sent into a self-attention and a feed forward layer in Eq. 3 and Eq. 4. Here $f(\cdot)$ is softmax function. Finally, we will obtain the final contextual

[1] Here we refer the intent detection problem in dialogs as predicting the dialog acts for each utterance.

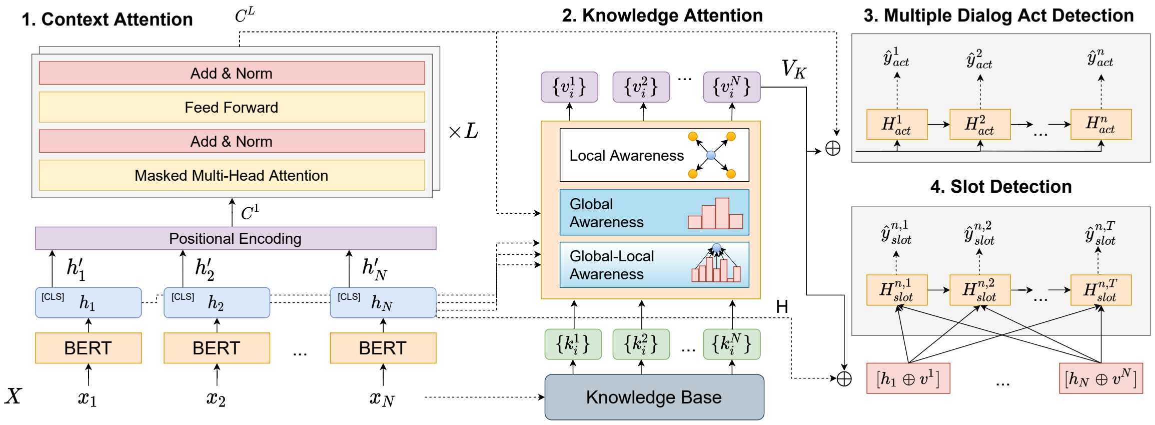

Figure 1: Illustration of our proposed framework for joint dialog act detection and slot filling in multi-turn dialogs. It consists of context and knowledge attention modules, and two LSTM-based decoders. The utterance-level representations will be encoded with the context attention module and token-level representations will interact with their corresponding knowledge in three proposed awareness submodules.

dialog representations $\mathbf{C^L}$.

$$\mathbf{C^l} = FFN(MHA(\mathbf{C^{l-1}}, \mathbf{C^{l-1}}, \mathbf{C^{l-1}})) \tag{2}$$

$$MHA(\mathbf{C_i^Q}, \mathbf{C_i^K}, \mathbf{C_i^V}) = f(\frac{\mathbf{C_i^Q}(\mathbf{C_i^K})^T}{\sqrt{H_b}})\mathbf{C_i^V} \tag{3}$$

$$FFN(x) = max(0,\ x\mathbf{W_1} + b_1)\mathbf{W_2} + b_2 \tag{4}$$

3.2 Knowledge Attention

To simulate the human awareness of coherently relating current contexts to background knowledge, for each utterance $x_n = \{w_1^n, w_2^n, \dots, w_T^n\}$, we also retrieve a T length knowledge sequence $\mathbf{K_n} = \{k_1^n, k_2^n, \dots, k_T^n\}$. Each k_T^n is retrieved from the knowledge base K_G using similar word matching on w_i^n, the i-th word in the utterance x_n. Each k_i^n is a collection of multiple related triples $\gamma = \{h, r, t\}$, as head entity, relation, and tail entity. We propose three different ways of inducing our reasoning module's attention on the associated knowledge, which are also illustrated in Figure. 2:

3.2.1 Local awareness

For each word w_i^n, we have k_i^n representing the commonsense knowledge related to it. We could then directly adopt an attention mechanism to dynamically perceive importance of knowledge triples based on its local word relevance and obtain the knowledge-aware vector $v_i^{n\prime}$.

$$v_i^{n\prime} = \sum_{j=1}^{M} \alpha_{ij}[r_{ij}^n; t_{ij}^n] \tag{5}$$

$$\alpha_{ij} = exp(\beta_{ij}) / \sum_{m=1}^{M} exp(\beta_{im}) \tag{6}$$

$$\beta_{ij} = (h_i^n \mathbf{W^H})(tanh(r_{ij}^n \mathbf{W^R} + t_{ij}^n \mathbf{W^T}))^T \tag{7}$$

r_{ij}^n, t_{ij}^n are relation and tail entity vectors. $\mathbf{W^H}, \mathbf{W^R}, \mathbf{W^T}$ are learnable matrices during training. M is the number of knowledge triples. $[;]$ is the concatenation of two vectors. Given the token-level representations for each word h_i^n in the utterance x_n, attention weights are assigned to reveal the relevance of each knowledge triple under current contexts.

Knowledge triples are mostly associated with name entities and not non-alphabetic words. We instead replace triple vectors of these words as zero vectors to represent agnosticism of knowledge, which will nonetheless introduce redundant noises. Therefore, we propose a gated mechanism for each word h_i^n to regulate the degree of knowledge $v_i^{n\prime}$ induced for downstream tasks.

$$v_i^n = g_i \cdot h_i^n + (1 - g_i) \cdot v_i^{n\prime} \tag{8}$$

$$g_i = \sigma(\mathbf{W_g}[h_i^n; v_i^n] + b_g) \tag{9}$$

3.2.2 Global awareness

The above mechanism restricts its scope of exploring the relevance in intra-word knowledge to current contexts. However, several semantic slots

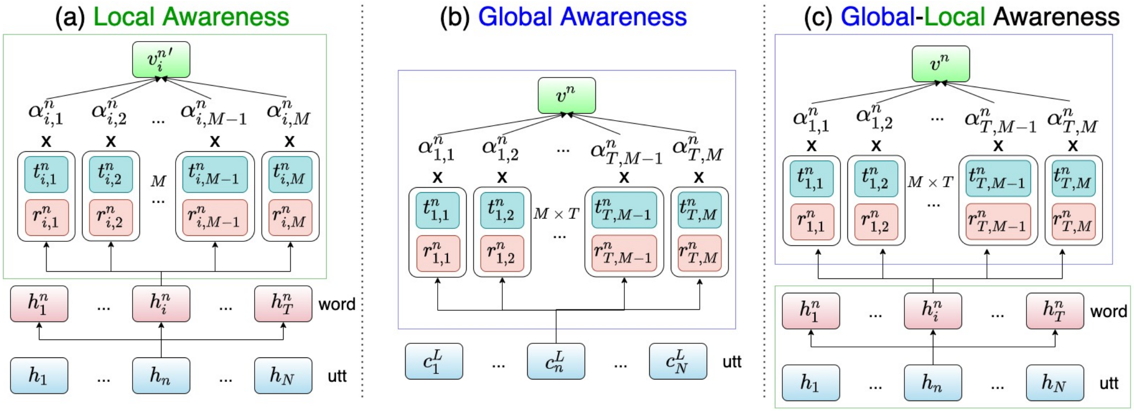
(a) Local Awareness
$v_i^{n\prime}$
$\alpha_{i,1}^n$ $\alpha_{i,2}^n$ ... $\alpha_{i,M-1}^n$ $\alpha_{i,M}^n$
x
$t_{i,1}^n$ $t_{i,2}^n$ M ... $t_{i,M-1}^n$ $t_{i,M}^n$
$r_{i,1}^n$ $r_{i,2}^n$ $r_{i,M-1}^n$ $r_{i,M}^n$
h_1^n ... h_i^n ... h_T^n word
h_1 ... h_n ... h_N utt
(b) Global Awareness
v^n
$\alpha_{1,1}^n$ $\alpha_{1,2}^n$ ... $\alpha_{T,M-1}^n$ $\alpha_{T,M}^n$
$t_{1,1}^n$ $t_{1,2}^n$ $M \times T$ ... $t_{T,M-1}^n$ $t_{T,M}^n$
$r_{1,1}^n$ $r_{1,2}^n$ $r_{T,M-1}^n$ $r_{T,M}^n$
c_1^L ... c_n^L ... c_N^L utt
(c) Global-Local Awareness
v^n
$\alpha_{1,1}^n$ $\alpha_{1,2}^n$ ... $\alpha_{T,M-1}^n$ $\alpha_{T,M}^n$
$t_{1,1}^n$ $t_{1,2}^n$ $M \times T$ ... $t_{T,M-1}^n$ $t_{T,M}^n$
$r_{1,1}^n$ $r_{1,2}^n$ $r_{T,M-1}^n$ $r_{T,M}^n$
h_1^n ... h_i^n ... h_T^n word
h_1 ... h_n ... h_N utt

Figure 2: Three submodules to induce knowledge awareness. (a) Local awareness performs attention at token-level with intra-word knowledge. (b) Global awareness takes all knowledge related to the utterance for context-based attention. (c) Global-Local awareness performs attention at token-level but with all inter-word knowledge.

may be expressed as phrases rather than individual words and the overall intent should spread across the entire utterance. Therefore, instead of attending knowledge locally, we aggregate the knowledge triples from all words $\mathbf{K_n} = \{k_i^n\}$ into a dense matrix and directly find the attention weights for our utterance-level contexts c_n^L.

$$v^n = \sum_{t=1}^{T} \sum_{j=1}^{M} \alpha_{tj} [r_{tj}^n; t_{tj}^n] \qquad (10)$$

$$\alpha_{tj} = exp(\beta_{tj}) / \sum_{t=1}^{T} \sum_{m=1}^{M} exp(\beta_{tm}) \qquad (11)$$

$$\beta_{tj} = (c_n^L \mathbf{W^H})(tanh(r_{tj}^n \mathbf{W^R} + t_{tj}^n \mathbf{W^T}))^T \qquad (12)$$

3.2.3 Global-Local awareness

At last, we combine the view of global and local awareness by considering local attention mechanism mentioned in Eq. 13 with the global knowledge $\mathbf{K_n}$. We could avert the circumstances where some out-of-vocabulary words may not have relevant knowledge by considering knowledge from other words in proximity. Here the knowledge-aware vector v_i^n will be obtained by summing up all knowledge vectors in the sentence x_n:

$$v_i^n = \sum_{t=1}^{T} \sum_{j=1}^{M} \alpha_{tj} [r_{tj}^n; t_{tj}^n] \qquad (13)$$

where T is the number of words in the sentence x_n.

3.3 Semantic Decoder

After obtaining the knowledge-enriched representations $\mathbf{V_K} = \{v^n\}$ along with contextual dialog representations $\mathbf{C^L}$ and the initial token-level representations $\mathbf{H}$, we adopt a BiLSTM for slot filling and a LSTM to detect multiple dialog acts.

$$\mathbf{H_{slot}} = \mathbf{BiLSTM}([\mathbf{H}; \mathbf{V_K}], \mathbf{C^L}) \qquad (14)$$

$$\mathbf{H_{act}} = \mathbf{LSTM}([\mathbf{C^L}; \mathbf{V_K}]) \qquad (15)$$

For slot filling, $\mathbf{V_K}$ will be first concatenated with $\mathbf{H}$ and serve as the inputs of BiLSTM with initial hidden states of $\mathbf{C^L}$, where contexts will assist the slot prediction at each knowledge-enhanced time step. At the same time, $\mathbf{V_K}$ will also be concatenated with dialog contexts $\mathbf{C^L}$ to serve as inputs in a unidirectional LSTM. Finally, we can generate logits $\hat{y}_{act} = \sigma(\mathbf{H_{act}} \mathbf{W_{act}})$ by transforming $\mathbf{H_{act}}$ with $\mathbf{W_{act}} \in \mathbb{R}^{H_L \times |\mathcal{Y}^a|}$ and a sigmoid function σ. H_L is LSTM hidden size and $|\mathcal{Y}^a|$ is the size of dialog act set. Likewise, we compute $\hat{y}_{slot} = softmax(\mathbf{H_{slot}} \mathbf{W_{slot}})$. Total loss will be the combination between the binary cross entropy loss based on $\hat{y}_{act}$ and the cross entropy loss based on $\hat{y}_{slot}$ as shown in Eq. 16 and Eq. 17. Finally, the joint objective is formulated as the weighted sum of $\mathcal{L}_a$ and $\mathcal{L}_s$.

$$\mathcal{L}_a \triangleq -\sum_{n=1}^{N} \sum_{a=1}^{|\mathcal{Y}^a|} (y_a^n log(\hat{y}_a^n) + (1 - y_a^n) log(1 - (\hat{y}_a^n)) \qquad (16)$$

$$\mathcal{L}_s \triangleq -\sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{|\mathcal{Y}^s|} (y_s^{(n,t)} log(\hat{y}_s^{(n,t)})) \qquad (17)$$

4 Experiment Setting

4.1 Experimental setup

We evaluate our proposed framework on two large-scale dialog datasets, i.e. Microsoft Dialog Challenge dataset (MDC) (Li et al., 2018b) and Schema-Guided Dialog dataset (SGD) (Rastogi et al., 2019). **MDC** contains human-annotated conversations in three task-completion domains (movie, restaurant, taxi) with total 11 dialog acts and 50 slots. **SGD** entails large-scale task-oriented dialogs over 20 domains ranging from travel, weather to banks, etc. It has total 18 dialog acts and 89 slots. We randomly select 1k dialogs for each domain in MDC and two domains (restaurant, flights) from SGD for total 5k dialogs in 7:3 training and testing ratio. Each utterance is labeled with one or more dialog acts and several slots.

4.2 Baselines

We compare our models with several competitive baselines which sequentially include more features for better semantic considerations:

MID-SF (Rashmi Gangadharaiah, 2019) which first considers joint multi-intent and slot detection in use of BiLSTMs.

ECA (Chauhan A., 2020) which encodes the dialog context with LSTM for joint task prediction.

KASLUM (Wang et al., 2019) which extracts knowledge from the knowledge base and incorporates dialog history for joint tasks.

CASA (Gupta et al., 2019) which encodes the context with DiSAN sentence2token and we replace it with BERT to demonstrate its contributions.

We also denote several variations of our proposed framework with the following detailed descriptions.

LKA-Dense (Wang et al., 2021b): we use the filtering algorithm depicted in Wang et al. (2021b) to select the knowledge by concatenating each knowledge vector with dialog contexts for scoring.

LKA-Dot: it is the local awareness version of our model. We use the dot-product attention of each knowledge vector and hidden vectors of each word to determine attention weights on their associated knowledge.

GKA: it is the global awareness version of our model. We globally collect knowledge from each word to fuse with dialog contexts as global information for both intent detection and slot filling.

GLKA: it is the global-local awareness version of our model. We globally collect knowledge from each word for slot local attention.

4.3 Implementation details

We adopt the pretrained $\mathbf{BERT}_{base}$ (Devlin et al., 2019) as our utterance encoder. Context attention transformer has $L = 6$-layer attention blocks with 768 head size and 4 attention heads. The max sequence length is 60. We use simple string matching of words to extract relevant knowledge triples from the ConceptNet. Then, TransE (Bordes et al., 2013) is adopted to represent head, relation and tail as 100-dim vectors. We retrieve 5 most related knowledge from each word based on weights assigned on the edges. Both LSTMs have 256 hidden units. We use the batch size of 4 dialogs for MDC and 2 for SGD. In all training, we use Adam optimizer with learning rate as 5e-5. The best performance on validation set is obtained after training 60 epochs on each model. For metrics, we report the dialog act accuracy and slot filling F1 score. Here we only consider a true positive when all BIO values for a slot is correct and forfeit 'O' tags.

5 Main Results

5.1 Main results

Table. 2 shows our main results on the joint task performances of several advanced neural network based frameworks. MID-SF with only LSTMs has relatively inferior performances on both datasets especially in SGD. ECA by taking dialog contexts into consideration has much greater increase in SGD than in MDC and further knowledge induction gives 3.5 % increase in KASLUM. Leveraging BERT-based encoder seems to substantially increase semantic visibility in CASA and our proposed frameworks. Eventually, all of our knowledge-enhanced models beat all baselines both in MDC and substantially in SGD, by more efficiently incorporating external knowledge and dialog contexts with the proposed mutual attention mechanism. Interestingly, LKA with local attention seems to perform substantially well on the slot filling task in SGD dataset, which alludes that the model should rely more on the local knowledge restricted to specific words when making slot decisions in SGD. Knowledge from other words may disturb the attention on some key related knowledge. But for dialog act detection, we could see a performance increase in GLKA which jointly induces global knowledge across the sentence to allow the aggregation of the overall semantics in determining acts.

5.2 Ablation analysis

To better estimate the effectiveness of each module of our best model: GLKA, we conduct ablation experiments in Table. 3. We sequentially ablate

Dataset	**MDC**						**SGD**			
Domain	**Movie**		**Restaurant**		**Taxi**		**Restaurant**		**Flights**	
Model	MDA	SL	MDA	SL	MDA	SL	MDA	SL	MDA	SL
MID-SF	76.56	67.56	77.35	65.77	85.03	70.03	74.26	81.38	84.74	84.48
ECA	77.10	69.72	77.56	66.85	86.61	71.28	87.98	84.87	95.16	87.91
KASLUM	81.86	73.32	80.76	68.36	88.31	74.07	86.81	87.82	92.87	90.05
CASA	84.22	79.59	83.17	74.89	90.00	78.54	92.54	94.20	95.00	91.79
LKA-Dense†	85.25	79.46	83.27	74.89	90.05	79.59	96.84	94.61	97.17	91.14
LKA-Dot†	85.63	80.03	83.69	75.36	**90.95**	79.18	97.70	**96.63**	98.10	**94.02**
GKA†	85.94	80.56	83.64	**75.94**	90.28	79.08	98.44	94.75	98.74	91.71
GLKA†	**86.09**	**80.58**	**84.01**	75.27	90.80	**79.60**	**98.47**	94.86	**99.22**	92.67

Table 2: Experimental Results on several SLU models including our proposed frameworks which are specified in percentage (%). MDA indicates the dialog act detection accuracy by counting corrects when all acts are predicted correctly. SL indicates the slot filling F1 score. † denotes our proposed frameworks for the experiments.

Dataset	**MDC**						**SGD**			
Domain	**Movie**		**Restaurant**		**Taxi**		**Restaurant**		**Flights**	
Model	MDA	SL	MDA	SL	MDA	SL	MDA	SL	MDA	SL
GLKA	**86.09**	**80.58**	**84.01**	**75.27**	**90.80**	**79.60**	**98.47**	**94.86**	**99.22**	**92.67**
w/o KG	86.01	79.92	83.53	74.76	90.56	78.29	97.53	94.83	97.73	92.23
w/o CA	84.87	79.79	81.33	74.68	89.00	78.50	95.88	94.36	97.17	91.94
w/o LSTM	84.57	79.14	82.70	74.35	89.65	79.00	90.96	93.64	94.80	91.33

Table 3: Ablation Results of joint tasks (%) by removing some key components of our proposed frameworks GLKA.

each component from GLKA to observe the performance drops. By removing the knowledge attention module, we can see more obvious reduction in slot filling tasks denoting the necessity of external knowledge in enriching the current word representations. By substituting a unidirectional LSTM on top of BERT for our context attention module (CA), we obtain poorer performance in dialog act detection instead. Finally, we see dialog contexts are more crucial in SGD where drop seems significant by removing all context fusion modules. Overall, we observe dialog act detection relies more on contexts while slot filling tasks may concentrate on inter-utterance relations where external knowledge benefits more instead.

5.3 Further Discussion

Could knowledge amend the data scarcity? We also study how knowledge could contribute to the joint tasks when resources are scarce. Figure. 3 shows the performance changes with different numbers of training data. We found that overall inducing the knowledge will have the positive effect both on dialog act detection and slot detection. When number of training data starts to drop, the performance difference gradually increases especially when we have around 5% of training data. In such few-shot setting, knowledge is beneficial for model to enrich the external information aside from data itself, particularly assisting to capture overall intent semantics. However, knowledge becomes less useful when we have extreme low dataset particularly for slot detection.

Does global knowledge helps non-alphabetic slots? We are interested if knowledge for other words would also help with the slot prediction of the non-alphabetic words. Table. 4 shows the results for each non-alphabetic slot for our local and global attention models. Since there is no knowledge for the non-alphabetic words, we observe an overall 2% increase by inducing global attention. Contexts are beneficial especially for slots associated with rating, money and address, which should be likely inferred by other keywords near them. However, time and zip code are rather independent to contexts which may be disturbed by introducing more irrelevant noises.

5.4 Knowledge Attention

In Figure. 4, we visualize the attention heatmap of tokens with their slot labels vs. all knowledge triples from each token. First, we focus on the rows of the heat map. Without attached knowledge for the words like numbers or punctuations, their attention weights are perceived blank across all tokens in the utterance. Second, for valid attention weights, we found the knowledge corresponding to keywords like *'you', 'with', 'restaurant'* and *'antioch'* are most adopted for overall knowledge representations across all the utterance. It reck-

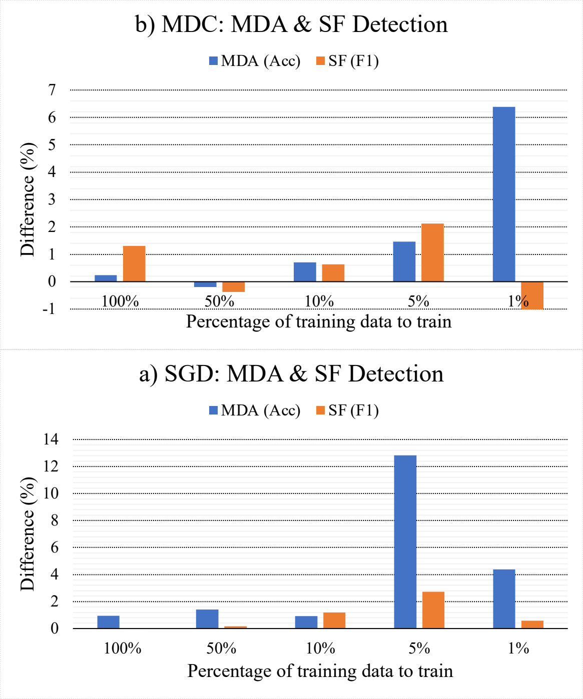

Figure 3: SLU performance by training GLKA with a subsample (%) of the original training data of two datasets: MDC and SGD. We show the results with or without the knowledge induced.

Slot	GLKA (%)	LKA (%)	Δ (%)
address	17.39	0.00	+17.39
price	66.67	50.00	+16.67
critic_rating	34.48	23.08	+11.41
dress_code	50.00	44.44	+5.56
rating	52.17	49.32	+2.86
cost	95.54	95.29	+0.26
numberofpeople	95.63	95.51	+0.12
date	86.96	86.99	-0.02
pricing	42.55	43.14	-0.58
starttime	76.80	77.68	-0.88
numberofkids	73.68	77.78	-4.09
mpaa_rating	76.92	83.33	-6.41
zip_code	77.65	84.44	-6.80
pickup_time	75.19	82.29	-7.09
total	**65.83**	**63.80**	**+2.03**

Table 4: F1 scores for GLKA and LKA of non-alphabetic slots in overall MGD dataset.

ons that the model will highly grasp knowledge in words especially tagged as valued slots (non-O tag) for overall semantic understanding. Interestingly, this collection of knowledge is more emphasized on predicting a word to be non-valued than those words with valued slots. For the columns, we could see for each individual word, for non-valued words, they somehow will accentuate on knowledge of valued words like *'restaurant'* and *'antioch'*, than the knowledge related to itself. It substantiates the belief that the overall semantics of the utterance may be driven by these valued words. For valued words, we instead see a more concentration on their own knowledge to predict specific slots.

In Table. 5, we further illustrate the utterance example with some highlighted words with their extracted knowledge and weights for semantic detection. Here we visualize the knowledge to keywords *'you'*, *'restaurant'* and *'Antioch'*. Here, we take the average of all attention weights across all tokens for the specific knowledge triple; then normalized across the knowledge triples in the same word (head). We could see *'you'* as an object is most adopted to clarify the user being offered and informed counts. Then we observe that the knowledge triple *(restaurant, atl, city)* where *restaurant is at a location of the city* is most recognized to illustrate the relations of restaurant and city tags. Finally, knowledge for *'Antioch'* keyword is mostly relevant to a country which provides additional information when the system may seldom see this word during the training phase. But without further contexts, our model believes *'Antioch'* is more of a part of Turkey.

6 Related Work

Intent detection and slot filling are two main tasks in spoken language understanding (Weld et al., 2021). Many classification-based approaches (Sarikaya et al., 2011; Raymond and Riccardi, 2007; Liu et al., 2017) had been proposed to solve single intent detection problems. On the other hand, hidden markov models (HMM) or conditional random field (CRF) were first to treat slot filling as a tagging problem (Pieraccini et al., 1992) and RNN becomes a popular structure to solve. However, treating two tasks separately may experience error propagation. Liu and Lane (2016) first proposed an attention-based LSTM network to model the correlations between intents and slots. Li et al. (2018a) proposed the gating mechanism for better self-attention on joint tasks. However simply relying on the gate function is not ideal for long sequences. Wang et al. (2018) instead proposed the bi-model to directly model the cross impacts and Zhang et al. (2019) utilized capsule neural networks. Memory networks are also popular choices to model long-range dependency (Wu et al., 2021a). However, a single utterance may have many intents. Rychalska et al. (2018) first proposed hierarchical structures to explore multiple intents. Qin et al.

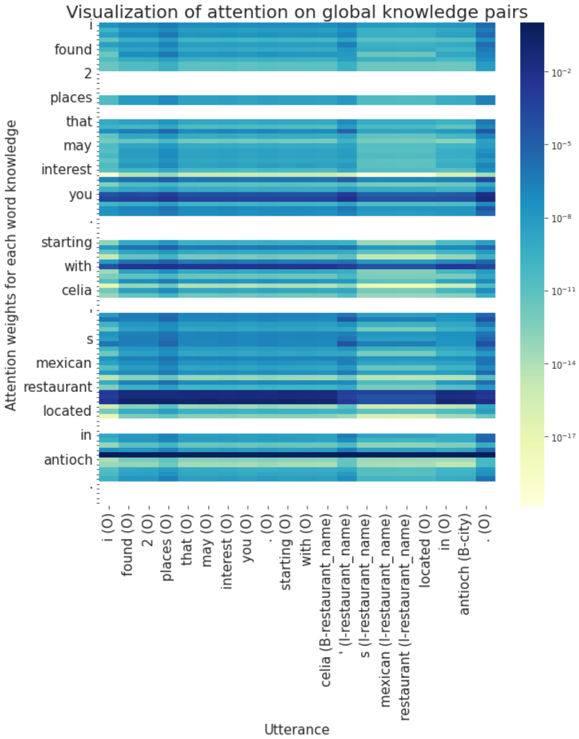

Figure 4: Attention visualization of a single utterance example with respect to all knowledge related to each word. We denote a token followed by their predicted tag in x-axis to construct an utterance. For y-axis, each word will have five knowledge triples with each as a single tick. The blank area is where attention weights are zero.

Utterance Example in Figure 4	
Utterance	I found 2 places that may interest you. Starting with **Celia's Mexican restaurant** located in **Antioch**.
Dialog acts	Offer, Inform Count
Slots	O O O O O O O O O O O **B-res I-res I-res** O O **B-city**
Keyword	**Knowledge**
you	(hc, noun) (0.29), **(hc, object) (0.7)** (rel, guys) (6e-4), (hc, object) (8e-5)
restaurant	(isa, establishment) (8e-9), (atl, hotel) (0.2) (atl, town) (0.14), **(atl, city) (0.65)**
Antioch	(rel, orontes) (4e-5), (rel, swiss) (2e-2) (rel, usa) (5e-2), **(ptof, turkey) (0.9)**

Table 5: The utterance example in Figure 4 of utilizing knowledge for joint task prediction. Knowledge (Relation, Tail) related to three **keywords** as head are presented with their attention weights (number after the knowledge). We only show the top four knowledge adopted for each keyword based on the attention weights. 'hc' represents 'has context', 'rel' represents 'related to', 'atl' represents 'at location' and 'ptof' represents 'part of'.

(2019) proposed a stack-propagation networks to predict intents on each token. Rashmi Gangadharaiah (2019) and (Qin et al., 2020) considered the dynamic interactions between two tasks by jointly detecting multiple intents. Wu et al. (2021b) extended the multiple intent scenario with zero-shot cases. These methods nevertheless restrict their resources to current utterances for prediction.

Contexts and knowledge With respect to dialogs, contexts are also critical for semantic understanding. Bertomeu et al. (2006) first studied the contextual phenomena in words. Bhargava et al. (2013) and Shi et al. (2015) then introduced contextual signals to the joint intent-slot tasks. Advanced hierarchical structures are also emphasized to encode multi-turn dialog contexts efficiently (Chauhan A., 2020; Wang et al., 2019; Gupta et al., 2019; Wu et al., 2021c). Knowledge is also another important resource to induce commonsense for understanding. In task-oriented dialogs, Main emphasis lies in the interaction with task-related knowledge bases (Madotto et al., 2020; Yang et al., 2020). Most of works also focus on open-domain dialog response generation (Zhao et al., 2020; Wang et al., 2021b; Rashkin et al., 2021; Zheng et al., 2021) or task-specific responses (Wang et al., 2021a). Wang et al. (2019) also tried to apply knowledge in SLU but it is not suitable for complex dialog modeling. To amend the gap in modeling knowledge and context interactions of SLU, we follow these previous works' paradigms and explore the mechanisms of characterizing their mutual effects in details.

7 Conclusion

In this paper, we propose a novel BERT-based knowledge augmented network to both consider dialog history and external knowledge in the joint SLU tasks. We propose three different approaches of inducing knowledge awareness, which are capable of selecting relevant knowledge triples and adopt the attention mechanism to acquire useful knowledge representation. We found that our model with local attention (LKA) is useful for slot filling task while the global-local attention (GLKA) reveals its power in dialog act detection. The effectiveness of our proposed model is verified in two multi-turn dialog datasets. We visualize how our models adopt the knowledge from words spreading across the utterance instance to provide better interpretability for decision making. These knowledge fusion vectors could be easily applied to downstream dialog state tracking or management tasks.

References

Leonard Abbeduto. 1983. Linguistic communication and speech acts. kent bach, robert m. harnish. cambridge: M.i.t. press, 1979, pp. xvii 327. *Applied Psycholinguistics*, 4(4):397–407.

Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database QA dialogues: results from a Wizard-of-Oz experiment. In *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8, New York, NY, USA. Association for Computational Linguistics.

A. Bhargava, A. Celikyilmaz, D. Hakkani-Tür, and R. Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8337–8341.

A Bordes, N Usunier, A Garcia-Duran, J Weston, and O Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Singh A. Arora J. Shukla S. Chauhan A., Malhotra A. 2020. Encoding context in task-oriented dialogue systems using intent, dialogue acts, and slots. In *Saini H., Sayal R., Buyya R., Aliseri G. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 103. Springer, Singapore.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Arshit Gupta, Peng Zhang, Garima Lalwani, and Mona Diab. 2019. Casa-nlu: Context-aware self-attentive natural language understanding for task-oriented chatbots.

Changliang Li, Liang Li, and Ji Qi. 2018a. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833, Brussels, Belgium. Association for Computational Linguistics.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018b. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling.

Ting Liu, Xiao DING, Yue QIAN, and Yiheng CHEN. 2017. Identification method of user's travel consumption intention in chatting robot. *SCIENTIA SINICA Informationis*, 47:997.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems.

R. Pieraccini, E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.-H. Lee, and J.G. Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 193–196 vol.1.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features.

Balakrishnan Rashmi Gangadharaiah. 2019. *Joint multiple intent detection and slot labeling for goal-oriented dialog*. Proc. of NAACL.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.

Christian Raymond and Giuseppe Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. In *Proc. Interspeech 2007*, pages 1605–1608.

B. Rychalska, H. Glabska, and A. Wroblewska. 2018. Multi-intent hierarchical natural language understanding for chatbots. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 256–259.

Ruhi Sarikaya, Geoffrey E. Hinton, and Bhuvana Ramabhadran. 2011. Deep belief nets for natural language call-routing. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5680–5683.

Yangyang Shi, Kaisheng Yao, Hu Chen, Yi-Cheng Pan, Mei-Yuh Hwang, and Baolin Peng. 2015. Contextual spoken language understanding using recurrent neural networks.

Ruolin Su, Ting-Wei Wu, and Biing-Hwang Juang. 2021. Act-Aware Slot-Value Predicting in Multi-Domain Dialogue State Tracking. In *Proc. Interspeech 2021*, pages 236–240.

Qingyue Wang, Yanan Cao, Junyan Jiang, Yafang Wang, Lingling Tong, and Li Guo. 2021a. Incorporating specific knowledge into end-to-end task-oriented dialogue systems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yanmeng Wang, Ye Wang, Xingyu Lou, Wenge Rong, Zhenghong Hao, and Shaojun Wang. 2021b. Improving dialogue response generation via knowledge graph filter. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7423–7427.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bimodel based rnn semantic frame parsing model for intent detection and slot filling.

Yufan Wang, Tingting He, Rui Fan, Wenji Zhou, and Xinhui Tu. 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)*, pages 960–967.

H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding.

Jie Wu, Ian Harris, and Hongzhi Zhao. 2021a. Spoken language understanding for task-oriented dialogue systems with augmented memory networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 797–806, Online. Association for Computational Linguistics.

Ting-Wei Wu, Ruolin Su, and Biing Juang. 2021b. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4884–4896, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021c. A Context-Aware Hierarchical BERT Fusion Network for Multi-Turn Dialog Act Detection. In *Proc. Interspeech 2021*, pages 1239–1243.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online. Association for Computational Linguistics.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2019. Joint slot filling and intent detection via capsule neural networks.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent advances and challenges in task-oriented dialog system.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2972–2983, Online. Association for Computational Linguistics.