Infusing Context and Knowledge Awareness in Multi-turn Dialog Understanding

Ting-Wei Wu
Georgia Institute of Technology
Electrical & Computer Engineering
waynewu@gatech.edu

Biing-Hwang Juang
Georgia Institute of Technology
Electrical & Computer Engineering
juang@ece.gatech.edu

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 in natural language understanding (NLU), instead leaving them for updating dialog states only. Moreover, humans usually associate relevant background knowledge with the current dialog contexts to better illustrate slot semantics revealed from word connotations, where previous works have explored such possibility mostly in knowledge-grounded response generation. In this paper, we propose to amend the research gap by equipping a BERT-based NLU framework with knowledge and context awareness. We first encode dialog contexts with a unidirectional context-aware transformer encoder and select relevant inter-word knowledge with the current word and previous history based on a knowledge attention mechanism. Experimental results in two complicated multi-turn dialog datasets have demonstrated significant improvements of our proposed framework. Attention visualization also demonstrates how our modules leverage knowledge across the utterance.

1 Introduction

In conventional task oriented dialog systems, natural language understanding (NLU) 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. Humans usually associate relevant knowledge and previous contexts with current utterance's entities to understand an utterance. Similarly, models' prediction of overall intent semantics and slot values can benefit from act relations such as *'Inform'* may follow *'Request'* acts, and background knowledge which is usually represented as triples in knowledge graphs (Wang et al., 2021a).

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.

However such intuition has not been emphasized when automating NLU tasks. In early attempts of NLU systems, utterances were isolated and analyzed separately for user intents and semantic slots (Raymond and Riccardi, 2007; Liu et al., 2017). Models that maximize the joint distribution likelihood were proposed to allow transitions between two tasks (Liu and Lane, 2016; Wang et al., 2018; Wu et al., 2021a; Li et al., 2018a). While driven by large pretrained corpora, these methods still fall short of employing complete dynamic interactions within dialogs, especially in multiple intent cases (Qin et al., 2019; Rashmi Gangadharaiah, 2019; Qin et al., 2020). Some works have then integrated dialog contexts for more robust NLU (Wang et al., 2019; Gupta et al., 2019; Su et al., 2021; Wu et al., 2021c). However, many of them could not capture dialog flows well with RNN encoders or explain how contexts should affect the slot filling task.

Publicly available models like BERT or XLNet

Findings of the Association for Computational Linguistics: EACL 2023, pages 254–264

provide universal contextualized representations that could be adapted for learning task-oriented contexts. However, it may not give full play to its value when tagging some rare words like *Foxtrot* together with *Tango* as *Movie* in Table 1 that may appear in a domain-specific dataset. One can pretrain these models beforehand emphasizing such phrase relationship which nevertheless tends to be time-consuming and computationally expensive. Therefore, directly integrating external knowledge like a knowledge graph (KG) becomes a more tractable solution (Liu et al., 2019; Zhang et al., 2019b; Wu and Juang, 2022b).

However, there are mainly three challenges lying in the way of such integration: (1) **Heterogeneous information fusion**: the vector space of KG entities is inconsistent with that of the pretrained models. (2) **Knowledge noise**: overwhelming knowledge for models may adversely cause redundant noises for more ambiguity. Many works in knowledge grounded dialog generation has applied term-level denoising (Zheng et al., 2021) or filtering techniques (Wang et al., 2021b) to refine the adopted knowledge for better semantic considerations. (3) **Inter-token knowledge sharing**: Wang et al. (2019) predicts a slot for a given word along with its own associated knowledge. However, real sentences may contain phrases where knowledge between words should be shared to probably enrich the entire utterance semantics. To overcome these challenges and ground knowledge in contextual NLU, which is less explored in the research community, we propose a **C**ontext and **K**nowledge **A**wareness **NLU** Framework (**CKA-NLU**) to effectively incorporate relevant knowledge and dialog history in dialog understanding.

The key ingredients lie in how we can efficiently integrate relevant knowledge and previous history for understanding. We first introduce a context attention module to retrieve context-aware representations. Different from previous works of determining a given word's slot based on its own knowledge, our objectives require models to aggregate both previous dialog contexts and all intra-sentence knowledge facts together to formulate context-attended knowledge vectors in the same space. Such vectors are a weighted combination of all knowledge facts based on the aggregated information until the current turn. We use attention masks and filtering to remove adversarial effects from redundant knowledge noises. Finally we adopt these context-attended vectors for NLU tasks with RNN decoders. Experiment results have shown superior performances of our methods that beat all competitive baselines.

Our contributions are as follows:

1. We propose a novel CKA-NLU framework that incorporates inter-word knowledge with inter-sentence contexts to fill the void of relevant knowledge exploration for important NLU tasks.
2. We demonstrate the benefits of adopting knowledge for token-level slot filling and dialog history for sentence-level intent detection.
3. Experimental and attention visualization results show that our model achieves superior performances over several competitive baselines and demonstrates how our model adopts the knowledge.

2 Problem Formulation

For each utterance $x_n = \{w_1^n, w_2^n, \ldots, w_T^n\}$ 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 $\{s_1^n, s_2^n, \ldots, s_T^n\}$ 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 $\mathbf{C_n} = \{x_1, \ldots, x_{n-1}\}$ 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 represented as triples (head h, relation r, tail t) and $\phi(\cdot)$ helps to extract related knowledge pairs for x_n (§3.2.1). 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 allow information flow across the dialog, we first encode the entire dialog with a token-level BERT (Devlin et al., 2019) encoder and a turn-level context-aware transformer encoder. Instead of concatenating all sentences which may cause an extreme sequence length, we first generate the token-level representations $\mathbf{H} = \{h_1, h_2, \ldots, h_N\}$

[1] Dialog acts and intents are equivalent and interchangeably used in this paper.

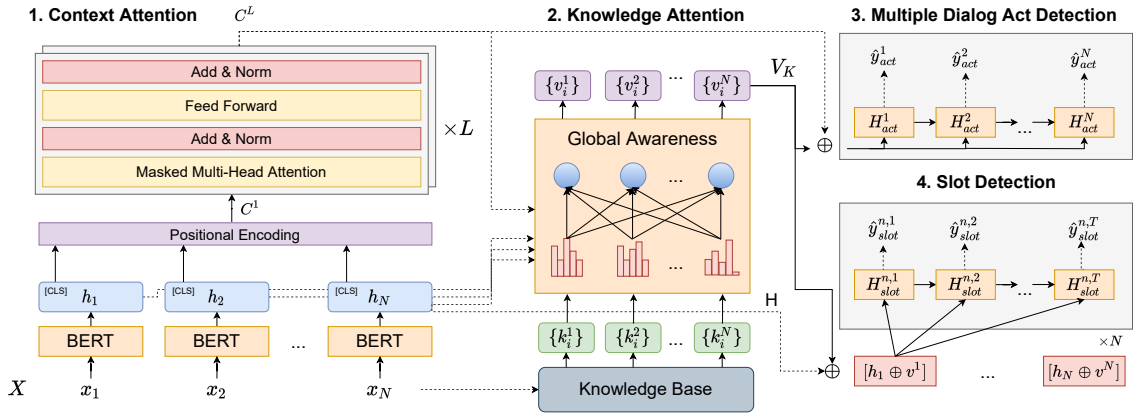

Figure 1: Illustration of our proposed framework for joint dialog act detection and slot filling in multi-turn dialogs. It consists of context and knowledge attention modules, and two LSTM-based decoders. The utterance-level representations will be encoded with the context attention module and token-level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

for each utterance x_n in a dialog X by taking vectors from each `[CLS]` token. During testing at turn n, we may directly reuse these calculated representations $\{h_1, h_2, \ldots, h_{n-1}\}$ until turn $n-1$.

In contrast with other contextual NLU (Wang et al., 2019; Gupta et al., 2019) with hierarchical components, we introduce a GPT-like unidirectional transformer encoder with the hidden size H_a to encode $\mathbf{H} \in \mathbb{R}^{N \times H_b}$. It consists L layers of masked multi-head self-attention (MHA), point-wise feed forward network (FFN), residual sublayer and layer normalization. The future time steps are masked for training since we will not have access to future utterances during testing. We will send $\mathbf{H}$ as the first layer input $\mathbf{C^1}$ and iteratively encode it with two sublayers in Eq 2. Each head $\mathbf{C_i} \in \mathbb{R}^{N \times (H_a/h)}$ will be first mapped into a query $\mathbf{C^Q}$, a key $\mathbf{C^K}$ and a value $\mathbf{C^V}$ which participate in the multi-head self-attention. Here $f(\cdot)$ is softmax function. Finally, we will obtain the final contextual 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

Humans could naturally associate contexts with relevant knowledge to predict semantics. Here we elaborate on how we can leverage current contexts $\mathbf{C^L} = \{c_n^L\}$ and a relevant knowledge base K_G to induce the intents and slots for each utterance x_n.

3.2.1 Knowledge extraction

The first step is to gather all necessary knowledge triples $\gamma = \{h, r, t\}$, which are head h and tail t entities with their relation r, related to the current utterance $x_n = \{w_1^n, w_2^n, \ldots, w_T^n\}$. For each word w_i^n, we first retrieve a list of triples with the exactly same head entity being w_i^n from a knowledge base K_G. If no head entities are matched, we instead seek entities that has a substring of w_i^n. Each triple in the pretrained K_G (Bordes et al., 2013) has a pre-given relation weight $w_r \in [0, 1]$. For each w_i^n, we select $|K|$ triples that have the largest $|K|$ weights as the final word-level knowledge k_i^n. We will finally obtain a T length knowledge sequence $\mathbf{K_n} = \{k_1^n, k_2^n, \ldots, k_T^n\}$ gathered from each word w_i^n. In case of non-alphabetic or out-of-vocabulary (OOV) words with no match in K_G, we instead replace their $\mathbf{K_n}$ as zero vectors to represent agnosticism of knowledge.

3.2.2 Global awareness

To improve the heterogeneous information fusion between contexts and knowledge, after obtaining the knowledge sequence $\mathbf{K_n} = \{k_i^n\}$ (i.e. total $T \times |K|$ triples $\gamma = \{h, r, t\}$), we aim to obtain the context-attended knowledge sequence $\mathbf{V_K} = \{v_i^n\}$ by selecting the most appropriate knowledge (i.e., removing redundant knowledge noise) within the entire sentence, given each word w_i^n and its previous dialog history c_n^L. Different from the term-level denoising like Zheng et al. (2021) and Wang et al. (2019), to allow phrase-level knowledge sharing,

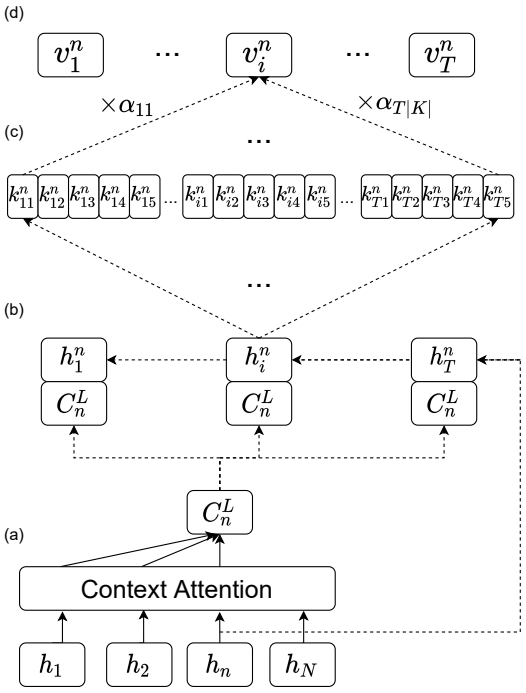

Figure 2: Knowledge Attention Diagram. (a) Context Attention module will first process the dialog history and produce context-aware vectors for each utterance. (b) Token-level representation will be concatenated with the context-aware vector. (c) The fused vector will be used to calculate the attention weights for every knowledge vector in the utterance. (d) The final context-attended knowledge vector will be the weighted combination of all knowledge vectors.

for each word, we aim to globally select all related knowledge in the sentence after seeing previous turns c_n^L. This will allow us to possibly consider knowledge of words in the same phrase.

Shown in Figure 2, we calculate the vector v_i^n where r_{ij}^n, t_{ij}^n are j-th relation and j-th tail entity vectors for the word w_i^n. $\mathbf{W^H}, \mathbf{W^R}, \mathbf{W^T}$ are learnable matrices during training. $[;]$ is the concatenation of two vectors:

$$v_i^n = \sum_{i=1}^{T} \sum_{j=1}^{|K|} \alpha_{ij} [r_{ij}^n; t_{ij}^n] \quad (5)$$

$$\alpha_{ij} = exp(\beta_{ij}) / \sum_{i'=1}^{T} \sum_{j'=1}^{|K|} exp(\beta_{i'j'}) \quad (6)$$

$$\beta_{ij} = (\tilde{h}_i^n \mathbf{W^H})(tanh(r_{ij}^n \mathbf{W^R} + t_{ij}^n \mathbf{W^T}))^T \quad (7)$$

$$\tilde{h}_i^n = [h_i^n; c_n^L] \quad (8)$$

We first concatenate the token-level representations for each word h_i^n in the utterance x_n with its context vector c_n^L, which entails the embedded information from previous turns (Eq 8). Then we use $\tilde{h}_i^n$ to calculate the attention weight α_{ij} with any of the knowledge (r_{ij}^n, t_{ij}^n) related to this utterance (Eq 6, 7). Eventually, we linearly combine all knowledge vectors together to formalize the context-attended knowledge vector v_i^n (Eq 5). Additionally, to avert the noise from zero-vectors of non-alphabetic word knowledge, we introduce an attention mask to calculate α_{ij} only on the non-zero knowledge vectors.

3.3 Semantic Decoder

After obtaining the context-attended knowledge $\mathbf{V_K} = \{v_i^n\}$, context vectors $\mathbf{C^L}$ and initial token-level vectors $\mathbf{H}$, we adopt two BiLSTMs to predict multiple dialog acts and slots which exhibit the sequential information in BIO scheme.

$$\mathbf{H_{act}} = \mathbf{BiLSTM}([\tilde{\mathbf{H}}; \mathbf{V_K}]) \quad (9)$$

$$\mathbf{H_{slot}} = \mathbf{BiLSTM}([\mathbf{H}; \mathbf{V_K}]) \quad (10)$$

For dialog act detection, we concatenate $\mathbf{V_K}$ with the fused context $\tilde{\mathbf{H}} = ([\mathbf{H}; \mathbf{C^L}])\mathbf{W^H}$ from the attention mechanism and serve as the inputs of BiLSTM. For slot filling, since the task focuses more on token-level information for decision, we only concatenate raw token-level representations and $\mathbf{V_K}$ to be inputs of another BiLSTM, which empirically works better. 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 11, 12. Finally, the joint objective is formulated as the sum of $\mathcal{L}_a$ and $\mathcal{L}_s$.

$$\mathcal{L}_a \triangleq -\sum_{n=1}^{N} \sum_{a=1}^{|\mathcal{Y}^a|} (y_a^n log(\hat{y}_a^n) + (1 - y_a^n) log(1 - (\hat{y}_a^n)) \quad (11)$$

$$\mathcal{L}_s \triangleq -\sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{|\mathcal{Y}^s|} (y_s^{(n,t)} log(\hat{y}_s^{(n,t)})) \quad (12)$$

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. To compare the relevant knowledge usage in different domains and save computational resources, we randomly select 1k dialogs for each domain in MDC and two restaurant and flights domains from SGD for total 5k dialogs in 6:1:3 train, validation, test ratio. For SGD, Restaurant domain is chosen to compare with that of MDC and Flights domain is the one not existing in MDC. 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:

- **MID-SF** (Rashmi Gangadharaiah, 2019) considers joint multi-intent and slot detection in use of BiLSTMs.
- **ECA** (Chauhan A., 2020) encodes the dialog context with LSTM for joint tasks.
- **KANLUM** (Wang et al., 2019) extracts knowledge from the knowledge base and incorporates dialog history for joint tasks.
- **ERNIE** (Zhang et al., 2019b): We take ERNIE backbone to integrate knowledge entities and take the token and entity outputs for intent detection and slot filling directly.
- **LABAN** (Wu et al., 2021b) leverages label information to construct a latent semantic space for utterance projection. It is mainly for the multiple intent detection task only.
- **CASA-BERT** (Gupta et al., 2019) encodes the context with sentence2token and DiSAN which we replace with BERT for fair comparison with other BERT-based models.

We also perform several variations of our proposed framework to conduct the ablation study with the following detailed descriptions.

- **Less-Relevant knowledge triples (LR-KA)**: We replace the top $|K|$ knowledge triples with the less related knowledge triples ranked from $|K| \sim 2|K|$ (from relation weights in K_G) to perform sensitivity analysis on the quality of knowledge.
- **Word-Level knowledge attention (WL-KA)**: We use the attention-based filter (AF) (Wang et al., 2021b) to perform token-level knowledge attention instead of sentence-level attention in our framework.
- **Transformer decoder (Trans)**: We replace the semantic decoder (§ 3.3) with a transformer decoder to both predict dialog acts and slots.

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 ConceptNet knowledge base (Speer et al., 2018) to obtain relevant knowledge for attention. It involves many crowdsourced and expert-created resources like DBPedia, OpenCyc and WordNet with 1.5M word entities connected with weighted edges (relation). Each word or relation is represented as a dense 100-dim vectors by adopting TransE (Bordes et al., 2013) learning mode. Each knowledge also contains an ExternalURL to represent the external source. We retrieve $|K| = 5$ 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 2 dialogs for MDC and 1 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 30 epochs on each model. For metrics, we report the dialog act accuracy (exact match) 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 performance. MID-SF with only LSTMs has relatively inferior performance on both datasets especially in SGD. ECA by taking dialog contexts into consideration has much greater increase in SGD than in MDC. ERNIE and KANLUM have better slot filling performance which suggests the importance of further knowledge induction. Leveraging BERT-based encoder seems to substantially increase semantic visibility in ERNIE, CASA-BERT and our proposed framework, while introducing dialog contexts additionally gives better dialog act detection performance in CASA-BERT and our model. Eventually, our proposed framework beats all baselines both in MDC and substantially in SGD, by more

Dataset	MDC						SGD			
Domain	Movie		Restaurant		Taxi		Restaurant		Flights	
Model	MDA	SL	MDA	SL	MDA	SL	MDA	SL	MDA	SL
MID-SF (Rashmi Gangadharaiah, 2019)	76.56	67.56	77.35	65.77	85.03	70.03	74.26	81.38	84.74	84.48
ECA (Chauhan A., 2020)	77.10	69.72	77.56	66.85	86.61	71.28	87.98	84.87	95.16	87.91
KANLUM (Wang et al., 2019)	81.86	73.32	80.76	68.36	88.31	74.07	86.81	87.82	92.87	90.05
ERNIE (Zhang et al., 2019b)	81.52	79.18	80.60	74.68	87.72	76.85	88.53	91.37	89.33	90.50
LABAN (Wu et al., 2021b)	82.05	-	82.28	-	88.19	-	90.51	-	94.23	-
CASA-BERT (Gupta et al., 2019)	84.22	79.59	83.17	74.89	90.00	78.54	92.54	94.20	95.00	91.79
CKA-NLU	**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 NLU 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. † indicates the significant improvement of p-value < 0.05, compared with CASA-BERT.

Dataset	MDC						SGD			
Domain	Movie		Restaurant		Taxi		Restaurant		Flights	
Model	MDA	SL	MDA	SL	MDA	SL	MDA	SL	MDA	SL
CKA-NLU	**86.09**	**80.58**	**84.01**	75.27	**90.80**	79.60	**98.47**	**94.86**	**99.22**	**92.67**
w/ LR-KA	85.63	80.26	83.43	**75.76**	89.77	**80.03**	98.38	94.31	98.93	91.99
w/ WL-KA	85.25	79.46	83.27	74.89	90.05	79.59	96.84	94.61	97.17	91.14
w/ Trans	85.98	79.94	83.27	75.19	90.40	78.33	97.35	94.34	98.20	91.95
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 model: CKA-NLU.

efficiently incorporating external knowledge and dialog contexts with the proposed global awareness attention mechanism.

5.2 Ablation analysis

To better estimate the effectiveness of each module of our best model, we conduct ablation experiments in Table 3. We ablate or replace each component from CKA to observe the performance drops. First, we could see knowledge quality may affect the performance of joint tasks where most performance drops are observed with LR-KA, while we found that slot accuracy may increase if the overall extracted knowledge is less relevant to utterances. To note, the word matching accuracies in the knowledge base are 78.12% (MDC) and 80.97% (SGD), which indicates that there is still about 20% of zero vectors introduced as redundant noises. Second, considering global knowledge across the entire sentence has overall better performance than only word-level knowledge, where knowledge of some phrases should be treated jointly. Finally, we see a single transformer decoder may still entangle the act and slot information by updating gradients simultaneously with poorer performance.

By removing the entire knowledge attention module, we could see a larger accuracy decrease in slot filling tasks, denoting the necessity of external knowledge in enriching the current word representations. By substituting a LSTM on top of BERT for our context attention module (CA), we obtain poorer performance in dialog act detection. By replacing two LSTMs with fully connected layers after knowledge attention, the performance drops especially in SGD. 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 inducing the knowledge will have the positive effect on both tasks. In the few-shot setting, we see the performance difference enlarges where knowledge becomes beneficial to enrich the external information aside from data itself. However, knowledge becomes less useful when we have extreme low dataset particularly for slot detection in MDC. Introducing more MDC data at a certain point may contradict with the external knowledge data base that possibly makes models hard to generalize, while it helps dialog act detection that amends the training instability from data scarcity.

Does global knowledge help non-alphabetic slots? We are interested if knowledge for other words would also help with the slot prediction of

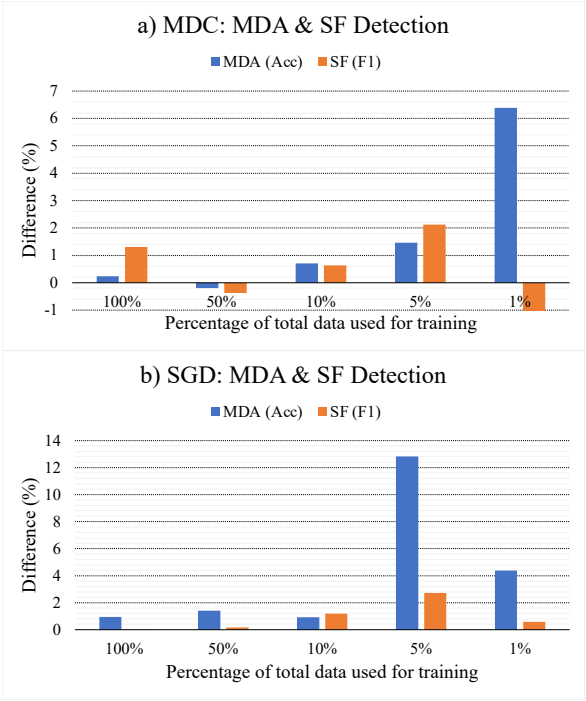

Figure 3: NLU performance gain by using knowledge in CKA-NLU with a subsample (%) of the original training data of two datasets: MDC and SGD.

non-alphabetic words. Table 4 shows the results for each non-alphabetic slot for our global and local 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, introducing more knowledge noises may not help to predict time and zip code since they are rather independent to contexts.

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 elucidates that the model will mostly 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 non-valued words, they will rely 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.

Slot	CKA-NLU (%)	WL-KA (%)	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 of non-alphabetic slots in overall SGD dataset when using all (CKA-NLU) or word-level (WL-KA) knowledge.

In Table 5, we further show an utterance example with some highlighted words including *'you'*, *'restaurant'* and *'Antioch'* with their extracted knowledge and weights for semantic detection. We take the average of all attention weights across all tokens for that 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 is conducive when the system seldom sees this word during training. 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 NLU tasks (Weld et al., 2021). Many classification or clustering approaches (Sarikaya et al., 2011; Raymond and Riccardi, 2007; Liu et al., 2017; Wu and Juang, 2022a) had been proposed for single intent detection. However, treating two tasks separately may experience error propagation. Liu and Lane

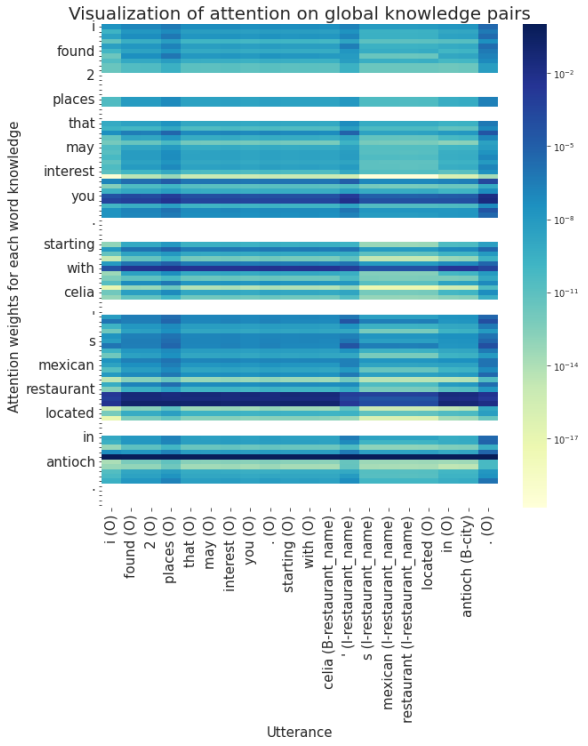

Figure 4: Attention visualization of a single utterance example with respect to all knowledge related to each word. We denote an utterance with tokens followed by their predicted tag in x-axis. For y-axis, each word will have five knowledge triples with each as a single tick. The blank area is where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

(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, which is not scalable for longer sequences. Wang et al. (2018) instead proposed the bi-model to directly model the cross impacts and Zhang et al. (2019a) 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. Qin et al. (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 where we consider the multi-turn dialogs jointly where dialog acts could be context-sensitive (Bothe et al., 2018).

Contexts and knowledge Contexts are also critical for dialog 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. It is widely adopted for knowledge-enhanced pretraining to enrich representations (Liu et al., 2019; Zhang et al., 2019b). 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). However, commonsense knowledge is seldom adopted in NLU. Wang et al. (2019) tried to apply knowledge in NLU but it is not suitable for complex dialog modeling. To amend the gap in modeling such knowledge and context interactions, we follow these previous works' paradigms and explore the mechanisms of characterizing their mutual effects.

7 Conclusion

In this paper, we propose a novel BERT-based knowledge-augmented network to effectively incorporate dialog history and external knowledge in the joint NLU tasks. Compared to recent works

which consider only intra-word knowledge, we instead raise the knowledge awareness by selecting all relevant knowledge triples in an utterance with the current dialog contexts. We found that our framework is verified to be effective in two complex multi-turn dialog datasets where contexts and knowledge are crucial in dialog act detection and slot filling respectively. The visualization shows that our models adopt some key knowledge in particular words and learn to grasp useful information for better interpretability. These context-attended knowledge vectors could be easily applied to downstream dialog state tracking or management tasks.

Limitations

The possible limitations for our works are twofolds. First, the scalability of our method is subject to the size of the knowledge base and the number of incorporated knowledge since selecting from larger knowledge candidates may require more computational memory and training time but with higher performance. Exact string matching between context words and knowledge entities is relatively simple and could be replaced with more advanced semantic matching techniques, which nevertheless may increase model complexity. Second, depending on the domains of datasets to apply, too many out-of-vocabulary words (OOV) with no match in the knowledge base may affect the model performance and our future works will investigate a better solution to replace zero-vectors that are associated with non-alphabetic words.

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.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019. K-bert: Enabling language representation with knowledge graph.

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.

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.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge.

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 bi-model 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 and Biing Juang. 2022a. Induce Spoken Dialog Intents via Deep Unsupervised Context Contrastive Clustering. In *Proc. Interspeech 2022*, pages 1081–1085.

Ting-Wei Wu and Biing-Hwang Juang. 2022b. Knowledge augmented bert mutual network in multi-turn spoken dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7487–7491.

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. 2019a. 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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. Ernie: Enhanced language representation with informative entities.

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.

A Additional Experimental Setting

We use huggingface transformers (Wolf et al., 2020) to implement our framework and we use two Nvidia 2080Ti GPUs for all model training. The number of model parameters is around 146M. It takes 30 minutes to train 30 epochs for a single model.