KNOWLEDGE AUGMENTED BERT MUTUAL NETWORK IN MULTI-TURN SPOKEN DIALOGUES

Ting-Wei Wu[1], Biing-Hwang Juang[1]

[1]Georgia Institute of Technology
Department of Electrical and Computer Engineering
waynewu@gatech.edu, juang@ece.gatech.edu

ABSTRACT

Modern spoken language understanding (SLU) systems rely on sophisticated semantic notions revealed in single utterances to detect intents and slots. However, they lack the capability of modeling multi-turn dynamics within a dialogue particularly in long-term slot contexts. Without external knowledge, depending on limited linguistic legitimacy within a word sequence may overlook deep semantic information across dialogue turns. In this paper, we propose to equip a BERT-based joint model with a knowledge attention module to mutually leverage dialogue contexts between two SLU tasks. A gating mechanism is further utilized to filter out irrelevant knowledge triples and to circumvent distracting comprehension. Experimental results in two complicated multi-turn dialogue datasets have demonstrate by mutually modeling two SLU tasks with filtered knowledge and dialogue contexts, our approach has considerable improvements compared with several competitive baselines.

Index Terms— Multi-turn Dialogues, Slot Filling, Knowledge base, BERT, Context

1. INTRODUCTION

Recent advances of spoken language understanding (SLU) modules prompt the success of task oriented dialogue systems, in transforming utterances into structured and meaningful semantic representations for dialogue management [1, 2]. It mainly detects associated dialogue acts or intents and extracts key slot information as so-called *'semantic frames'* [3], shown in Table 1. Some knowledge triples in a knowledge base may be related to specific keywords in the dialogue which may accelerate the understanding process.

In early attempts of SLU tasks, isolated utterances in dissected dialogues were analyzed separately for user intents and semantic slots [4, 5]. 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 [6, 7, 8], with most studying the benefits of intent information for the later slot filling task. Some works also predicted multiple intents

Table 1. Snippet of a single turn within a dialogue with corresponding dialogue acts, slots and knowledge samples related to **keywords** in the utterance.

Speaker	Utterance
1. User	Is there something that's maybe a good intelligent **comedy**?
Act & Slots:	*Request (genre: **comedy**)*
Knowledge:	*(**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 constraints: in your **area**)* *Confirm_question*
Knowledge:	*(**foxtrot**; related to; dance)* *(**area**; is a; region)*

[9, 10, 11]. While driven by large pretrained corpus, these methods still fall short of employing complete dynamic interactions within dialogues. In contrast, humans can naturally adopt history contexts to identify intentions with their background knowledge. Some works have integrated previous dialogue contexts for more robust SLU [12, 13, 14].

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 dialogue generation [15, 16, 17], where relevant documents or a knowledge base auxiliarily guide the language autoregressive progress. Term-level denoising [17] or filtering techniques [15] 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 dialogue context understanding, which recent SLU works often neglect. [13] has proposed to adopt knowledge attention for joint tasks. However, it adopts a single LSTM layer to couple all knowledge without filtering
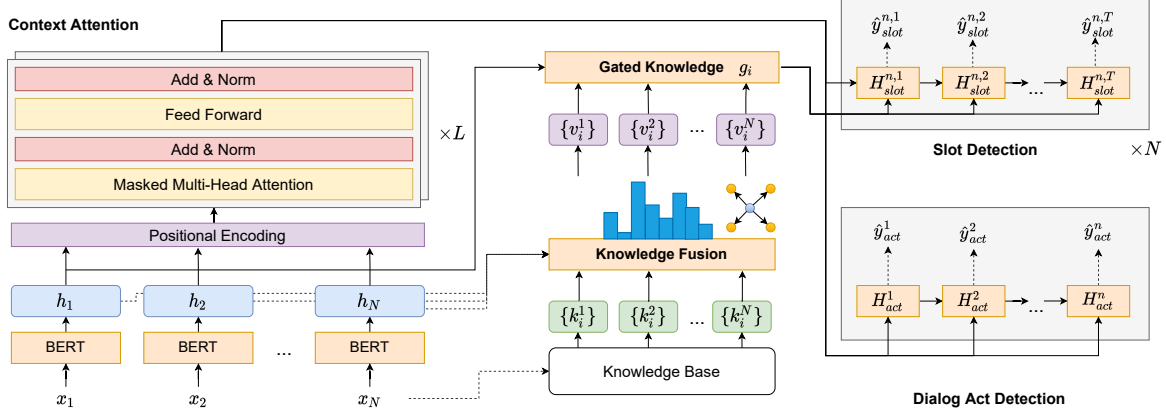
Fig. 1. Illustration of our proposed framework for joint dialogue act detection and slot filling in multi-turn dialogs.

and contexts, which cannot model complex interactions well.

To solve above concerns, we propose a new **K**nowledge **A**ugmented **BER**T **M**utual Network (KABEM) to effectively incorporate dialogue history and external knowledge in joint SLU tasks. Encoded knowledge is further gated to abate useless information redundancy. We then respectively induce dialogue contexts and knowledge to mutually predict intents and slots coherently with two LSTM decoders. Experiment results have shown superior performance of our methods in manipulating contexts and knowledge for joint tasks and beat all competitive baselines. Our contributions are as follows:

1. We propose KABEM to incorporate external knowledge and previous dialogue history for joint multiple dialogue act and slot filling detection, where previous SLU works usually isolate the utterances without knowledge grounded.

2. We demonstrate the effectiveness of knowledge attention and the gating mechanism to reinforce the knowledge transitions between dialogue act and slot detection.

3. Experimental results show that our model achieves superior performances over several competitive baselines with more comprehensive knowledge consideration.

2. METHODOLOGY

2.1. Problem Statement

In a dialogue $X = \{x_1, \ldots, x_N\}$ of total N user utterances and system responses, we would like to detect one or more dialogue acts A and slots S for each x_n. We denote the dialogue history $C_n = \{x_1, \ldots, x_{n-1}\}$ and associated knowledge $K_n = \phi(K_G, x_n)$ for the current utterance x_n. K_G is an external large knowledge base with knowledge triples and $\phi(\cdot)$ is the filter function. In essence, the joint probability distributions of predicting dialogue acts and slot labels are given as $A, S = arg \max P(A, S|x_n, C_n, K_n)$. For an utterance of T words $x_n = \{w_1^n, w_2^n, \ldots, w_T^n\}$, we will finally obtain a corresponding dialogue act set $\{a_i\}$ and a sequence of slot tags $\{s_1^n, s_2^n, \ldots, s_T^n\}$.

2.2. Context Attention

To fully leverage the dialogue context information, we propose to encode the dialogue at token and turn levels respectively. At token level, we adopt BERT [18], a powerful NLP representation model, to extract semantic representations. For each utterance x_n in a dialogue X, we encode it with BERT and obtain token-level representations $H = \{h_1, h_2, \ldots, h_N\}$ from [CLS] tokens for N utterances.

At turn level, to better capture semantic flows within a dialogue, we further encode H with a context-aware unidirectional transformer encoder [19], 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 $H \in R^{N \times H_b}$ as the first layer input C^1 and iteratively encode with two sublayers in Eq. 1. For each layer, it will first project the input C with weight matrices: $W^Q, W^K, W^V \in R^{H_b \times H_a}$ to be $C^Q = CW^Q$, $C^K = CW^K$, $C^V = CW^V$. Then each of them will be separated into h heads, with each head i to be $C_i \in R^{N \times (H_a/h)}$, H_a is the hidden size for the attention module and H_b is BERT hidden size. These C_i will be sent into a self-attention and a feed forward layer in Eq.2 and Eq.3. Finally, we will obtain the final contextual dialogue representations C^L.

$$C^l = FFN(MHA(C^{l-1}, C^{l-1}, C^{l-1})) \quad (1)$$

$$MHA(C_i^Q, C_i^K, C_i^V) = softmax(\frac{C_i^Q (C_i^K)^T}{\sqrt{H_b}})C_i^V \quad (2)$$

$$FFN(x) = max(0, \ xW_1 + b_1)W_2 + b_2 \quad (3)$$

2.3. Knowledge Fusion

To simulate the human awareness of coherently relating current contexts to background knowledge, the knowledge subgraph k_i^n corresponding to the i-th word w_i^n in n-th utterance x_n is retrieved from the knowledge base K_G using similar word matching. Each k_i^n is a collection of multiple related

triples $\gamma = \{h, r, t\}$, as head entity, relation, and tail entity. For each word, we then adopt an attention mechanism to dynamically filter irrelevant knowledge triples based on word contexts and obtain the knowledge-aware vector v_i^n.

$$v_i^n = \sum_{j=1}^{M} \alpha_{ij}[r_{ij}; t_{ij}] \qquad (4)$$

$$\alpha_{ij} = exp(\beta_{ij}) / \sum_{m=1}^{M} exp(\beta_{im}) \qquad (5)$$

$$\beta_{ij} = (h_i^n W^H)(tanh(r_{ij}W^R + t_{ij}W^T))^T \qquad (6)$$

r_{ij}, t_{ij} are relation and tail entity vectors. W^H, W^R, 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.

2.4. Gated Knowledge

Knowledge triples are mostly associated with name entities, where stochastic numbers or dates mentioned in utterances may not be relevant. We instead replace the triple vectors 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 induced for downstream tasks and prevent information from overloading.

$$h_i^{n\prime} = g_i \cdot h_i^n + (1 - g_i) \cdot v_i^n \qquad (7)$$

$$g_i = \sigma(W_i[h_i^n; v_i^n] + b_i) \qquad (8)$$

Information from word hidden states and corresponding knowledge is introduced in a trainable fully-connected layer with a sigmoid layer to produce a knowledge gated score. Then the network will balance the degree of knowledge influencing the decoding outputs.

2.5. Semantic Decoder

After obtaining the knowledge-enriched representations $H_K = \{h_i^{n\prime}\}$ along with contextual dialogue representations C^L, we adopt a BiLSTM for slot filling and a LSTM to detect multiple dialogue acts mutually. It will allow information to dynamically flow between two networks for understanding.

$$H_{slot} = BiLSTM(H_K, C^L) \qquad (9)$$

$$H_{act} = LSTM(C^L) \qquad (10)$$

Knowledge-enriched vectors H_K will be the inputs of BiLSTM with C^L as initial hidden states, where contexts will assist the slot prediction at each knowledge-enhanced time step. At the same time, we also input dialogue contexts C^L only to another unidirectional LSTM for dialogue act detection since our context attention module is shared and has learned H_K information implicitly. Finally, we can generate logits $\hat{y}_{act} = \sigma(H_{act}W_{act})$ by transforming H_{act} with $W_{act} \in R^{H_L \times |\mathcal{Y}^a|}$ and a sigmoid function σ. H_L is LSTM hidden size and $|\mathcal{Y}^a|$ is the size of dialogue act set. Likewise, we compute $\hat{y}_{slot} = softmax(H_{slot}W_{slot})$. Total loss will be the combination between the binary cross entropy loss based on $\hat{y}_{act}$ and the cross entropy loss based on $\hat{y}_{slot}$.

3. EXPERIMENTS

3.1. Experimental setup

We evaluate our proposed framework on two large-scale dialogue datasets, i.e. Microsoft Dialogue Challenge dataset (MDC) [21] and Schema-Guided Dialogue dataset (SGD) [22]. **MDC** contains human-annotated conversations in three domains (movie, restaurant, taxi) with total 11 dialogue acts and 50 slots. **SGD** entails dialogues over 20 domains ranging from travel, weather to banks etc. It has more structured annotations with total 18 dialogue acts and 89 slots. We randomly select 1k dialogues for each domain in **MDC** and the restaurant domain from **SGD** to compare that in MDC and a very different domain (flights) for total 5k dialogues in 7:3 training and testing ratio. Each utterance is labeled with one or more dialogue acts and several slots.

We compare our models with several competitive baselines which sequentially include more semantic features: **MID-SF** [10] which first considers multi-intent detection with slot filling tasks with BiLSTMs. **ECA** [20] which encodes the dialogue context with a LSTM encoder for joint tasks. **KASLUM** [13] which extracts knowledge from a knowledge base and includes dialogue history for joint tasks. **CASA** [14] which encodes the context with DiSAN sentence2token and we replace BERT encoder to demonstrate its contributions. **KABEM**$_{AF}$ [15] we replace only Knowledge Fusion part in KABEM (§ 2.3) with the attention-based filter (AF) in [15] to compare different knowledge attention.

We adopt the pretrained $BERT_{base}$ [18] 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 [23] 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 dialogues 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.

Table 2. Experimental Results on several SLU models and ablation study of KABEM (%). ID (Acc) indicates the dialogue act detection accuracy when all acts are predicted correctly. SL (F1) indicates the slot filling F1 score.

Dataset	MDC						SGD			
Domain	Movie		Restaurant		Taxi		Restaurant		Flights	
Model	ID (Acc)	SL (F1)	ID (Acc)	SL (F1)	ID (Acc)	SL (F1)	ID (Acc)	SL (F1)	ID (Acc)	SL (F1)
MID-SF [10]	76.56	67.56	77.35	65.77	85.03	70.03	74.26	81.38	84.74	84.48
ECA [20]	77.10	69.72	77.56	66.85	86.61	71.28	87.98	84.87	95.16	87.91
KASLUM [13]	81.86	73.32	80.76	68.36	88.31	74.07	86.81	87.82	92.87	90.05
CASA [14]	84.22	79.59	83.17	74.89	90.00	78.54	92.54	94.20	95.00	91.79
KABEM$_{AF}$ [15]	85.25	79.46	83.27	74.89	90.05	**79.59**	96.84	94.61	97.17	91.14
KABEM	85.63	**80.03**	**83.69**	**75.36**	**90.95**	79.18	**97.70**	**96.63**	**98.10**	**94.02**
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

4. RESULTS AND ANALYSIS

4.1. 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 with dialogue contexts enhanced 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 KABEM. Eventually, KABEM$_{AF}$ and KABEM beat all baselines both in MDC and substantially in SGD, while our knowledge fusion module incorporates external knowledge and dialogue contexts more efficiently.

To better estimate the effectiveness of each module of KABEM, we conduct ablation experiments following in Table. 2. We sequentially ablate each component from KABEM to observe the performance drops. By removing knowledge attention with gating (KG), we see more obvious reduction in slot filling tasks denoting the necessity of external knowledge. By substituting a unidirectional LSTM on top of BERT for our context attention module (CA), we obtain poorer performance in dialogue act detection instead. Finally, we see dialogue contexts are more crucial in SGD where drop seems significant by removing all context fusion modules. Overall, we observe dialogue act detection relies more on contexts while slot filling tasks may concentrate on inter-utterance relations where external knowledge benefits more instead.

4.2. Knowledge attention

In Table. 3, we visualize the extracted knowledge and their weights corresponding to three important keywords for semantic detection in the utterance. Here, the word 'cheap' is super related to 'affordable' which helps identifying the slot 'pricing'. Our model also leverages the fact of 'yesterday' and

Table 3. A utterance example of utilizing knowledge for joint task prediction. Knowledge (Relation, Tail) related to three **keywords** as head are presented with their attention weights. 'rel' represents 'related to' and 'ant' represents 'antonym'.

Utterance Example		
Utterance	I need a **cheap** food place for 3 people **tomorrow** at 1pm in **Seattle**.	
Dialog acts	Request	
Slots	O O O **B-pricing** O O O B-numberofpeople O **B-date** O B-starttime I-starttime O **B-city**	
Knowledge		
cheap	**tomorrow**	**Seattle**
rel, affordable (0.99)	rel, later_on (5e-2)	rel, city_usa (2e-2)
rel, chintzy (3e-7)	rel, morrow (7e-3)	rel, washington (1e-4)
rel, chinchy (2e-9)	is a, future (9e-7)	rel, emerald_city (9e-2)
rel, twopenny (5e-5)	is a, day (4e-6)	part of, wa (0.87)
rel, gimcrack (8e-6)	ant, yesterday (0.9)	is a city_wa (8e-3)

'tomorrow' to identify a 'date' slot. Eventually, Knowledge related to 'city' assists the city identification for 'Seattle', especially beneficial when model has never seen 'Seattle' in the training data. To notice, numbers or time are not valid entities inside the knowledge base, where equal weights are assigned to each zero vector and our gating mechanism will circumvent from using it for prediction.

5. CONCLUSION

In this paper, we propose a novel BERT-based integrated network to both consider dialogue history and external knowledge in joint SLU tasks. The model is capable of selecting relevant knowledge triples and adopts the attention mechanism to acquire useful knowledge representation. Fused information is then mutually induced between the prediction of dialogue acts and slots. The effectiveness of our proposed model is verified in two multi-turn dialogue datasets and knowledge fusion vectors could be easily applied to downstream dialogue state tracking or management tasks.

6. REFERENCES

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

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

[3] Leonard Abbeduto, "Linguistic communication and speech acts. kent bach, robert m. harnish. cambridge: M.i.t. press, 1979, pp. xvii 327.," *Applied Psycholinguistics*, vol. 4, no. 4, pp. 397–407, 1983.

[4] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proc. Interspeech*, 2007, pp. 1605–1608.

[5] Ting Liu, Xiao Ding, Yue Qian, and Yiheng Chen, "Identification method of user's travel consumption intention in chatting robot," *SCIENTIA SINICA Informationis*, vol. 47, pp. 997, 08 2017.

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

[7] Yu Wang, Yilin Shen, and Hongxia Jin, "A bi-model based rnn semantic frame parsing model for intent detection and slot filling," 2018.

[8] Jie Wu, Ian Harris, and Hongzhi Zhao, "Spoken language understanding for task-oriented dialogue systems with augmented memory networks," in *Proc. of NAACL*, Online, June 2021, pp. 797–806, Association for Computational Linguistics.

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

[10] Rashmi Gangadharaiah and Balakrishnan, "Joint multiple intent detection and slot labeling for goal-oriented dialog," in *Proc. of NAACL*, 2019, pp. 564–569.

[11] Ting-Wei Wu, Ruolin Su, and Biing Juang, "A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding," in *Proc. of EMNLP*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 4884–4896, Association for Computational Linguistics.

[12] Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang, "A Context-Aware Hierarchical BERT Fusion Network for Multi-Turn Dialog Act Detection," in *Proc. Interspeech 2021*, 2021, pp. 1239–1243.

[13] Yufan Wang, Tingting He, Rui Fan, Wenji Zhou, and Xinhui Tu, "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)*, 2019, pp. 960–967.

[14] A. Gupta, P. Zhang, G. Lalwani, and M. Diab, "Casanlu: Context-aware self-attentive natural language understanding for task-oriented chatbots," 2019.

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

[16] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," 2020.

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

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

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in NIPS*, 2017, vol. 30.

[20] Singh A. Arora J. Shukla S. Chauhan A., Malhotra A., "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.*, 2020.

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

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

[23] A Bordes, N Usunier, A Garcia-Duran, J Weston, and O Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in NIPS*. 2013, vol. 26, Curran Associates, Inc.