Knowledge Augmented BERT Mutual Network in Multi-turn Spoken Dialogues

Spoken Language understanding

ICASSP 2022 presentation

Author: **Ting-Wei Wu**, Biing-Hwang Juang

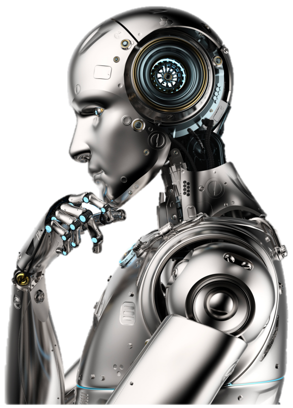
22-27 MAY 2022
ICASSP 2022
SINGAPORE

Task-oriented Dialogs

Restaurant booking scenario

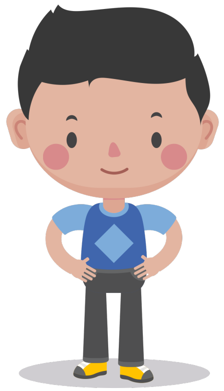
I had 10 restaurants. 2g Japanese Brasserie is great for you.

Offer	name: 2g Japanese Brasserie
Inform count	count: 10



System

User



Yes, 2g Japanese works. I want to reserve there.

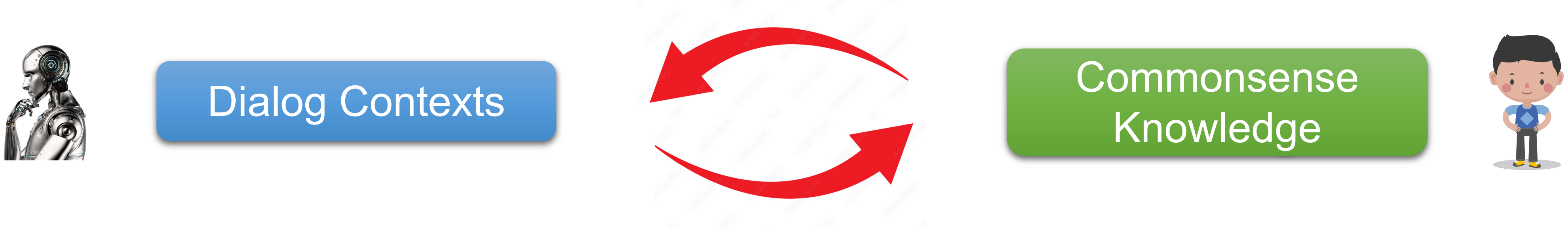
Inform Intent	reserve_restaurant: True
Select	name: 2g Japanese

Predict intents and slots for a given utterance.

Problems

Previous works rely only on single utterances for spoken language understanding.

→ In multi-domain dialogs, it requires back-and-forth interactions to reduce ambiguity.



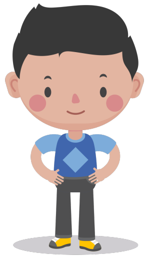
Previous work

1. Model Joint distribution on intents and slots (Liu et al ‘17).
   -> No contexts.
2. Use the previous turn to compare.
   -> Insufficient to model history.
3. Memory network (Chen et al ‘16).
   CASA-NLU (Gupta et al ‘19).
   -> No temporal information.
4. Sequential Dialogue Network (Bapna et al ‘17).
   -> Contexts are condensed.

Previous work

1. Response generation
   (Zhao et al ‘20, Zheng et al ‘21).
   -> SLU is important as well.
2. Knowledge attention (Wang et al ‘19).
   -> Single LSTM to encode all knowledge and contexts.

Example



Is there something that's maybe a good intelligent comedy?

Commonsense Knowledge

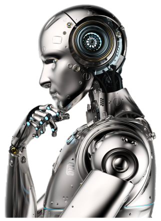
(maybe; related to; uncertainty)

(comedy; is a; drama)

Intent/Slots

Request

genre: comedy



Whiskey Tango Foxtrot is the only Adult comedy I see playing in your area. Would you like to try that

Commonsense Knowledge

(Foxtrot; related to; dance)

(adult; capable of; work)

(area; is a; region)

Intent/Slots

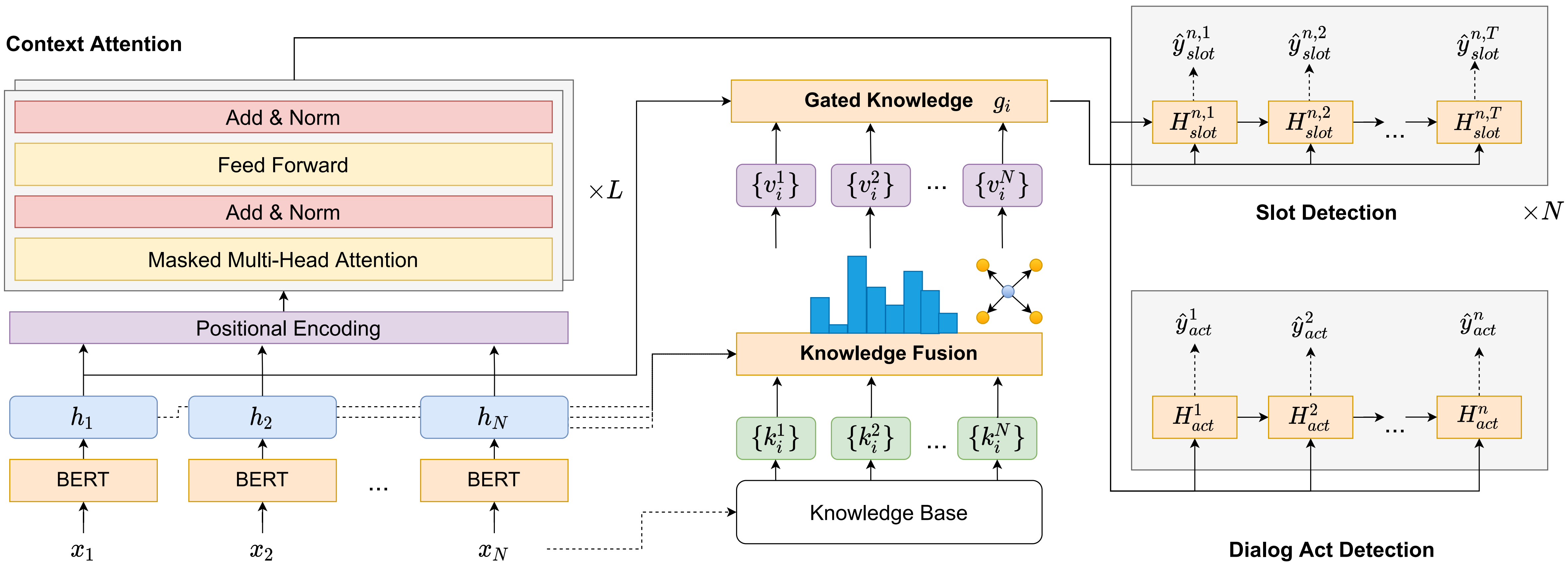
Inform

Movie: Foxtrot

Rating: Adult

Distance: area

Proposed Approach



Proposed Approach

Context Attention

- Masked transformer decoder
  - ➢ Remain chronological order.
  - ➢ Maintain contextual information.
  - ➢ Store previous calculation.

Knowledge Fusion

- Knowledge Attention with contexts.
  - ➢ Extract knowledge triples with word matching.
  - ➢ Context-based filtering.
  - ➢ Knowledge-enriched vectors.

Gated Knowledge

- Non-alphabetic words have no knowledge.
- Gating mechanism to remove noises.

$$h_i^{n\prime} = g_i \cdot h_i^n + (1 - g_i) \cdot v_i^n$$
$$g_i = \sigma(W_i[h_i^n; v_i^n] + b_i)$$

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

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

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

r_{ij}, t_{ij}: entity vectors.
W: learnable matrices.

M: Number of knowledge.
$[;]$: concatenation.
h_i^n: dialog contexts.

Experiments

Multi-turn Dialogue Datasets

1. MDC:
   Microsoft dialogue challenge dataset
2. SGD:
   Schema-Guided Dialogue dataset

	train/val/test (total)	Total Labels	Slots
MDC	45k/15k/15k	11	50
SGD	198k/66k/66k	18	89

Baselines

1. **MID-SF:**
   Multi-intent detection with BiLSTMs.
2. **ECA:**
   LSTM encoder to encoder dialog contexts.
3. **KASLUM:**
   Extract knowledge for joint tasks.
4. **CASA:**
   Encode contexts with DiSAN setnence2token and BERT.
5. **$\mathbf{KABEM_{AF}}$:**
   Replace our knowledge fusion part with attention filter in Wang et al '21.

We randomly select 1000 dialogues for 5 domains.
We use TransE embeddings in ConceptNet as initial knowledge vectors.

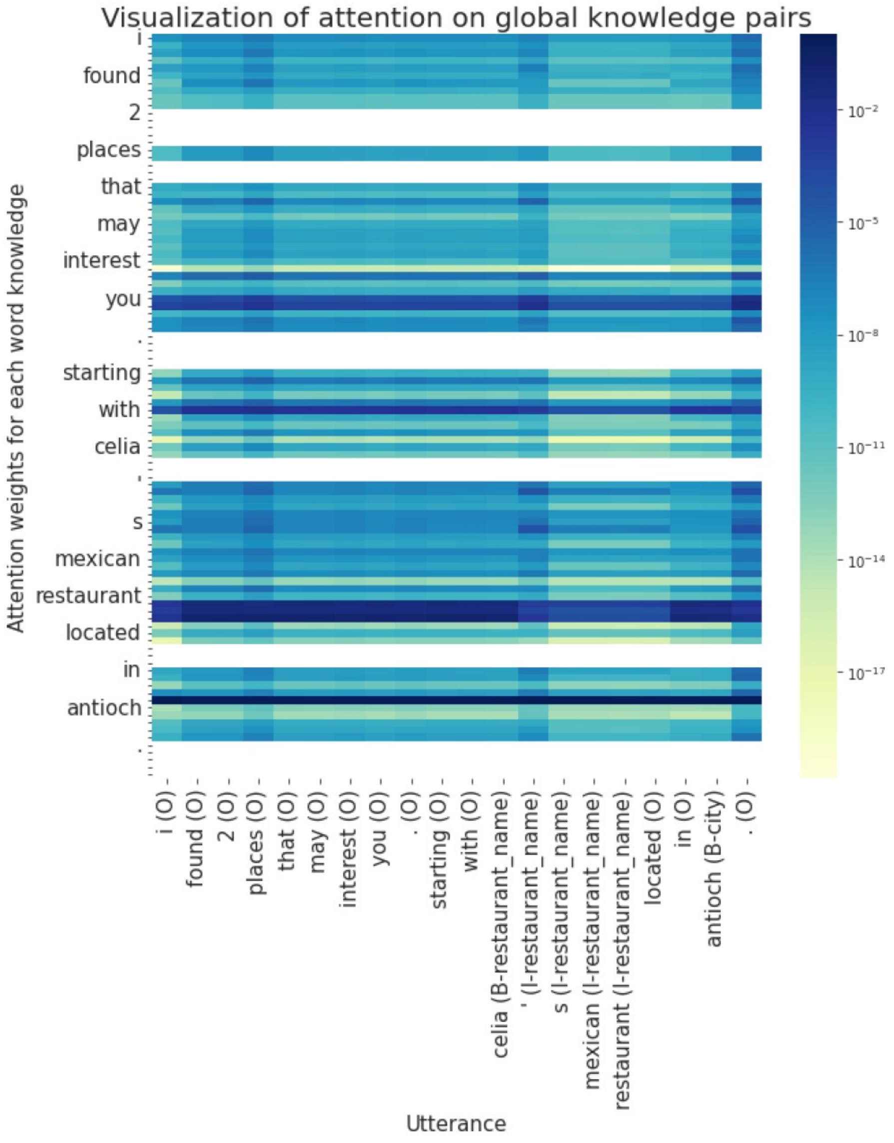
Main Results

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

- More powerful dialog context encoding network and interactions with knowledge.
- Contexts are useful for dialogue act detection.
- Knowledge is useful for slot filling.

Visualization

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)



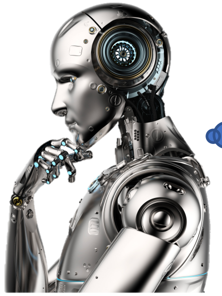
Conclusion

1. Human naturally refers commonsense knowledge to current contexts for understanding.

2. We propose:
   1. Context attention to encoder dialogs.
   2. Knowledge attention to take commonsense knowledge into account.

3. The results achieve the best results on joint multi-intent detection and slot filling tasks compared with several competitive baselines.

Yes, 2g Japanese works. I want to reserve there.



Let me figure out with contexts and knowledge