Multi-Level Memory for Task Oriented Dialogs

Revanth Reddy[1], Danish Contractor[2], Dinesh Raghu[2], Sachindra Joshi[2]

[1]Indian Institute of Technology, Madras [2]IBM Research AI, New Delhi

Motivation

- End-to-end task-oriented dialog systems use neural memory architectures to incorporate external knowledge.
- Current models break down the external KB results into the form of *Subject-Relation-Object* triples.
- This makes it hard for the memory reader to infer relationships across otherwise connected attributes.
- Existing models like *Mem2Seq* use a shared memory for copying entities from dialog context, as well as the KB results, thereby making inference harder.

Role	Turn	Utterance
User	1	Hi, I'm leaving Dallas for Manheim from Aug 26 – Aug 31
Agent	1	I have some options starting at $2800
User	2	How about to Santos?
Agent	2	I have a 3.0 star hotel for $2000
User	3	What is the name of the hotel?
Agent	3	Regal Resort

Origin	Destination	Hotel	Price	Cat.	...
Dallas	Mannheim	Globetrotter	$2800	4.0	...
Toronto	Calgary	Amusement	$1864	4.0	...
Dallas	Santos	Regal Resort	$2000	3.0	...
Dallas	Mannheim	Starlight	$4018	5.0	...
...	...	...	...	...	...

Figure: Sample dialog and its corresponding external KB.

t

Subject	Relation	Object
Globetrotter	Price	$2800
Globetrotter	Category	4.0
Globetrotter	Origin	Dallas
Starlight	Category	5.0
Starlight	Price	$4018

Subject	Relation	Object
Starlight	Origin	Dallas
Regal Resort	Category	3.0
Regal Resort	Price	$2000
Regal Resort	Origin	Dallas
...	...	...

Figure: Triple store in current models.

- We separate the memory used to store tokens from the input context and the results from the knowledge base.
- We propose a novel multi-level memory architecture which encodes the natural hierarchy exhibited in the KB results.
- We store the queries, their results and corresponding attributes in different levels.

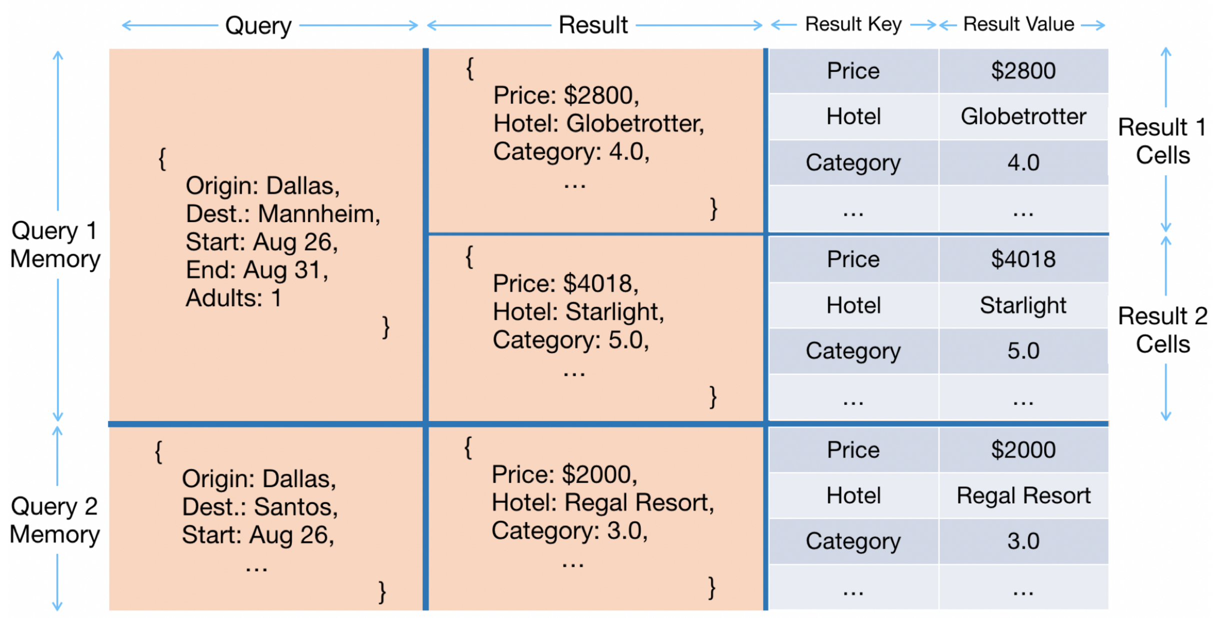


Figure: Multi-Level memory in our model.

Model Architecture

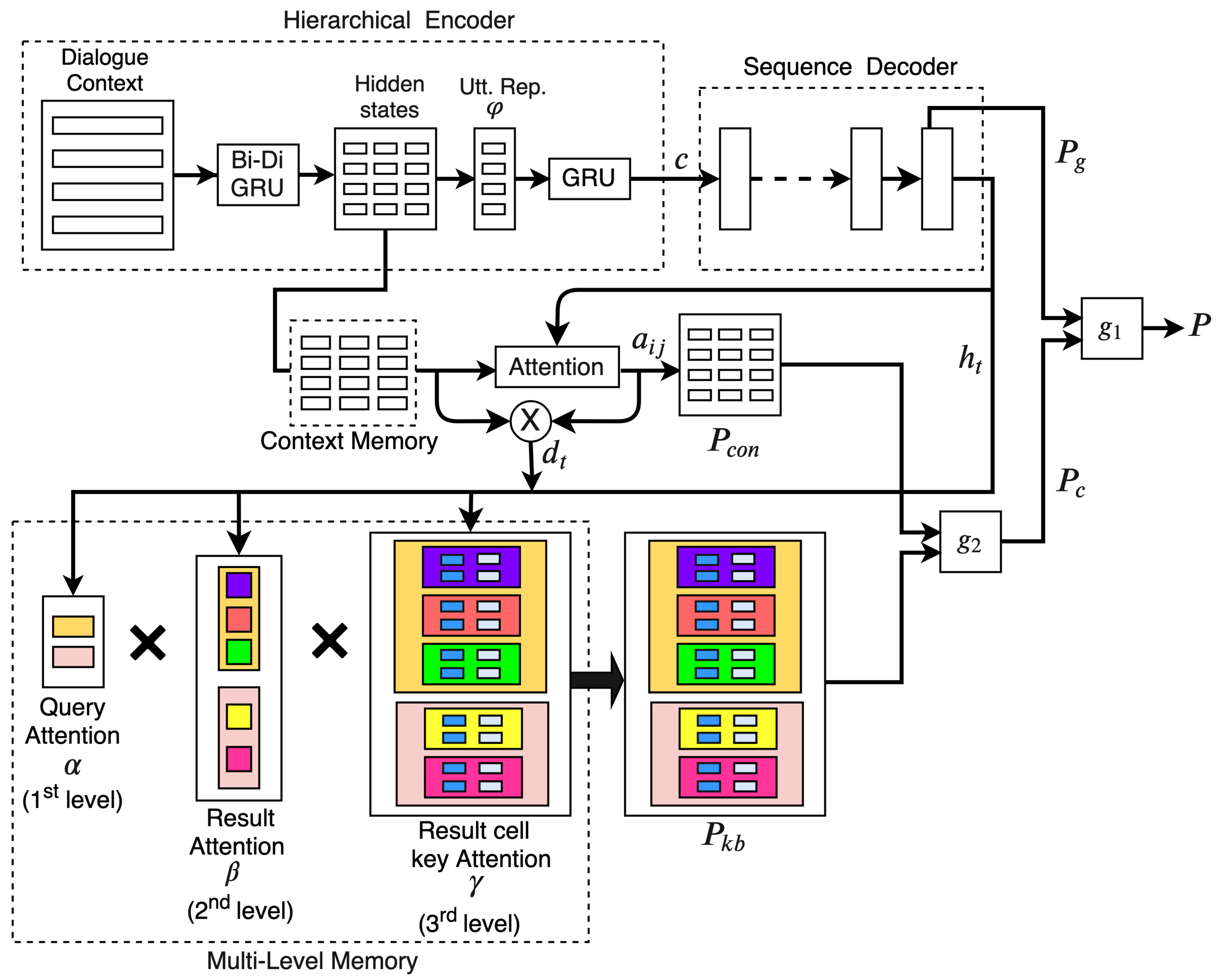


Figure: Model architecture with multi-level memory

Memory Representation

Every query q_i is a set of key-value pairs $\{k_a^{q_i} : v_a^{q_i}, 1 < a < n_{q_i}\}$.
q_i is represented by q_i^v = Bag of words over the word embeddings of values ($v_a^{q_i}$) in q_i.
Each result r_{ij} is also a set of slot-value pairs $\{k_a^{r_{ij}} : v_a^{r_{ij}}, 1 < a < n_{r_{ij}}\}$.
r_{ij} is represented by r_{ij}^v = Bag of words over the word embeddings of values ($v_a^{r_{ij}}$) in r_{ij}.

Equations

Query Attention: The first level attention which is applied over the query representations.

$$\alpha_i = \frac{\exp(w_2^T \tanh(W_4[d_t, h_t, q_i^v]))}{\sum_i \exp(w_2^T \tanh(W_4[d_t, h_t, q_i^v]))}$$

Result Attention: The second level attention which is applied over the result representations.

$$\beta_{ij} = \frac{\exp(w_3^T \tanh(W_5[d_t, h_t, r_{ij}^v]))}{\sum_j \exp(w_3^T \tanh(W_5[d_t, h_t, r_{ij}^v]))}$$

Result cell Attention: The third level attention which is applied over the keys in each result.

$$\gamma_{ijl} = \frac{\exp(w_4^T \tanh(W_6[d_t, h_t, \phi^{emb}(k_l^{r_{ij}})]))}{\sum_l \exp(w_4^T \tanh(W_6[d_t, h_t, \phi^{emb}(k_l^{r_{ij}})]))}$$

KB copy distribution: The product of attention over the three levels gives the final attention score of the values in each result.

$$P_{kb}(y_t = w) = \sum_{ijl: v_l^{r_{ij}} = w} \alpha_i \beta_{ij} \gamma_{ijl}$$

Context copy distribution: Obtained from the attention scores over the input dialog context.

$$P_{con}(y_t = w) = \sum_{ij: w_{ij} = w} a_{ij}$$

Copy Distribution: The copy distribution over the memory is gated sum of the copy distributions over KB and context.

$$P_c(y_t) = g_2 P_{kb}(y_t) + (1 - g_2) P_{con}(y_t)$$

Output Distribution: The final output distribution is gated sum of the generate distribution and the copy distribution over memory.

$$P(y_t) = g_1 P_g(y_t) + (1 - g_1) P_c(y_t)$$

Experiments

Our experiments on three publicly available datasets show a substantial improvement of 15-25% in both entity F1 and BLEU scores as compared to existing state-of-the-art approaches.

	InCar					CamRest		Frames	
Model	BLEU	F1	Calendar F1	Weather F1	Navigate F1	BLEU	F1	BLEU	F1
Attn seq2seq	11.3	28.2	36.9	35.7	10.1	7.7	25.3	3.7	16.2
Ptr-UNK	5.4	20.4	22.1	24.6	14.6	5.1	40.3	5.6	25.8
KVRet	13.2	48.0	62.9	47.0	41.3	13.0	36.5	10.7	31.7
Mem2Seq	11.8	40.9	61.6	39.6	21.7	14.0	52.4	7.5	28.5
Multi-level Memory Model	**17.1**	**55.1**	**68.3**	**53.3**	**44.5**	**15.9**	**61.4**	**12.4**	**39.7**

Table: Comparison of our model with baselines

We investigate the gains made by (i) Using separate memory for context and KB triples (ii) Replacing KB triples with a multi-level memory.

	InCar					CamRest		Frames	
Model	BLEU	F1	Calendar F1	Weather F1	Navigate F1	BLEU	F1	BLEU	F1
Unified Context and KB memory (Mem2Seq)	11.8	40.9	61.6	39.6	21.7	14.0	52.4	7.5	28.5
Separate Context and KB Memory	14.3	44.2	56.9	54.1	24.0	14.3	55.0	12.1	36.5
+Replace KB Triples with Multi-level memory	17.1	55.1	68.3	53.3	44.5	15.9	61.4	12.4	39.7

Table: Ablation study: Effect of separate memory and multi-level memory design.

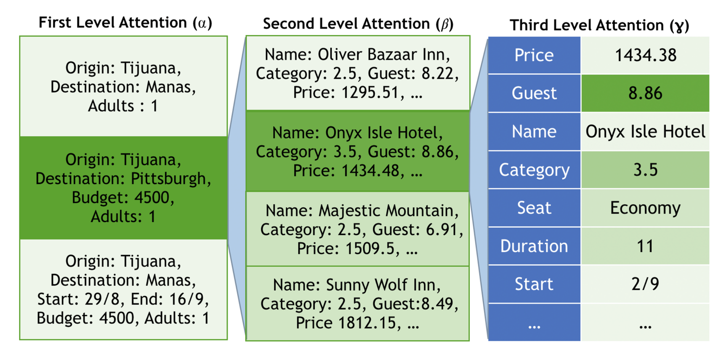
In a human evaluation study, users were asked to score the models in terms of accuracy of information in response and quality of language.

	CamRest			Frames		
	Info.	Lang.	MRR	Info.	Lang.	MRR
KVRet	2.49	4.38	0.57	2.42	3.31	0.64
Mem2Seq	2.48	3.72	0.51	1.78	2.55	0.50
Our Model	**3.62**	**4.48**	**0.76**	**2.45**	**3.93**	**0.69**

We visualize the attention weights to understand how the model is inferencing over the memory. The figures below show the attention heatmap when generating the word '8.86'.

Role	Turn	Utterance
Agent	1	hello ! how can i help you today ?
User	2	*i just need to dream a bit. i have 4500 dollars but no vacation days. i'm wondering what a hypothetical trip to Manas would be like*
Agent	2	i can help with that, where are you departing from
User	3	*i'm in Tijuana, not departing any time soon unfortunately though*
...	...	...
Agent	5	[API_CALL manas tijuana 29/8 16/9 4500 ...]
...	...	...
User	7	*that's fine, what about pittsburgh*
Agent	7	[API_CALL pittsburgh tijuana ... 4500 ...]
...	...	...
Agent	8	that i do have! would you be satisfied with a 3.5 star hotel ?
User	9	*as of now i'm satisfied with a tent and a backpack . what are the hotel details?*
Agent	9	the hotel is situated near a park, comes with free parking, wifi and breakfast ! does this sound good
User	10	*sounds quite nice, what about the guest rating ?*
Gold		the guest rating is impressive : **8.86**
Mem2Seq		this is the only available with the same
KVRet		it's a 2.5 star hotel near a park, museum and airport.it has a guest rating of 8.22
Our Model		the onyx isle hotel has a **8.86** guest rating and offers free parking, breakfast and wifi.

(a) Comparing the responses generated by various models on an example in Frames dataset.



(b) Attention over the multi-level KB memory.

Word	rating	guest	3.5	1	park	...
Score	0.9132	0.0721	0.0108	0.0030	0.0005	...

(c) Decreasing order of attention scores over words in context.

Gate	Value
g_1 (Generate from vocabulary)	0.08
g_2 (Copy from KB memory)	0.99

(d) Probability values of the gates

Conclusion

- Our model separates the context and KB memory and combines the attention on them using a gating mechanism.
- The multi-level KB memory reflects the natural hierarchy present in KB results. This also allows our model to support non-sequential dialogs.
- In future work, we would like to incorporate better modeling of latent dialog frames so as to improve the attention signal on our multi-level memory.
- Model performance can be improved by capturing user intent better in case of non-sequential dialog flow.