Enhanced Knowledge Selection for Grounded Dialogues via Document Semantic Graphs

Sha Li1∗ , Madhi Namazifar2, Di Jin2, Mohit Bansal2, Heng Ji2,
Yang Liu2, Dilek Hakkani-Tur2
1University of Illinois at Urbana-Champaign, 2 Amazon Alexa AI
shal2@illinois.edu
{madhinam, djinamzn, mobansal, jihj,
yangliud, hakkanit}@amazon.com

2022. 6. 12 • ChongQing

—— NAACL 2022

Code: https://github.com/LeqsNaN/KEC

Leibniz-Institut für Sozialwissenschaften

Reported by Sijin Liu

Artificial

Introduction

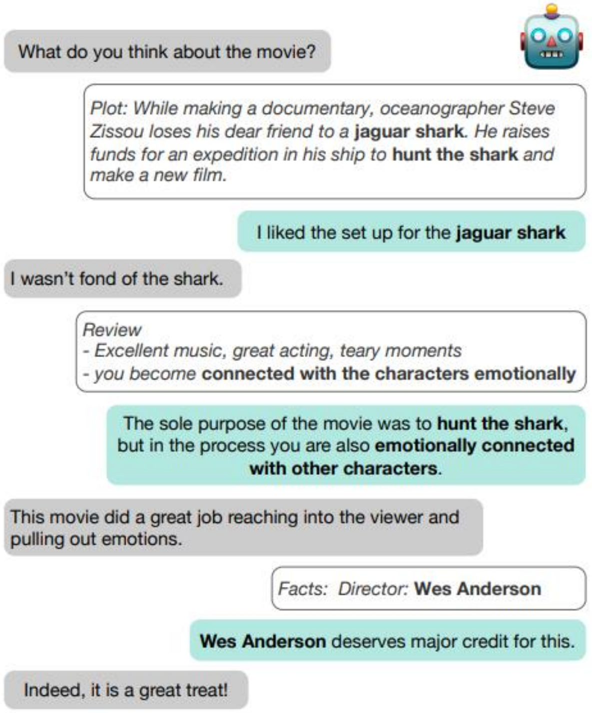


Figure 1: An example of knowledge-grounded dialog. Semantic connections between sentences improve coherence and not imposing knowledge boundaries allows the system to utilize multiple knowledge snippets. The used knowledge is in bold. *The jaguar shark is a character.

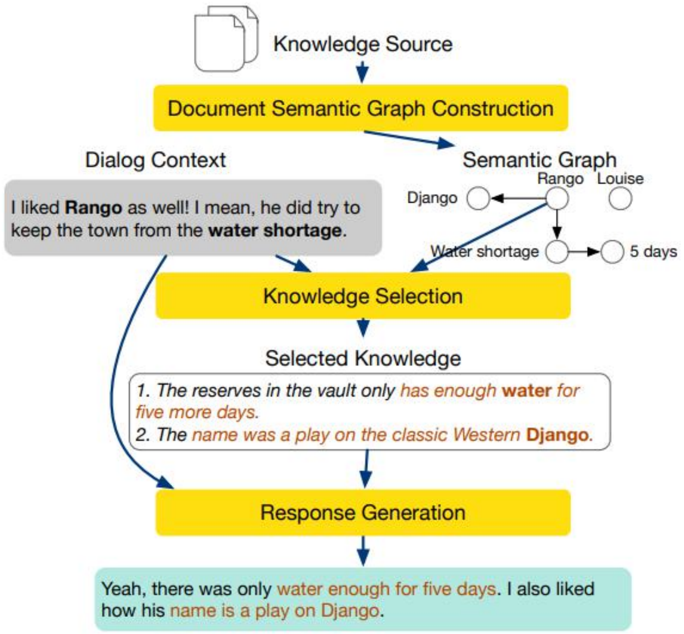


Figure 2: The pipeline for generating responses based on a given knowledge source.

This setting has two inherent draw-backs:
(1) it ignores the semantic connections between sentences and (2) it imposes an artificial constraint over the knowledge boundary.

Hence, to bridge these two worlds of sentence-based knowledge selection and KG-based knowledge selection, we introduce knowledge selection using document semantic graphs.

Method

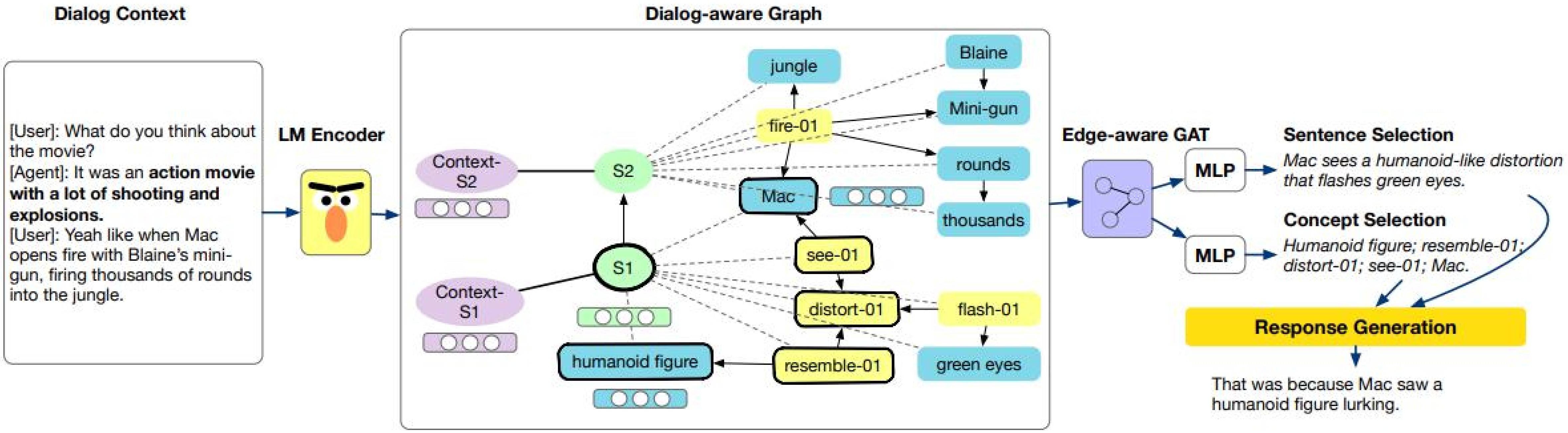


Figure 4: The knowledge selection model. We encode the dialog context using a pretrained language model and represent the dialog context along with each candidate sentence as a context node. We then use an edge-aware graph attention network to encode the dialog-aware graph. Finally, we classify each node on the graph to be relevant or not based on the learned node embedding, effectively performing both sentence selection and concept selection. The selected nodes are outlined in black.

Method

Document Semantic Graph Construction

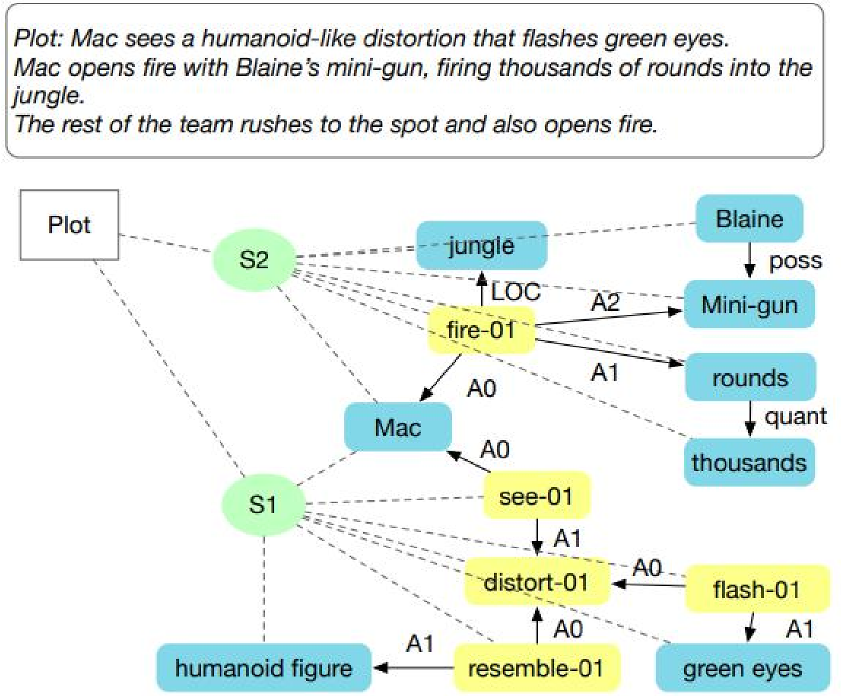


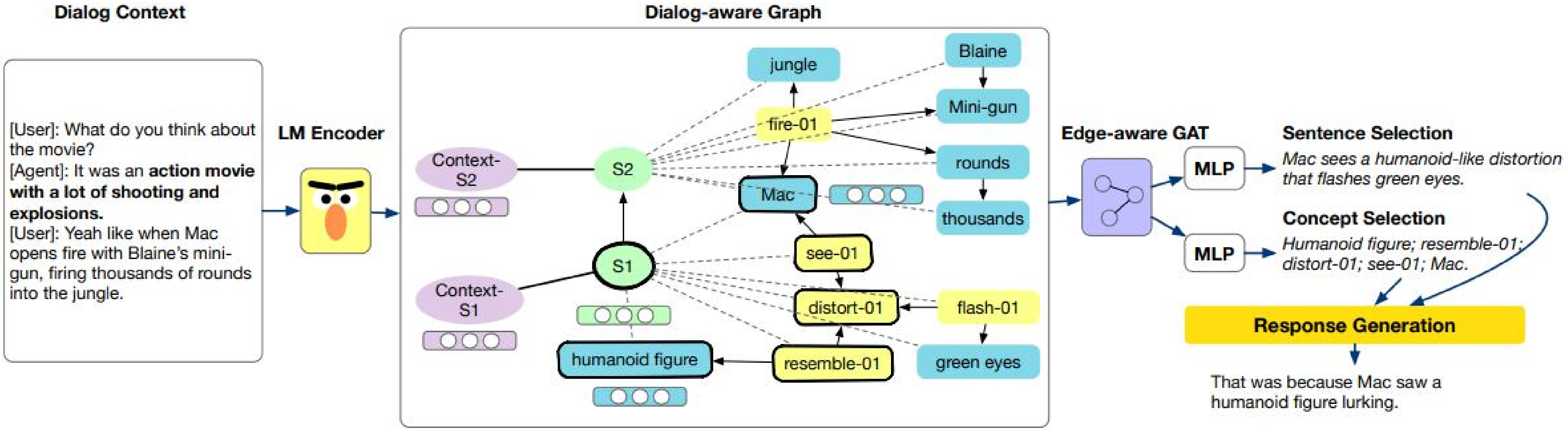
Figure 3: Part of the document semantic graph for the shown plot. The graph includes the source node (white rectangle), the sentence nodes (green circles), and the concept nodes (yellow and blue rectangles). Directional edges with labels (e.g., A0, A1) are from AMR parsing, dotted edges are from the document structure.

We first process the sentences in the background knowledge documents using the Stack Transformer AMR to obtain sentence-level AMR graphs.

Based on the AMR output, we consider all of the concepts that serve as the core roles (agent, recipient, instrument etc.) for a predicate as mention candidates.

Then, we run a document-level entity coreference resolution system to resolve coreference links between such mentions.

Method



Knowledge Selection

$$h_{c_i} = \text{Pooling}\left(f_{\text{LM}}([s_i; x])\right) \quad (1)$$

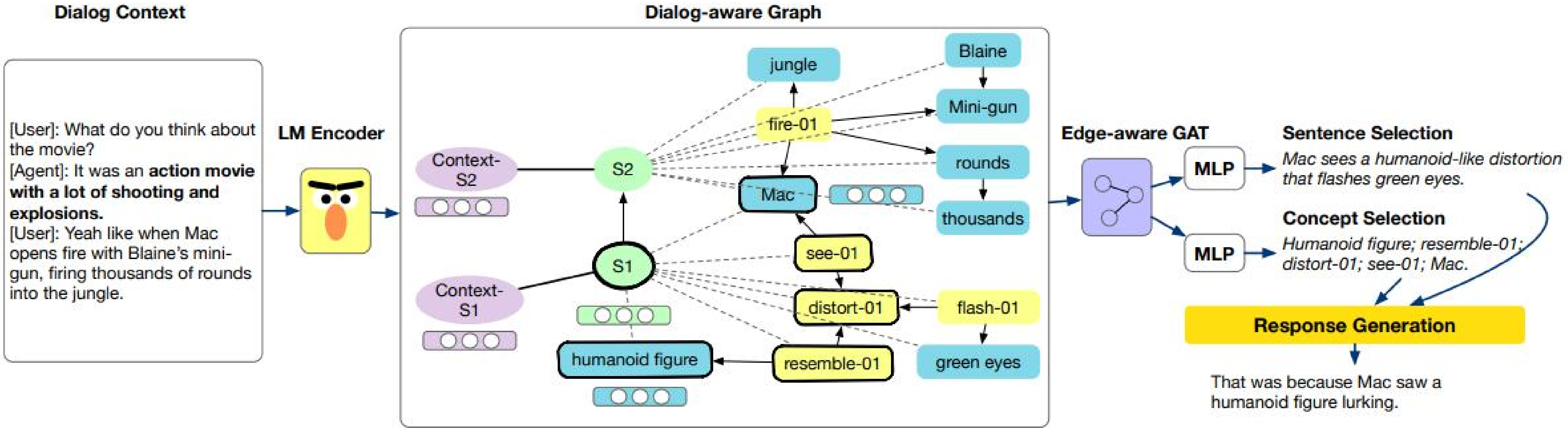
$$m_{s\to t} = W_v([h_s^l; h_{T(v)}]) + W_e h_{T(e)} \quad (2)$$

$$\begin{aligned} q_s &= W_q([h_s^l; h_{T(s)}]) \\ k_t &= W_k([h_t^l; h_{T(t)}; h_{T(e)}]) \\ \alpha_{s\to t} &= \text{Softmax}_{s\in\mathcal{N}_t}\left(\frac{q_s^T k_t}{\sqrt{D}}\right) \end{aligned} \quad (3)$$

$$h_t^{l+1} = \text{GELU}\left(\text{MLP}\left(\sum_{s\in\mathcal{N}(t)} \alpha_{s\to t} m_{s\to t}\right) + h_t^l\right)$$

After L layers, we obtain embeddings for our context nodes h_c^L, sentence nodes h_s^L and concept nodes h_n^L.

Method



Knowledge Selection

$$\text{score}(c) = \text{MLP}([h_c^L; h_c^0]) \qquad (4)$$

$$\text{score}(n) = \sigma\left(\text{MLP}(h_n^L)\right) \qquad (5)$$

$$\mathcal{L}_c = -\log \frac{\exp\left(\text{score}(c^+)\right)}{\exp_{c \in \{c^+\} \cup C^-}\left(\text{score}(c)\right)} \qquad (6)$$

$$\mathcal{L}_n = -\frac{1}{N} \sum_{n \in G} r_n \log \text{score}(n) \qquad (7)$$

$$\mathcal{L} = \mathcal{L}_c + \beta \mathcal{L}_n \qquad (8)$$

Response Generation

$$y = \text{GPT2}([\hat{s}; x]) \qquad (9)$$

Experiments

Dataset		Train	Dev	Test
HollE	Dialogs	7,228	930	913
	# turns	34,486	4,388	4,318
WoW	Dialogs	18,430	981/967	965/968
	# turns	61,263	3401/3186	3246/3360

Table 1: Dataset statistics for WoW and HollE. For WoW, the first column is the seen split and the second column is the unseen split.

Experiments

Model	Single Reference		Multiple Reference		
	MAP	Acc	MAP	MRR	Acc
Ranking	0.493	34.3	0.527	0.526	45.3
Graph Paths	0.497	35.0	0.527	0.579	45.8
Ours	0.513	37.7**	0.514	0.580	46.1

Table 2: Knowledge selection results on the HollE dataset. For single references, MRR is the same as MAP. Acc is reported in percentage%. ** indicates significance compared to the second best model with $p < 0.005$ under the paired t-test.

Model	Test Seen		Test Unseen	
	MAP	Acc	MAP	Acc
Ranking	0.472	30.1	0.436	26.3
Graph Paths	0.469	29.5	0.436	26.4
Ours	0.469	29.4	0.486	30.8**

Table 3: Knowledge selection results on WoW using the topic passage and passages retrieved at the first turn. Acc is reported in percentage%. ** indicates significance compared to the second best model with $p < 0.005$ under the paired t-test.

Experiments

Model	Single Reference			Multiple Reference		
	R1	R2	RL	R1	R2	RL
Transformer MemNet (Dinan et al., 2019)	20.1	10.3	-	24.3	12.8	-
E2E BERT †	25.9	18.3	-	31.1	22.7	-
SKT (Kim et al., 2020)	29.8	23.1	-	36.5	29.7	-
SKT+PIPM+KDBTS (Chen et al., 2020)	30.8	23.9	-	37.7	30.7	-
MIKe (Meng et al., 2021)	37.78	25.31	32.82	44.06	31.92	38.91
GPT2 + Ranking	40.22	31.78	38.73	47.53	39.31	45.89
GPT2 + Graph Paths	40.76	32.32	39.12	47.71	39.33	45.90
GPT2 + Graph Selection	42.49	34.37	41.01	47.89	39.58	46.25
GPT2 + Gold knowledge	75.92	72.82	75.37	75.92	72.82	75.37

Table 4: Response generation results ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL) and knowledge selection accuracy (Acc%) on HollE. † results taken from (Kim et al., 2020). Other results with citations are taken from their respective papers.

Experiments

Model	Preferred	Approp.	Know.	Engaging
Ours	69%	3.54	3.42	3.32
Ranking	56%	3.47	3.39	3.28
MIKe	34.5%	2.88	3.02	2.82

Table 5: Human evaluation results. "Preferred" includes cases where annotators choose multiple systems as the best. 'Approp.' is short for Appropriate, 'Know.' is short for Knowledgeable.

Model	R1	R2	RL
GPT2 + Ranking	19.95	4.70	16.33
GPT2 + Graph Paths	19.83	4.89	16.37
GPT2 + Graph Selection	20.43	5.31	16.97
GPT2 + Gold knowledge	30.53	11.94	25.61

Table 6: End-to-end results (in %) on the unseen split of WoW using first turn retrieved passages as background knowledge.

Model	Acc(%)	MAP	Concept MAP	Concept MRR
Full	37.7	0.513	0.420	0.495
Sent. graph	35.6	0.494	-	-
Coref. graph	37.0	0.510	0.420	0.421
Homog. graph	37.3	0.516	0.409	0.398
Sent. loss	36.0	0.500	0.063	0.151

Table 7: Model ablations for knowledge selection on Holl-E using single reference.

Experiments

Topic	The Matrix	Bad Ass	Rocky V
History	**S**: My favorite character was Morpheus, because he truly believed in what he was doing. **U**: Yeah. It was a story of a reluctant Christ-like person ... The Matrix ... should be the benchmark for all sci-fi films to come.	**S**: What was your favorite scene in this movie? **U**: I liked the one in which Frank fights the skinheads. **S**: Yes. They messed with the wrong senior citizen! **U**: No doubt! Gotta love the viral fame angle they had as well.	**U**: My favorite character was Adrian ... **S**: She sure did protect her family in this great movie. **U**: I have watched the rocky series probably 100 times each ... I would feel offended that Rocky V is rated lower than Rockys III and IV.
Ours	And so much interpretation in these movies. Have the directors ever come out and just explained it? That would make things a lot easier.	Oh yeah, it goes viral & he is proclaimed by the public as "Bad Ass."	Yeah. This movie is a 10, but I feel I need to give it that bit extra due to the very low rating this movie has polled on this site.
Ranking	Yeah. With The Matrix, both elements are perfectly en sync.	Right. It wasn't all that bad when Frank becomes an instant celebrity, and he and Mrs Vega (Tonita Castro) are interviewed at the morning programme LA Mornings by a journalist (Sam Rubin).	My favorite character is Adrian too.
MIKe	I agree , and I loved the scene where he throws Neo into the subway tracks then drops down there.	Danny Trejo's Grand Torino.	A wonderful movie about father and son.

Table 8: Generated responses from our system and baselines on HollE. **S** stands for system turn and **U** for user turn.

Experiments

User | **System** | **Knowledge**

I never realized since recently that the American football rules became popular because of college football.

I didn't know that. I have always wondered where and how the game got started. It can be a dangerous game.

Ground truth: Yes, it can be. College football is considered the second tier of American football in the US.

Therefore, college football is generally considered to be the second tier of American football in the US.

Ours: It was also the first indoor game since 1902.

The modern history of American football can be considered to have begun after the 1932 Playoff game, which was the first American football game to feature hash marks… it was also the first indoor game since 1902.

Ranking: Yeah, I'm not too sure, but I know that the team with the most points wins!

The team with the most points at the end of the game wins.

Figure 5: An example of selected knowledge and generated responses from our model on WoW.