Enhanced Knowledge Selection for Grounded Dialogues via Document Semantic Graphs

Sha Li[1*], **Madhi Namazifar**[2], **Di Jin**[2], **Mohit Bansal**[2], **Heng Ji**[2],
Yang Liu[2], **Dilek Hakkani-Tur**[2]
[1]University of Illinois at Urbana-Champaign, [2] Amazon Alexa AI
shal2@illinois.edu
{madhinam, djinamzn, mobansal, jihj,
yangliud, hakkanit}@amazon.com

Abstract

Providing conversation models with background knowledge has been shown to make open-domain dialogues more informative and engaging. Existing models treat knowledge selection as a sentence ranking or classification problem where each sentence is handled individually, ignoring the internal semantic connection among sentences in background document. In this work, we propose to automatically convert the background knowledge documents into *document semantic graphs* and then perform knowledge selection over such graphs. Our document semantic graphs preserve sentence-level information through the use of sentence nodes and provide concept connections between sentences. We apply multi-task learning for sentence-level knowledge selection and concept-level knowledge selection jointly, and show that it improves sentence-level selection. Our experiments show that our semantic graph based knowledge selection improves over sentence selection baselines for both the knowledge selection task and the end-to-end response generation task on HollE (Moghe et al., 2018) and improves generalization on unseen topics in WoW (Dinan et al., 2019).[1]

1 Introduction

Natural language generation models have seen great success in their ability to hold open-domain dialogues without the need for manual injection of domain knowledge. However, such models often degenerate to uninteresting and repetitive responses (Holtzman et al., 2020), or hallucinate false knowledge (Roller et al., 2021; Shuster et al., 2021). To avoid such phenomena, one solution is to provide the conversation model with relevant knowledge to guide the response generation (Parthasarathi and Pineau, 2018; Ghazvininejad et al., 2018; Dinan et al., 2019). Figure 1 illustrates such knowledge grounded generation.

Relevant knowledge is often presented in the form of documents (Moghe et al., 2018; Zhou et al., 2018b; Ghazvininejad et al., 2018; Dinan et al., 2019; Gopalakrishnan et al., 2019) and the task of identifying the appropriate knowledge snippet for each turn is formulated as a sentence classification or ranking task (Dinan et al., 2019). Although more advanced methods have been proposed by modeling knowledge as a latent variable (Kim et al., 2020; Chen et al., 2020), or tracking topic shift (Meng et al., 2021), they abide by the setting of sentence-level selection. This setting has two inherent drawbacks: (1) it ignores the *semantic connections* between sentences and (2) it imposes an artificial constraint over the *knowledge boundary*.

A document is not simply a bag of sentences, in fact, it is the underlying semantic connections and structures that make the composition of sentences meaningful. Two examples of such connections are coreference links and predicate-argument structures.[2] These connections are vital to the understanding of the document and also beneficial to knowledge selection. In many cases, we can draw information from multiple sentences to create the response, breaking the *knowledge boundary*. For instance, in Figure 2, the connections among the character “Rango”, the plot point “water shortage" and the name “Django" help us generate a response with a smooth topic transition.

A related line of work (Liu et al., 2018; Moon et al., 2019; Xu et al., 2020; Young et al., 2018; Zhou et al., 2018a) that seemingly overcomes the aforementioned issues is knowledge selection from existing knowledge graphs (KGs) such as

[*] Work done as an intern at Amazon Alexa AI.

[1] See https://www.amazon.science/publications/enhanced-knowledge-selection-for-grounded-dialogues-via-document-semantic-graphs for an updated paper with information about code and resources.

[2] Another example would be discourse relations between sentences, which we do not explore here.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2810 - 2823
July 10-15, 2022 ©2022 Association for Computational Linguistics

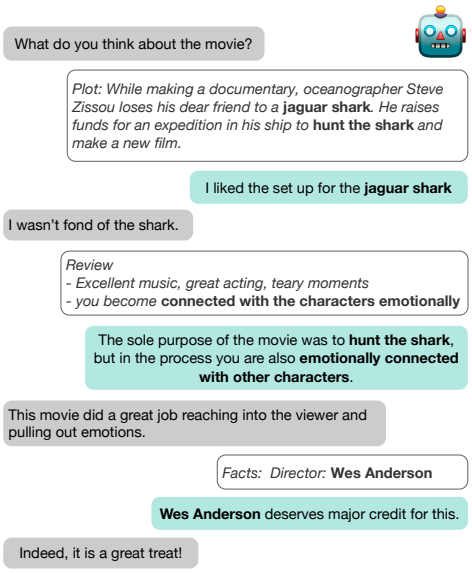

Figure 1: An example of knowledge-grounded dialog. Semantic connections between sentences improve coherence and not imposing knowledge boundaries allows the system to utilize multiple knowledge snippets. The used knowledge is in bold. *The jaguar shark is a character.

Wikidata (Vrandecić and Krötzsch, 2014), DBpedia (Lehmann et al., 2015), and ConceptNet (Speer et al., 2017). If the character "Rango" were in the KG, it would have been represented as an entity node and be connected to respective events. On the KG, we are also free to select as many concepts as needed, without being restricted to a single sentence as the source. However, KGs are known to have limited coverage of real world entities, let alone emerging entities in works of fiction such as books and movies (Razniewski et al., 2016).

Hence, to bridge these two worlds of sentence-based knowledge selection and KG-based knowledge selection, we introduce knowledge selection using *document semantic graphs*. These graphs are automatically constructed from documents, aiming to preserve the document content while enhancing the document representation with semantic connections. To create such a document semantic graph, we first obtain the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) for each sentence. AMR detects entities, captures predicate-argument structures, and provides a layer of abstraction from words to concepts.[3] Compared to existing knowledge base construction methods, AMR covers a wide range of relations and fine-grained semantic roles, and can fully reflect the semantics of the source text. Since AMR graphs only represent single sentences, we utilize coreference resolution tools to detect coreferential entity nodes and merge them to build graphs for documents. On top of this content representation, we also add sentence nodes and passage nodes to reflect the structure of the document. This allows for traversal across the graph by narrative order or concept association.

Given the document semantic graph, knowledge selection can be seen as identifying relevant nodes on the graph, sentence nodes or concept nodes. As knowledge selection in dialog models is conditioned on the dialog context, for each dialog turn, we create a *dialog-aware graph* derived from the document graph. It contains context nodes representing contextualized versions of the sentence and concept nodes. We design an edge-aware graph neural network model to propagate information along the dialog-aware graph and finally score the context nodes (or concept nodes) on their relevance to the dialog turn (as shown in Figure 4).

We validate our model on two widely used datasets HollE (Moghe et al., 2018) and Wizard of Wikipedia (Dinan et al., 2019) by constructing a semantic graph from relevant background documents[4] for each dialog. The use of document graphs improves both knowledge selection and response generation quality on HollE and boosts generalization to unseen topics for WoW. From our ablation tests we find that in terms of the graph structure, the key component is the use of coreference edges that stitches sentences together.

Our contributions include: (1) We propose to perform knowledge selection from document semantic graphs that are automatically constructed from source documents and can reflect the implicit semantic connections between sentences without being limited to a pre-defined set of entities and relations as KGs do. Our approach bridges the gap between sentence-based knowledge selection and KG-based knowledge selection. (2) We show that joint selection over sentences and concepts can model more complex relations between sentences and boost sentence selection performance. (3) We build a pipeline for converting documents (document collections) into semantic graphs through the use of AMR parsing and coreference. We hope that

[3] In AMR, every node is a concept. This includes events, objects, attributes, etc.

[4] On WoW we use the passages retrieved from the first turn for graph construction. Our method will need to be extended to online graph construction to support per-turn retrieval of documents.

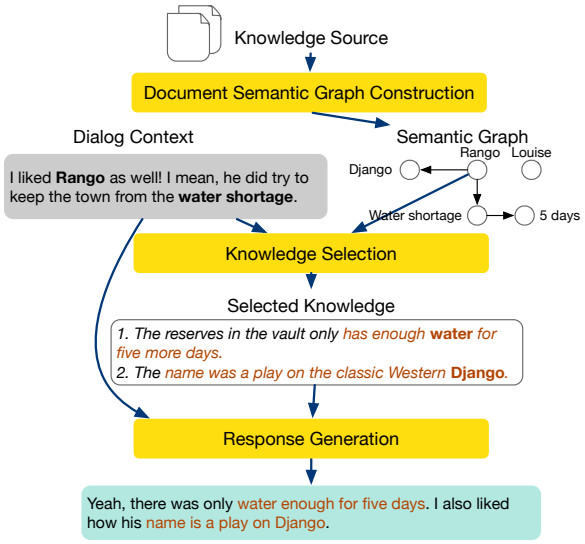

Figure 2: The pipeline for generating responses based on a given knowledge source.

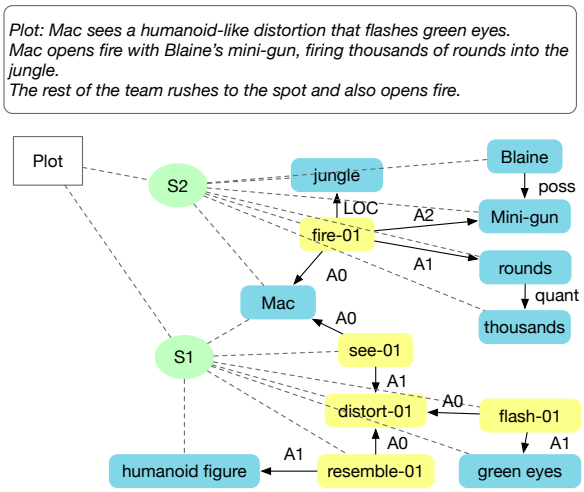

Figure 3: Part of the document semantic graph for the shown plot. The graph includes the source node (white rectangle), the sentence nodes (green circles), and the concept nodes (yellow and blue rectangles). Directional edges with labels (e.g., A0, A1) are from AMR parsing, dotted edges are from the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

our tool can facilitate future work on graph-based representations of documents.

2 Method

We show an overview of our knowledge-grounded dialog system in Figure 2. The system consists of three modules, namely semantic graph construction, knowledge selection and response generation.

2.1 Document Semantic Graph Construction

We first process the sentences in the background knowledge documents using the Stack Transformer AMR parser (Fernandez Astudillo et al., 2020) 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 (Wen et al., 2021) to resolve coreference links between such mentions. When joining sentence-level AMR graphs to form the document graph, entity mentions that are predicted to be coreferential are merged into one node, and we keep the longest mention as the node's canonical name. We show an example of our constructed document semantic graph in Figure 3.

On top of this content representation, we also add additional nodes to represent documents (or passages) and sentences. A document (or passage) node is linked to sentence nodes that are from this document (or passage). Each sentence node is directly connected to all the concept nodes that originate from that sentence. In addition, we add edges between neighboring sentences following the narrative order in the document.

Since each node is grounded in text, in order to create embeddings for the document semantic graph, we initialize the embedding of each node with their contextual embeddings from a frozen pretrained language model RoBERTa (Liu et al., 2019a). For sentence nodes, we use the embedding of the `[CLS]` token. For concept nodes, we average the embeddings of the tokens in the span. Note that the document semantic graphs can be created offline and indexed by topics to be used at knowledge selection inference time.

2.2 Knowledge Selection

The task of knowledge selection is to identify relevant knowledge snippets that can be used to produce an appropriate and informative response for each turn. Since our document semantic graph is based on the background knowledge source alone, we first create a *dialog-aware graph* that is conditioned on the given dialog turn. We then encode the dialog-aware graph by an edge-aware graph attention network and predict relevance scores for sentences and concepts as shown in Figure 4.

Dialog-Aware Graph. The dialog-aware graph is a copy of the document semantic graph with additional context nodes (c), each representing a dialog-contextualized knowledge sentence. For each candidate knowledge sentence s_i, to obtain the em-

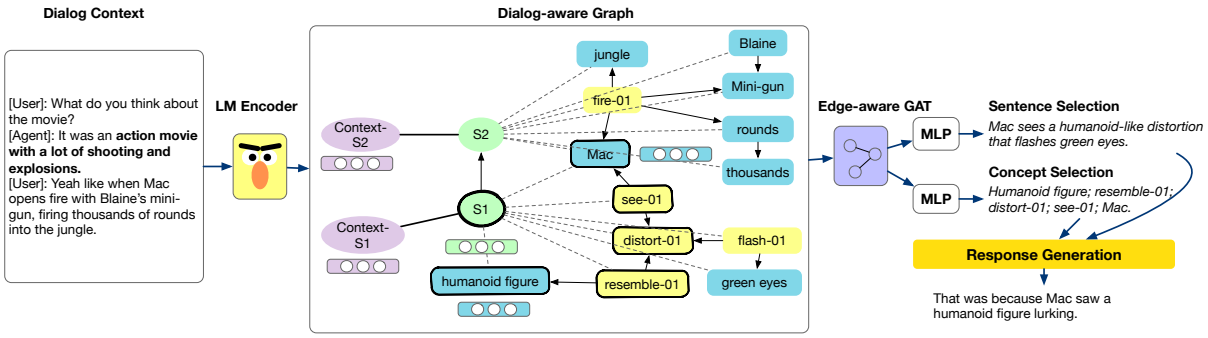

Figure 4: The knowledge selection model. We encode the dialog context using a pretrained language model and represent the dialog context along with each candidate sentence as a context node. We then use an edge-aware graph attention network to encode the dialog-aware graph. Finally, we classify each 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.

bedding h_c of the context node c_i, we encode the dialog context x and the candidate knowledge sentence s_i through a pretrained language model f_{LM}. We define the dialog context as the most recent two turns in the dialog history.

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

For the pooling operation, we simply take the first token (namely the `[CLS]` token) as the representation for the sequence. Since we want to enable message passing between the context node and the rest of the graph, we add an edge between the context node c_i and the sentence node s_i.

Edge-Aware Graph Attention Network. At this point, although our dialog-aware graph captures both the dialog context and the knowledge source, there is no interaction between the two. To this end, we apply an edge-aware graph attention network (EGAT) model to allow information to be propagated along the graph. Note that our dialog-aware graph is a heterogeneous network with multiple node types and edge types. To capture the semantics of the node and edge types, we use an extension of the graph attention network (Velickovic et al., 2018) that includes edge type embeddings $h_{T(e)}$ and node type embeddings $h_{T(v)}$ (Yasunaga et al., 2021). These embeddings are learnt along with the model parameters and are used to compute the vector "message" that is passed along edges.

In general, a graph neural network consists of L layers with shared parameters. We denote the initial embeddings for each node as h^0. Each layer l involves a round of nodes sending out "messages" to their neighbors and then aggregating the received "messages" to update their own embeddings from h^l to h^{l+1}. Consider a pair of nodes s and t with embeddings h_s and h_t respectively, the message $m_{s\to t}$ that is passed from s to t through edge e is computed as the sum of the edge-aware message and the node-aware message, where W_v and W_e are projection matrices.

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

Then we compute the attention weight $\alpha_{s\to t}$ from node s to node t as:

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

Here W_q and W_k are learnt projection matrices and D is the embedding dimension of h_s. $\mathcal{N}_t$ is the neighbor node set of node t. Finally, the messages from the surrounding neighbors are aggregated to compute the updated node embedding h_t^{l+1}.

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

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

Knowledge Selection Training. For each context node c that represents a pair of the knowledge sentence and dialog context, we compute their relevance score as

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

For each concept node n, we compute its relevance score as

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

where σ is the sigmoid function.

Each context node c needs to be encoded with the language model f_{LM}, but we are unable to fit all context nodes into memory.[5] Hence, during training, we randomly sample k negatives for each positive knowledge sentence and compute cross-entropy loss over the samples.

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

For concept nodes, we treat knowledge selection as a binary classification problem and compute the binary cross entropy loss.

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

Here $r_n \in \{0, 1\}$ is the relevance label for the vertex n, and N is the total number of concept nodes. When the dataset does not directly provide concept-level labels for training, we derive them from the ground truth knowledge snippet by assigning any concept that is mentioned in the snippet with a relevant label $r_n = 1$. The overall loss is the weighted sum of the above sentence-level and the concept-level loss:

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

During inference, we compute $\text{score}(c)$ for all knowledge sentence candidates and take the highest scored sentence for knowledge grounded response generation.

2.3 Response Generation

We fine-tune a left-to-right language model GPT2 (Radford et al., 2019) to perform response generation given the dialog context x and the chosen knowledge snippet $\hat{s}$.

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

During training we use teacher-forcing and use the ground truth knowledge snippet. This response generation model is independent from the knowledge selection model and trained with negative log-likelihood loss.

3 Experiments

3.1 Datasets

We evaluate our model on two publicly available datasets: Wizard of Wikipedia (Dinan et al., 2019) and Holl-E (Moghe et al., 2018). Both datasets are in English.

[5] On average, we have 60 knowledge sentences per turn.

Dataset		Train	Dev	Test
HollE	Dialogs	7,228	930	913
	# turns	34,486	4,388	4,318
WoW	Dialogs	17,629	941/936	924/952
	# turns	22,715	3257/3085	3104/3298

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.

Wizard of Wikipedia (WoW) is an open-domain dialog dataset, spanning multiple topics including famous people, works of art, hobbies, etc. The test set in WoW consists of two splits that are named "Test Seen" and "Test Unseen" based on the overlap of topics with the training set. In order to build our document graph, we use the selected topic passage and the passages retrieved in the first turn as background knowledge.

Holl-E is a movie domain dialog dataset. Each dialog discusses one movie, and the background knowledge includes the plot, reviews, comments and a fact table. Holl-E additionally provides multiple references for the test set so we report performance for both single and multiple references.

3.2 Implementation Details

Knowledge Selection. We only use the turns that utilize knowledge for training and prediction. To map the ground truth knowledge to a set of concept nodes, we choose all nodes with mention offsets contained within the span. We acquire the sentence-level labels following (Kim et al., 2020).

We use Roberta-base (Liu et al., 2019a) as the language model f_{LM}. We set $k = 5$ for negative sampling. The EGAT model is trained with 200 hidden dimensions and 2 layers. Edge features and node features are represented with 20 dimensional vectors. We train our model with a batch size of 16 and learning rate $3e^{-5}$ for 3 epochs.

Response Generation. Our response generation model is based on GPT2 (Radford et al., 2019) and is further fine-tuned with a batch size of 16 and learning rate of $3e^{-}5$. We truncate the dialog context to 128 tokens. During inference, we adopt top-k and top-p sampling with $k = 20$ and $p = 0.95$. The maximum generation length is limited to 286 tokens, including the input tokens.

3.3 Baselines

For knowledge selection, we also implemented the following two baseline methods:

- Roberta Ranking. We use a cross-encoder based on Roberta to represent the dialog context and the knowledge candidate, and a classification layer on top of it.
- Graph Paths. This model is built on top of the previous model. The graph paths are from the document semantic graph and obtained by breadth-first traversal starting at the candidate context node. In order to utilize the graph paths, we linearize it into tuples of (subject, predicate, object) or (modifier, subject) according to the AMR edge label and concatenate it with the candidate sentence.

For the end-to-end pipeline, we use the GPT2 response generation with our knowledge selection module and the two methods above. In addition, we compare against the following previous methods:

- Transformer MemNet (Dinan et al., 2019) is the combination of a Transformer memory network for knowledge selection and another Transformer decoder for generation.
- E2E BERT is a variant of the previous model using BERT (Devlin et al., 2019).
- Sequential Knowledge Transformer (SKT) (Kim et al., 2020) models knowledge as a latent variable and considers the posterior distribution of knowledge given the response.
- SKT+PIPM+KDBTS (Chen et al., 2020) is an improvement upon SKT with an additional Posterior Information Prediction Module (PIPM) and trained with knowledge distillation.
- Mixed Initiative Knowledge Selection (MIKe) (Meng et al., 2021) uses two knowledge selection modules to capture user-driven turns and system-driven turns respectively.

3.4 Evaluation Metrics

Knowledge Selection. To compare with previous methods, we use Accuracy, or Precision@1 as the main metric for evaluating knowledge selection. Additionally, we compute sentence ranking metrics, namely the mean average precision (MAP) and mean reciprocal rank (MRR)[6] for more fine-grained analysis of knowledge selection quality.

[6] https://github.com/usnistgov/trec_eval

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.

Response Generation. For automatic evaluation of responses, we use ROUGE-1, ROUGE-2 and ROUGE-L metrics (Lin, 2004).[7] As our response generation model is trained with gold-standard knowledge, we only report perplexity scores when using gold-standard knowledge, as a measure for the quality of the response generator alone.

For our human evaluation, we randomly sample 200 turns from the output of MIKe (Meng et al., 2021), our ranking model and our graph-based model. Annotators are asked to select which system's response is the best among the three (allowing for ties), and which system's knowledge is the most relevant. In addition, annotators score each response based on whether it is appropriate, knowledgeable and engaging on a scale of 1-4. Our annotators agreed with each other 54.2% on a single system and 91.7% when accounting for ties. The Krippendorff's alpha score for the normalized appropriate/knowledgeable/engaging scores is 0.537/0.634/0.470.

3.5 Main Results

We show our knowledge selection results in Table 2 and 3, and end-to-end results in Table 4 and 6.

From Table 2 we can see that our document semantic graph is helpful for the knowledge se-

[7] We use the `torchmetrics` package, which follows `rouge-score` package Python ROUGE implementation.

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.

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.

lection task and our edge-aware graph attention network is more effective in utilizing the graph structure compared to simply enumerating graph paths. In particular, when the graph is used, there is a large improvement in MRR when multiple gold-standard references are provided, showing that in cases where the top 1 result does not match the reference, we are able to rank the gold-standard knowledge at a high position.

For the end-to-end evaluation in Table 4, our model stands favorably among previous published results, with improvements in both knowledge selection accuracy and response quality.

We report human evaluation results in Table 5. Our system scores the best in all aspects and is voted by annotators as the most preferred response in the majority of the cases.

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.

On the WoW dataset (Table 3 and Table 6), the basic ranking model performs slightly better on the seen split and our graph-based knowledge selection method shows benefits for generalizing to unseen topics.

3.6 Analysis

Model Ablations. We investigate whether our design of the document semantic graph is effective by exploring different variants of the document graph, including: (1) **sentence graph** with only sentence nodes and source nodes, (2) **coreference graph** that removes all AMR role edges, and (3) **homogeneous graph** that treats all edges and nodes as the same type. The results are presented in Table 7. In particular, the **sentence graph** does not make use of AMR parsing nor coreference resolution, so it only reflects the document structure. This makes it the least effective in knowledge selection and unable to perform concept selection at all. The **coreference graph** does not perform as well as the full graph, but largely closes the gap. This suggests that entity recognition and coreference resolution are essential to the effectiveness of the document graph. When using the **homogeneous graph**, our

edge-aware graph attention network falls back to a regular graph attention network. We can see that without edge and node semantics, both sentence and concept selection are negatively impacted.

An important characteristic of our model is that it is trained to perform joint sentence selection and concept selection through a multi-task objective. We compare our full model with a variant, which is only trained with sentence-level supervision signal. Our results show that adding the concept selection loss not only enables concept-level knowledge selection, but also improves sentence-level knowledge selection.

Case Studies. We present some examples of the generated responses on HollE in Table 8. In the first example, the system started out with a comment on the character "Morpheus", the user agreed, and then shifted the topic towards a general comment on the movie. Both our model and the ranking model are able to follow the user's topic and make comments on the movie while the MIKe model continues the previously initiated topic. In the second example, we see that our model and the ranking model both capture the "viral fame" keyword in the user's response, but our model is able to produce a more appropriate response instead of directly copying the plot. In the last example, the ranking model repeats what the user said while our model and MIKe pick knowledge that is relevant to the rating of the movie. In this case, our model produces a more engaging response.

Figure 5 visualizes an example from the WoW dataset about the topic "Football". In this conversation, although the knowledge selected by our model is not the same as the ground truth, it is relevant to the user's question of "where and how the game (of football) got started". The ground truth, on the other hand, follows up on the wizard's own initiated topic of "college football".

Discussions on Limitations. (1) *Concept selection.* Current datasets were annotated with sentence selection in mind and only provided sentence level references. This makes it hard to directly demonstrate the utility of concept selection. (2) *Better utilization of history.* We have used the dialog history in a primitive way by concatenating the latest turns with the candidate knowledge. This ignores earlier turns, and leads to cases of repetition of history, or contradiction of persona. (3) *Limitation of preprocessing tools.* Our document semantic graphs rely on AMR parsing, which might not be available for other languages, or not be of high quality.

4 Related Work

Knowledge Selection for Dialog. Knowledge selection can be tightly coupled with the response generator (Ghazvininejad et al., 2018) or performed separately prior to response generation. Some approaches adapted question answering models (Moghe et al., 2018; Qin et al., 2019; Wu et al., 2021) or summarization models (Meng et al., 2020a) for knowledge selection. With a pool of knowledge candidates, knowledge selection has been commonly set up as a sentence classification or ranking problem (Dinan et al., 2019; Lian et al., 2019; Kim et al., 2020; Chen et al., 2020; Meng et al., 2020b; Zhao et al., 2020). Some work has modeled the underlying knowledge as a latent variable (Lian et al., 2019; Kim et al., 2020; Chen et al., 2020). Others have explored modeling the knowledge transition over dialog turns to improve selection accuracy (Kim et al., 2020; Meng et al., 2020b; Zheng et al., 2020; Zhan et al., 2021). In comparison, we model knowledge selection as a node selection task on the document semantic graph.

Graph-based Knowledge Sources. Knowledge graphs are popular choices for integrating knowledge into dialog systems (Liu et al., 2018; Moon et al., 2019; Xu et al., 2020; Jung et al., 2020; Zhou et al., 2020). However, their applicability is limited by the coverage of both entities and relations. For example in (Moon et al., 2019), for books and movies, the knowledge base only contains metadata such as title and genre, making it impossible to conduct conversation about the actual content. The closest work to ours is AKGCM (Liu et al., 2019b), which starts from an existing general knowledge graph and then augments the knowledge graph with unstructured text by performing entity linking on the sentences. In comparison, our document semantic graph is created from knowledge documents and during knowledge selection we select both sentences and concept nodes.

Application of Document Graphs. Document-level AMR graphs have been used for summarization (Liu et al., 2015; Dohare et al., 2018; Hardy and Vlachos, 2018; Lee et al., 2021) and document generation (Fung et al., 2021). Graphs constructed using OpenIE (Banko et al., 2008) have been ap-

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.

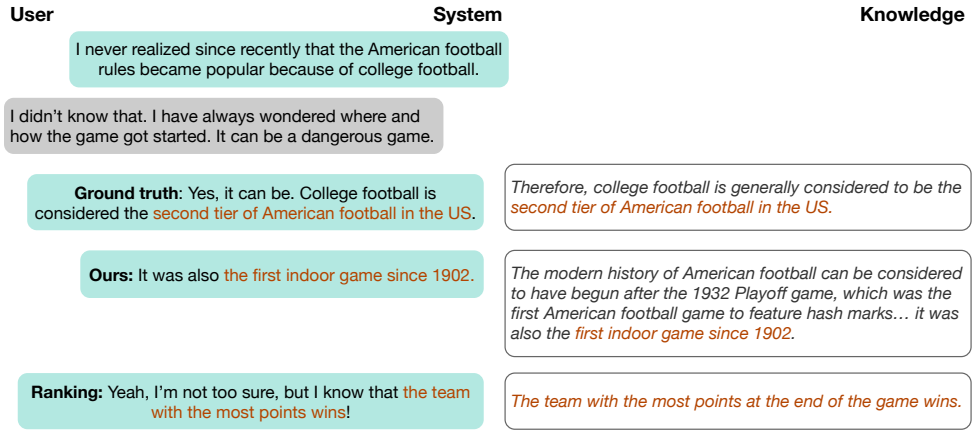

Figure 5: An example of selected knowledge and generated responses from our model on WoW.

plied to long-form question answering and multi-document summarization (Fan et al., 2019).

5 Conclusion and Future Work

In this paper, we introduce *document semantic graphs* for knowledge selection. Compared to existing document-based knowledge selection methods that typically treat sentences independently, our automatically-constructed document semantic graphs explicitly represent the semantic connections between sentences while preserving sentence-level information. Our experiments demonstrate that our semantic graph-based approach shows advantages over various sentence selection baselines in both the knowledge selection task and the end-to-end response generation task.

6 Ethical Considerations

The paper focuses on improving the knowledge selection component for dialog systems.

Intended use. The intended use of this grounded dialog system is to perform chit-chat with the user on topics such as books and movies. We also hope that our released system can help research in knowledge selection.

Bias. Our model is developed with the use of large pretrained language models such as RoBERTa (Liu et al., 2019a) and GPT2 (Radford et al., 2019), both of which are trained on large scale web data that is known to contain biased or discriminatory content. The datasets that we train on also include subjective knowledge (comments on movies) that may express the bias of the writers.

Misuse potential. Although our system is knowledge-grounded, the output from our system should not be treated as factual knowledge. It should also not be considered as advice for any critical decision-making.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2008. Open information extraction from the web. In *CACM*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and J. Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*.

Shibhansh Dohare, Vivek Gupta, and Harish Karnick. 2018. Unsupervised semantic abstractive summarization. In *ACL*.

Angela Fan, Claire Gardent, C. Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *EMNLP/IJCNLP*.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, Online. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, W. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Q. Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and D. Hakkani-Tur. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*.

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using abstract meaning representation. In *EMNLP*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ArXiv*, ICLR.

Jaehun Jung, Bokyung Son, and Sungwon Lyu. 2020. AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497, Online. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and G. Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *ICLR*, abs/2002.07510.

Fei-Tzin Lee, Christopher Kedzie, Nakul Verma, and Kathleen McKeown. 2021. An analysis of document graph construction methods for amr summarization. *ArXiv*, abs/2111.13993.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *IJCAI*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman M. Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *NAACL*.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019b. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and M. de Rijke. 2020a. Refnet: A reference-aware network for background based conversation. In *AAAI*.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and M. de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. 2020b. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Simon Razniewski, Fabian M. Suchanek, and Werner Nutt. 2016. But what do we actually know? In *AKBC@NAACL-HLT*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *EMNLP*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI*.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2018. Graph attention networks. *ICLR*.

Denny Vrandecić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57:78–85.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

J. Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. In *AAAI*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Tom Young, E. Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialog systems with commonsense knowledge. In *AAAI*.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *NAACL*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and A. Black. 2018b. A dataset for document grounded conversations. *EMNLP*.

A Experiment Details

Our experiments were run on a single V100 or RTX2080 GPU. We use gradient accumulation to reach an effective batch size of 16.

On average, the semantic graph construction takes 40s per document for the AMR parsing and 30s per document for coreference resolution. All documents were constructed before running knowledge selection experiments. Our knowledge selection model requires 10G of GPU memory and 6 hours to finish training. Our response generation model takes 1.5 hours to finish training.

We tuned our learning rate in the range of $[3e-6, 1e-5, 3e-5, 5e-5]$ and our batch size in the range of $[4, 8, 16]$. For the EGAT model, we experimented with hidden dimensions of $[50, 100, 200]$ and layers from $[2, 3, 4]$.

B Extra Case Studies

In Figure 6 we present an instance where the question from the user is quite open-ended and while our model's selection does not match the ground truth, it is still relevant to the dialog and can serve as the basis for an appropriate response.

In Figure 7 we show an example where knowledge selection performance does not directly translate to better dialog due to response generation errors. The selected knowledge from our model follows up on the "set routines" mentioned by the user but the response's stance is wrong. The baseline model selects a general statement about cheerleading as the relevant knowledge but the response is logically incorrect as the difficulty of cheerleading is not due to its geographical origin, but due to the moves.

C Human Eval Details

We show an example of the information provided to annotators in Figure 8. Annotators have access to the dialog history and the ground truth responses. System outputs are anonymized.

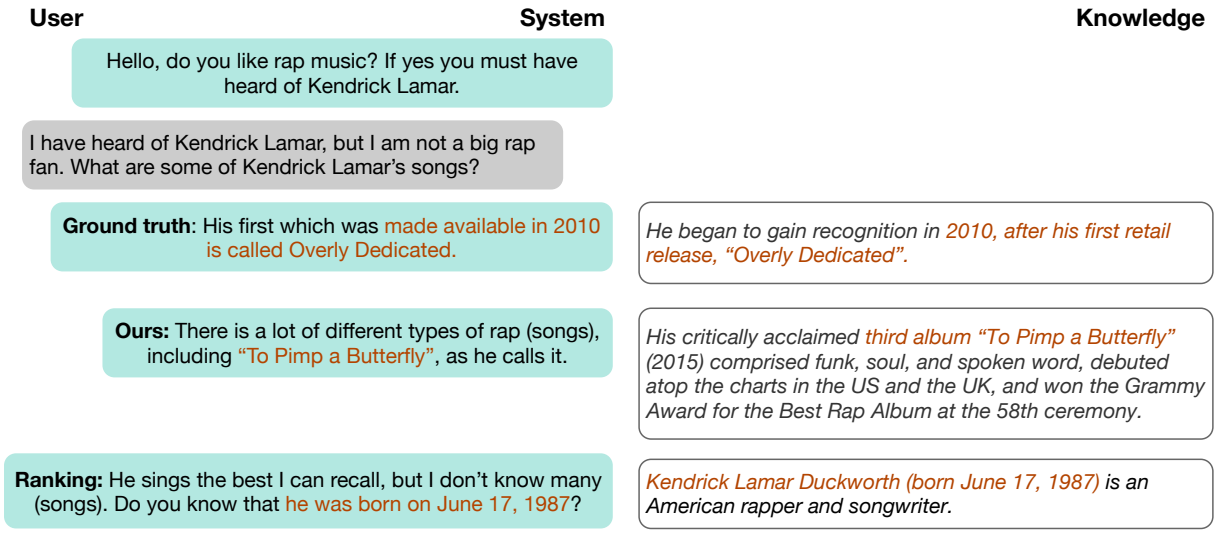

Figure 6: A case of mismatched but acceptable knowledge selection on WoW.

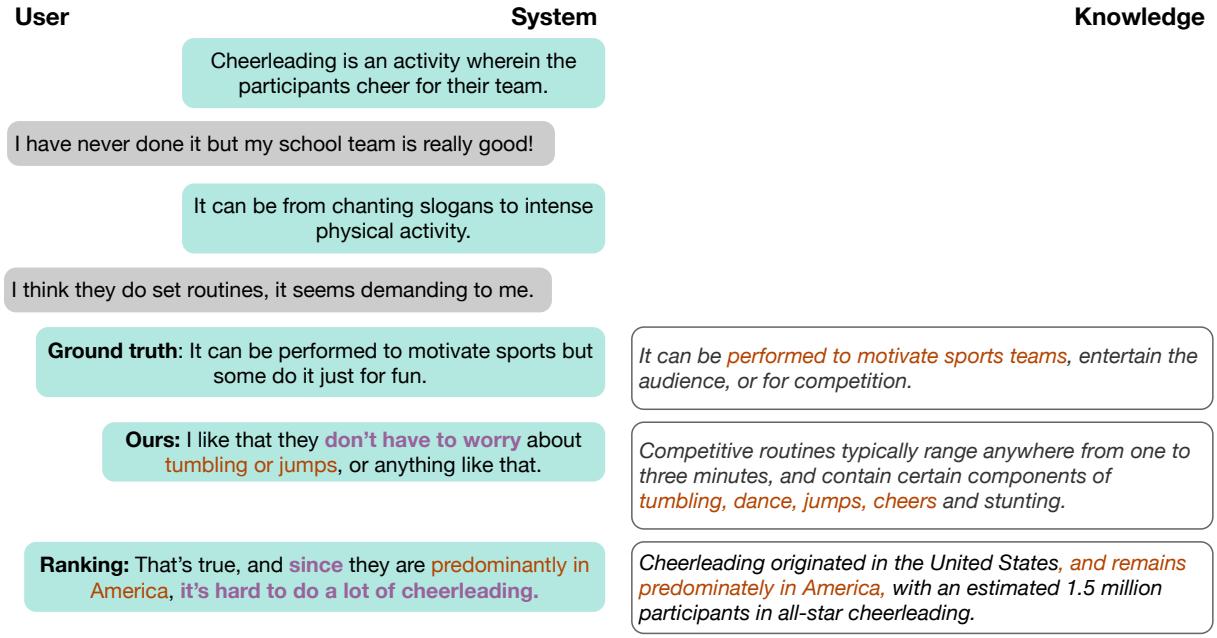

Figure 7: A case of response generation errors.The used knowledge is highlighted in brown and the generation error is marked in purple.

```
{'key': 'ts_521:5',
 'topic': 'The Sixth Sense',
 'history': ['What do you think about the characters in this movie?',
  'My favorite character was Cole',
  'The Young Cole Sear (Haley Joel Osment) was just amazing.',
  'He was more than amazing and really deserved the Teen Choice Awards2000Film -
Choice DramaFilm - Choice Breakout Performance',
  "He also won the Critics' Choice Movie Award for Best Young Performer.",
  ],
 'gold_responses': ['Awesome movie with a great story. Felt so sorry for Crowe m
ost especially because his relationship with his wife begins to fall apart due t
o his focus on work.'],
 'gold_knowledges': ['Meanwhile, his relationship with his wife begins to fall a
part due to his focus on work.',
  'Awesome movie with a great story. Felt so sorry for Crowe most especially bec
ause his relationship with his wife begins to fall apart due to his focus on wor
k.'],
 'system1': {'selected_knowledge': "When I first saw The Sixth Sense, I didn't k
now what to expect.",
  'response': 'And also won the MTV Movie Music Awards 2011-2014 for Best Female
Performer.'},
 'system2': {'selected_knowledge': 'The first scary movie I ever watched. Still
 gives me the chills. Fantastically written, filmed and acted.',
  'response': 'Yeah it was a scary movie. But the first scary movie I ever watch
ed.'},
 'system3': {'selected_knowledge': 'Empire Awards, UK 2000 Best Director.',
  'response': 'I think it won a empire awards, uk 2000.'}
```


Figure 8: Example of model output provided to annotators.