DeepMaven: Deep Question Answering on Long-Distance Movie/TV Show Videos with Multimedia Knowledge Extraction and Synthesis

Yi Fung\textsuperscript{1}, Han Wang\textsuperscript{2}, Tong Wang\textsuperscript{2}, Ali Kebarighotbi\textsuperscript{2}, Mohit Bansal\textsuperscript{2}, Heng Ji\textsuperscript{2}, Prem Natarajan\textsuperscript{2}

\textsuperscript{1}University of Illinois at Urbana Champaign, \textsuperscript{2}Amazon Alexa

yifung2@illinois.edu

\{wnghn,tonwng,alikeba,mobansal,jihj,premknat\}@amazon.com

Abstract

Long video content understanding poses a challenging set of research questions as it involves long-distance, cross-media reasoning and knowledge awareness. In this paper, we present a new benchmark for this problem domain, targeting the task of deep movie/TV question answering (QA) beyond previous work’s focus on simple plot summary and short video moment settings. We define several baselines based on direct retrieval of relevant context for long-distance movie QA. Observing that real-world QAs may require higher-order multi-hop inferences, we further propose a novel framework, called the \textsc{DeepMaven}, which extracts events, entities, and relations from the rich multimedia content in long videos to preconstruct movie knowledge graphs (movieKGs), and at the time of QA inference, complements general semantics with structured knowledge for more effective information retrieval and knowledge reasoning. We also introduce our recently collected \textsc{DeepMovieQA} dataset, including 1,000 long-form QA pairs from 41 hours of videos, to serve as a new and useful resource for future work. Empirical results show the DeepMaven performs competitively for both the new \textsc{DeepMovieQA} and the pre-existing MovieQA dataset.\textsuperscript{1}

1 Introduction

Our world tells an evolving story of people, objects, and their interactions. This storytelling may exist in various forms, from textual summaries and spoken dialogues, to accompanying images and videos. Because of its dynamically evolving and multi-media nature, long-distance video question answering on movies/TV shows provides a useful setting for studying the computational understanding of interconnected stories and events that aligns closely with real-world application scenarios. Yet, current intelligent systems still struggle with adequately processing the rich multimedia content in long videos and fail to answer many common inquiries that humans are interested in. If we ask a virtual assistant about the relationship between two main characters in a well-known movie/TV show, it likely defaults to some null response such as “\textit{hmm... I don’t know this one}”.

The research challenge of long video content understanding and question answering frameworks stems from the need to support information probing on certain specific details over a large multimedia context space. Propagating information across long-distance has been a well-known challenge due to the vanishing gradient and memory loss problem (Hochreiter, 1998). Long videos also can not fit its entire data in-memory typically for end-to-end feature extraction and neural network training. Meanwhile, retrieval-based methodologies, which first find sections in the video relevant to a query through semantic cues from the corresponding dialogues and/or frames, and then return the relevant textual dialogue and/or visual frame context for question answering, tend to narrowly isolate media instances semantically similar to the query inputs. Query inputs are short and succinct in nature, so simply considering data points that match the query semantics will overlook the larger picture behind the selected media instances \textit{i.e.} how they relate to each other as well as to other relevant media components that are initially missed out from retrieval as they involve additional reasoning.

For example, a question about “the hobby interests of Midge’s husband” in TV show “The Marvelous Mrs. Maisel” would likely be missed by text retrieval if the dialogue mostly uses the name of Midge’s husband - \textit{Joel}. A question about “Midge and Joel’s marriage breakdown” not only needs to identify the point of confrontation between \textit{Midge} and \textit{Joel}, but also other \textit{past, concurrent, or future} events connected to these two person entities,
<table>
<thead>
<tr>
<th>Category</th>
<th>Question (Q)</th>
<th>Natural Language Answer (A)</th>
<th>Visual Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>Who bailed Midge out of jail after she was arrested for public nudity? – “The Marvelous Mrs. Maisel”</td>
<td>Susie Meyerson, the manager of the Gaslight, bailed Midge out.</td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
<tr>
<td>Cross-Media</td>
<td>Why does Joel punch someone at Gaslight? – “The Marvelous Mrs. Maisel”</td>
<td>That person insulted Maisel during her act. Joel punches him and asserts that Maisel is good.</td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
<tr>
<td>Multi-Scene</td>
<td>How did Midge and Joel’s marriage fall apart? – “The Marvelous Mrs. Maisel”</td>
<td>Joel has been cheating on Midge, with his secretary Penny. One day, Joel got angry at Midge for giving suggestions that ended in a failed stand-up comedy performance for him. He packed up and left Midge.</td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
<tr>
<td>Multi-Hop</td>
<td>What does the Mayor oversee on the market? – “Les Miserables”</td>
<td>The mayor oversees which vendors are allowed on the market, what they are allowed to sell, how big their space is and what kind of stands should be added to the market. He charges the vendors his fees and have the police talk to vendors who act up.</td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
<tr>
<td>Background Knowledge</td>
<td>In real history, what ultimately happened to Malcolm X’s friend, singer Sam Cooke? – “One Night in Miami”</td>
<td>He was shot to death by a motel manager for attempting to molest a woman, but his death was controversial and involved conspiracy theories.</td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 1: Types of data annotated in our new DeepMovieQA benchmark, along with their statistics. The category pertains to the critical reasoning used in understanding the question, picking up relevant details from textual and/or visual media in the movie/TV show, and deriving to an answer.

Midge and Joel, relevant to their marriage dynamics, such as infidelity, pent-up frustration, and aftermaths. Previous work tends to focus on very high level questions with answers coming only from text summary, such as whether a character is good/bad, or focus on very superficial questions such as what object is behind this person, whereas we want to enable more interesting QAs such as from mimicking MovieClub conversations.

So in this work, we define a new task of deep movie question answering (DeepMovieQA), which gauges the content understanding of long movie/TV show videos that align with generic human inquiries. In contrast to previous work, the DeepMovieQA task consists of questions that naturally arise from watching the full length of a movie/TV show asset (as opposed to plot summary or short minute-long clips), and involve a more challenging set of reasoning summarized in Table 1, which make use of many different data modalities: plot summary, textual dialogues, visual frames, and background knowledge. DeepMovieQA additionally involves localizing the time frames that provide evidence for the question answering.

To tackle DeepMovieQA, we propose a novel framework, DeepMaven, which extracts context from multimedia video asset semantically relevant to the queries and generates a coherent answer conditioned on the retrieved context using a transformer-based backbone, with added benefits of transparency and explainability through this two-stage process. Observing that the new DeepMovieQA task involves more challenging types of reasoning, we further leverage a structured approach to incorporate wider context and better handle multi-hop queries, such as bridging the connection between the mention of Midge’s husband in a query and the canonical Joel entity.

Our key novel contributions can be summarized as follows: 1) we present a new research task of DeepMovieQA in the long-distance multimedia video understanding domain that involves reasoning beyond surface-level understanding of short moments and summaries; 2) we present DeepMaven, a novel multimedia approach that leverages complementary information from (local) dialogue passages and visual frames and (wider-context) movieKG subgraph to guide deeper content understanding and question answering, achieving 10+% absolute gain in answer extraction and relative gain in answer generation over single modality baselines; and finally, 3) we contribute a DeepMovieQA dataset that includes 1000 QA annotations, to serve as a new resource for studying
and evaluating this challenging but exciting task.

2 DeepMovieQA Data Collection

Pre-existing video/movie/TV show QA datasets fall into two extremes of either focusing on too narrow and minor scene-specific details such as an object on the couch that is irrelevant to the storyline development (Lei et al., 2018), or overly brief and simple details from summaries (Tapaswi et al., 2016) that align poorly with real-world needs such as the year in which Hook’s squad is sent to Belfast, which is written directly in the short plot summary. In comparison, we aim for QAs closer to the conversations in movie clubs or what we talk about when we step out of the movie theatre i.e. the detailed yet crucial information to main event developments. This means we want question annotations that involve longer context of video clips and answer length, compared to prior work as reported in Table 2.

With this motivation in mind, we collect our dataset, DeepMovieQA, on 4 movies and 5 TV shows listed in Table 3, chosen based on genre diversity. We instruct annotators to come up with QA pairs that arise naturally as they watch the movie/TV shows, through two rounds. In the initial pass, we provide ten annotators movie/TV show assets split into 20 minute video clips for focused attention on detailed plot content. In the second pass, we ask the annotators to re-watch the entire movie/TV show asset, review QAs previously annotated, and come up with interesting QAs that involve piecing information together from several scenes or sources – such as discussion of themes, messages, empathetic reactions\textsuperscript{2}, and societal background. This two-pass QA labeling process allows annotators to become more familiar with the movie/TV show content and come up with challenging QAs that better align with genuine human audience interests. DeepMovieQA provides the first long-distance, multi-scene QA benchmark in the multimedia setting, with long-form answers labeled to promote elaborative movie QA model capabilities.

3 Methodology

3.1 Direct Retrieval from Source Data

Given a movie or TV show \( V = \{V_f, V_{dial}, V_S\} \), consisting of the video frames, dialogues, and summaries, and a question, \( q \), a natural first step is to select relevant context through semantic matching of local features. On the text side, we draw inspirations from previous research (Qu et al., 2020; Mossad et al., 2020) to encode the query, \( q \), and candidate passages \( V_{dial,i} \), truncated from every \( n = 5 \) utterance exchange (a tweakable hyperparameter), using \( BERT \) (Devlin et al., 2019), which are expressive bidirectional transformers for capturing latent language representations, followed by two separate linear layers, \( \phi_q \) and \( \phi_p \). Then, we compute the retriever matching score through the cosine similarity of the encoded representations.

\[
h_q = \phi_q(ReLU(BERT(q)))
\]

\[
h_p = \phi_p(ReLU(BERT(V_{dial,i})))
\]

\[
s_{q, V_{dial,i}} = \text{cos}_\text{sim}(h_q, h_p)
\]

For visual retrieval, we utilize the powerful multimedia CLiP encoder release (Radford et al., 2021), which consists of a textual encoder component \( CLiP_t \) based on GPT (Radford et al., 2019) and a visual encoder component \( CLiP_v \) based on ViT.

\[
\text{Video Asset} & \text{# Eps} & \text{# Min.} \\
\hline
\text{Manchester by the Sea (MBTS)} & - & 137 \\
\text{Without Remorse (WORE)} & - & 109 \\
\text{One Night in Miami (ONIM)} & - & 114 \\
\text{Les Miserables (LMIS)} & - & 105 \\
\text{Marvelous Mrs. Maisel (MRSM), S1} & 8 & 419 \\
\text{Jack Ryan (JKRY), S1} & 8 & 400 \\
\text{The Boys (TBYZ), S1} & 8 & 361 \\
\text{Transparent (TP), S1} & 10 & 287 \\
\text{The Family Man (FMAN), S1} & 10 & 448 \\
\text{Total} & - & 2480 \\
\end{array}
\]

Table 2: Comparison of DeepMovieQA with respect to pre-existing datasets in the movie/TV domain. The # of mins and # of tokens listed are averaged numbers.

<table>
<thead>
<tr>
<th>Benchmark/ Dataset</th>
<th>Total Hrs</th>
<th># Min. per Video Clip</th>
<th># Tokens in Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MovieQA (Tapaswi et al., 2016)</td>
<td>280</td>
<td>-</td>
<td>5.6</td>
</tr>
<tr>
<td>TVQA (Lei et al., 2018)</td>
<td>461</td>
<td>1.3</td>
<td>5.1</td>
</tr>
<tr>
<td>TVQA+ (Lei et al., 2020a)</td>
<td>461</td>
<td>1.3</td>
<td>5.1</td>
</tr>
<tr>
<td>TVR (Lei et al., 2020b)</td>
<td>461</td>
<td>1.3</td>
<td>-</td>
</tr>
<tr>
<td>KnowIT VQA (Garcia et al., 2020)</td>
<td>69</td>
<td>0.3</td>
<td>4.5</td>
</tr>
<tr>
<td>DeepMovieQA (ours)</td>
<td>41</td>
<td>51.7</td>
<td>26.5</td>
</tr>
</tbody>
</table>

Table 3: The movies/TV shows in DeepMovieQA.

\textsuperscript{2}Similar to https://teachwithmovies.org/discussion-questions-for-use-with-any-film-that-is-a-work-of-fiction/
(Dosovitskiy et al., 2020), pretrained through contrastive learning on 400 million image caption pairs. We compare for semantic proximity between $q$, and a set of candidate visual frames, $V_f$, through cosine similarity, similar to text retrieval. But note that as each candidate $V_f$ is actually a set of visual frames corresponding to the time frame under $V_{dial}$, we take the mean of encoded image representations as the overall feature for $V_f$.

$$h_q = CLIP(q)$$

$$h_{V_f} = \frac{1}{|V_f|} \sum_{j \in 1..|V_f|} CLIP(V_f_j)$$

$$s_q V_f = \cos \_sim(h_q, h_{V_f})$$

As dialogue exchanges and visual scenes may contain cues that complement each other in signaling whether a section of the video is relevant to the query, we utilize a linear combination of the textual and visual retrieval scores for multimedia retrieval:

$$s_q(V_{dial}, V_f) = a \times s_q V_{dial} + b \times s_q V_f$$

We select the $V_{dial}$ and/or $V_f$ with top $k$ semantic matching score for query $q$ as the relevant context.

### 3.2 Retrieval from Structured Knowledge

Yet, structured knowledge provides benefits for long-distance and multi-hop information since events and interactions at separate time points (see Table 1 for examples) may then be directly connected through grounded nodes such as character entities.

**Multimedia KG Construction** To incorporate structured knowledge, we pre-construct movieKG using the open-source IE pipeline from (Wen et al., 2021) to extract events/entities/relations (Lin et al., 2020) from movie summary and dialogue, link entities to background knowledge base (Pan et al., 2015) where applicable, and perform event/entity coreference resolution (Lai et al., 2021). This leads to an initial sparse $KG$, in which the nodes consist of $N_{e}$ entities and $N_{r}$ while the edges consist of argument roles and relations $(E_{r})$, following a pre-defined ontology which inadequately covers the open-domain in diverse film genres, so we augment ontological-guided IE with Abstract Meaning Representation (AMR) parsing (Fernandez Astudillo et al., 2020; Zhang and Ji, 2021) on the movie dialogue and summaries. For example, let’s consider a subgraph extracted from IE containing events such as “<Midge, arrested (by), the police>”. IE has the advantage of performing entity/event linking, so the entity nodes have other direct connections in the constructed MovieKG as well, such as “<Midge, located in, Manhattan>” and “<police, (also) arrest, comedian Larry Bruce>”. However, other important events may not be captured by the IE ontology due to its more abstract or rarer occurrence in daily life and important news events, such as “<Midge, bailed out by, Gaslight manager Susie Meyerson>” and “<Gaslight manager Susie Meyerson, recognize, Midge’s talent>”, even though these information triplets can be directly extracted as the noun/verb-form concept nodes from AMR parsing on plot summary and dialogue transcripts. Hence, we add subgraphs from AMR parsing to the movieKG initially constructed from IE where there exists a coreferential event or entity node, such as “Midge”.

To further enrich this KG with visual information, we perform event extraction using grounded image situation recognition and localization (Pratt et al., 2020), which extracts the verb in action from the visual frame, as well as the semantic roles of objects detected (agent, item, destination, place, etc). We also perform character mapping for agents that play a role in visual events by finding the closest match from a bank of character profile images, using visual features extracted by an iResNet model (Duta et al., 2021) following Meng et al. (2021). Finally, we merge objects and events that occur across textual and visual media based on embedding similarity in a multimedia common semantic space computed from CLiP, introduced in Sec 4.1.

**Knowledge Subgraph Retrieval** Now, we cannot simply select the neighborhood surrounding seed nodes that best match question $q$ as the relevant subgraph because recurring entities contain many dense connections in the long story-telling domain. So instead, we select relevant subgraph based on context-aware saliency. Given a natural language question, we first represent it as a query graph that has undergone knowledge extraction for closer comparability with the movieKG. For instance, a question about “the initial encounter between Midge and Larry Bruce” becomes a graph with two connections: [ Midge—the initial encounter, the initial encounter–Larry Bruce]. At the time of probing, we compute contextualized embeddings of node mentions, from BERT

3These can be manually identified from video frames or automatically collected from www.imdb.com.
encoders with awareness of the sentence that they occur in, as the local features for both the query graph and the movieKG, concatenate with wider-context knowledge embeddings from a two-hop neighborhood aggregation of maximally aligning neighbor nodes’ local embedding with respect to the query graph nodes. For query graph nodes, we concatenate the sentence-level BERT embedding as their wider context features. We score and rank \(KG_{subgraph}\) selection based on the semantic (cosine) similarity of these knowledge embeddings between nodes in the movieKG and query graph.

### 3.3 Combining Context for Answer Fusion

In this work, we regard extractive QA as the task of selecting the relevant \(V_{\text{dia}},\) and \(V_{f}\) (as well as summary sentences \(V_{S}\) and movieKG \(KG_{subgraph}\) though these lack labels for evaluation) from a movie/TV show, given question \(q\). We formulate abstractive QA as the task of natural language answer generation, conditioned on these relevant \(V_{\text{dia}},\) \(V_{f},\) movieKG \(KG_{subgraph}\), and \(V_{S}\), providing complementary context to each other. Though it is tempting to learn an answer generation model that intakes context retrieved from different modalities through a straightforward common semantic space, we note the low-resource setting of our challenging task (in which annotations are expensive and time-consuming to obtain). Thus, we take advantage of pretrained conditional text generation transformers, in particular, BART, which has an encoder to extract context information and a generative decoder for sequential token generation, as a suitable backbone for answer generation (Lewis et al., 2020; Khashabi et al., 2020). Moreover, we aim to match the format of input that robust and high-performing conditional text generators have been pretrained on, which is natural language text. While \(q, V_{\text{dia}},\) and \(V_{S}\) directly fit this desirable input format, retrieved \(V_{f}\) and movieKG \(KG_{subgraph}\) should be projected into a textual semantic space for optimal alignment with our pretrained backbone answer generator. Hence, for the retrieved movieKG \(KG_{subgraph}\), we take a stringified representation of its structured connections following (Ribeiro et al., 2020). For the retrieved video frames, we make sure that knowledge elements in these images are included in or concatenated to the stringified movieKG subgraph representation. Although our transformer-based answer fusion mechanism may be relatively simple and straightforward, we observe it can handle QAs such as the ones in Table 6, detailed further in Sec 4. Figure 1 provides a walk-through illustration of our overall framework, which we refer to as the DEEPMAVEN. It is worthy to note that by nature of our information extraction-based QA reasoning, our textual answer generation is inherently supplemented with grounding to the visual context retrieved. This includes bounding box localization of the actions and participants (e.g., event – ‘drinking’, person – ‘Midge Maisel’, etc.) from the video frames.

### 4 Experiments

#### 4.1 Benchmark and Dataset

**DeepMovieQA Corpus** This is our newly constructed dataset from 4 TV shows and 4 movies. The QAs may involve deep content beyond plot summary information, background knowledge, and higher-order reasoning across different scenes and events, as well as cross-media inferences. The answers are designed to be conversational friendly in nature and have an average token length of 19. A separate expert annotator manually checked 100 random QA pairs and judged all of them as accurate, informative, and comprehensive. We used a 8:1:1 train/val/test data split due to the corpus size.

**Shallow MovieQA Corpus** This is a multiple choice QA dataset (Tapaswi et al., 2016) annotated solely from movie plot summaries, with a pre-established 66-13-21% train/val/test data split.

#### 4.2 Experimental Setting

For retrieval (extractive question answering), we include single modality approaches as the natural baselines to our proposed model, DEEPMAVEN.

- **Baseline 1:** Dialogue Component Only
  Here, we simply retrieve the relevant dialogue sections from BERT embedding similarity with respect to the input question, similar to the top performing approach (Mossad et al., 2020) in MovieQA leaderboard.

- **Baseline 2:** Visual Component Only
  Similarly, here, we use the features from CLiP for semantic matching between the question and visual frames.

- **Baseline 3:** Textual & Visual Component w/o Structured Knowledge
  We further perform comparison with a simplified version of our DeepMaven framework that focuses on the cross-media dialogue and
visual frames without structured knowledge from the movie KG\textsubscript{subgraph}.

For abstractive question answer generation, we consider the base variant of DEEPMAVEN, which finetunes BART (Lewis et al., 2020), using the textual dialogue retrieved from movies/TV shows as input context only. We also include the baseline of UniVL (Luo et al., 2020), which has been pretrained on video and language multimodal feature representation and text generation.

4.3 Evaluation Metrics

For long-distance retrieval, we report the hit@K metrics of whether the selected section of the dialogue or visual frames fall under the source time interval from the QA annotation. For natural language answer generation, we compute the ROUGE-L (Lin, 2004) F-scores\footnote{www.pypi.org/project/pycocoevalcap/}. However, there might be many variants of words and phrases with different level of granularities that can be used to describe the same answer in the same scenes. Take the question about “how Midge’s parent react to the news that Joel has left Midge” in “The Marvelous Mrs Maisel” for example. The ground truth answer annotation provides detailed descriptions on how “Midge’s dad starts to play the piano frantically and later blames Midge for marrying a weak man against his advice” while “Midge’s mother starts crying and whining heavily”. We observed that our system generated answers are more general but still correct, outputting “the way the parents are reacting is very upset and blaming each other”. A strict measurement of n-gram overlaps from ROUGE would not credit such an answer generation sufficiently. Therefore, we also include a semantic-based BLEURT \footnote{We use the BLEURT-20 scorer model from https://github.com/google-research/bleurt.} metric (Sellam et al., 2020). In addition, we conduct a human assessment on the extracted answers as well. For each pair of system extracted answer and ground-truth answer, we ask human assessors to judge whether the system answer is completely correct, partially correct or incorrect.

4.4 Quantitative Results and Analysis

Extractive Question Answer Retrieval As shown in Table 4, retrieval is a non-trivial task in the movie setting, which contains lengthy plot content, and a random approach has a very low hit@K=5 of 0.08. Single modality retrieval baselines perform similarly, with text retrieval using BERT and visual retrieval using CLiP having hit@K=5 of 0.32 and 0.28 respectively, while cross-media retrieval achieves a noticeable boost to hit@K=5 of 0.41, suggesting that textual and visual features offer similar levels of useful signals that complement each other for making sense of
Table 4: These are the video frame retrieval results, with * as the baselines, ∆ as the simplified DeepMaven without structured knowledge, and bold as the full DeepMaven retrieval approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Hit@K=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.08</td>
</tr>
<tr>
<td>Textual (BERT)*</td>
<td>0.32</td>
</tr>
<tr>
<td>Visual (CLIP)*</td>
<td>0.28</td>
</tr>
<tr>
<td>Textual+Visual</td>
<td>0.41</td>
</tr>
<tr>
<td>Textual+Visual+KG priming</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 5: Answer generation (%) scores on our DeepMovieQA dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Rouge-1</th>
<th>Rouge-L</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniVL</td>
<td>16.8</td>
<td>15.1</td>
<td>22.3</td>
</tr>
<tr>
<td>DeepMaven</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- dial. only</td>
<td>13.9</td>
<td>11.1</td>
<td>22.1</td>
</tr>
<tr>
<td>- all</td>
<td>21.6</td>
<td>17.4</td>
<td>31.7</td>
</tr>
</tbody>
</table>

Table 6: Example DeepMaven abstractive QA results, with the generated and ground truth answers.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who comes to the officers’ rescue?</td>
<td>Two United States marines, Dave Karnes and Jason Thomas.</td>
</tr>
<tr>
<td>Where are Stigman and Trench taken after being captured?</td>
<td>2 Guns</td>
</tr>
<tr>
<td>How does Joel leave the apartment after they fell asleep the night before?</td>
<td>Mexico</td>
</tr>
<tr>
<td>What were in the boxes in the back of the truck Sajid was driving?</td>
<td>They were filled with nerve gas.</td>
</tr>
</tbody>
</table>

4.5 Qualitative Analysis

Table 6 shows answer generation results while Table 7 shows content selection examples, performed by DeepMaven. We additionally conduct a human assessment on the DeepMovieQA answer generations from held-out data, rating answers on a likert scale of 0 – 5 (with 0 being nonsense and 5 being perfect). Thirteen percent of the generated answers scored ≥ 4, with a Kappa coefficient of 0.42, reflecting moderate inter-annotator agreement.

4.6 Remaining Challenges

In general, we found that DeepMaven retrieval performs worse on QAs that involve commonsense reasoning and reference to objects/entities that may not be easily picked up by our structure-guided cross-media semantic matching system. Some deeper movie QAs may also be more suitable for the conversational question answering setting due to insufficiently detailed question wording and open-ended answer form. Finally, we observed that long-form answer generation with a limited-sized training data is more prone to hallucination. Given questions such as “What did Susie promise to one of the waiters at the Copacabana Club in return for smuggling her into the club for free?”, the language model generator backbone tends to fill in some plausible pre-trained knowledge such as “Susie promises to buy one of the men a drink after every act” that may not match the actual detailed answer of “Susie promised him a prime slot for his singing act at the Gaslight Café for 2 weeks. Unfortunately Susie thinks his act is awful but keeps her word even since they had to watch the show from the kitchen”. Minimizing hallucination in long-form low-resource question answer generation is an important issue that merits future investigation.

5 Related Work

Text QA: QA has been a popular task (Rajpurkar et al., 2016), with significant advances made recently by attention-guided transformers (Qu et al., 2019). Several corpora have been proposed for QA in story comprehension. FriendsQA (Yang and Choi, 2019) bases QA on short dialogue exchanges. MovieQA (Tapaswi et al., 2016) and NarrativeQA (Kocisky et al., 2018) are annotated from plot summaries. (Zhou et al., 2018; Moon et al., 2019; Zhou...
et al., 2020) annotated open-ended conversations with groundings to either Wikipedia pages, or pre-existing structured knowledge graphs such as Freebase (Bollacker et al., 2008) and XLORE (Wang et al., 2013). These benchmarks overlook the full content from the original data, which we address. **Visual QA:** The visual question answering (VQA) task (Antol et al., 2015) aims to predict a natural language answer, given a natural language question and image(s) without other textual context. Various datasets have been constructed for this task, such as VQA 2.0 (Goyal et al., 2017), VCR (Zellers et al., 2019), and scientific PlotQA (Methani et al., 2020), in the setting of single images; as well as MSR-VTT-QA (Xu et al., 2016), MovieFiB (Maharaj et al., 2017), and VideoQA (Zhu et al., 2017), in the setting of visual frame sequences. The answers are typically multiple choices or from a predefined vocabulary. Simple baseline methods that only use question understanding (Kazemi and Elqursh, 2017) or sentiment analysis on answer options (Manjunatha et al., 2019) have proven surprisingly well on datasets such as VQA and VQA 2.0 but are unlikely to provide good answers for understanding complex events and person interactions.

**Multimedia QA:** There are increasing interest nowadays in using information from multiple modalities for answering questions, such as reasoning through text, images, and tables in Many-ModalQA (Hannan et al., 2020), MultiModalQA (Talmor et al., 2021), and MuMuQA (Reddy et al., 2022). The (movie/TV) video domain is a further step that involves dynamic events from dialogues and corresponding visual frames. TVQA (Lei et al., 2018) presents a multimedia QA dataset grounded on minute-long video clip snippets from popular TV show, with frame localization added in TVQA+ (Lei et al., 2020a). However, their QAs involve highly scene-specific discussion points, trivial to the deeper contents around the central plotline, e.g. “Q: What is on the couch behind Joey when he is at the counter? A: A soccer ball.” In contrast, our work highlights deeper knowledge and multi-frame information synthesis.

**Other Story-based Video Understanding:** Benchmarks such as MovieNet (Huang et al., 2020) and LVU (Wu and Krahenbuhl, 2021) gauge a variety of content classification tasks related to the ‘scene/place’, ‘cinematic style’, ‘genre’, ‘produced year’, ‘popularity’, etc. Moreover, Lei et al. (2020b); Huang et al. (2020); Bain et al. (2020) explore retrieval based on sentence-length

<table>
<thead>
<tr>
<th>Cue Category</th>
<th>Question</th>
<th>Retrieved Dialogue Passages and Visual Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Content</td>
<td>What did Kelly threaten Clay with after forcing him into his car? ←WORE</td>
<td>“Why my family? My wife? My daughter? Why kill my team? Rykov was all in, so I’d like a little bit of what you sold him... Know where we driving? You got a farm out in West Virginia, right? Your daughter should be home from college. Wesleyan. That’s a good school. Your wife. Your son too. So I suggest you start talking.”</td>
</tr>
<tr>
<td>Implicit Cross-Media Entity Grounding</td>
<td>What does Midge’s mother want to talk about after Midge arrives at her parent’s apartment? ←MRS</td>
<td>“Thanks for taking the kids last night. Were they okay? - We need to talk about the baby. Why? What’s the matter with her? That forehead is not improving. - What? Are you sure?”</td>
</tr>
<tr>
<td>Implicit Multi-hop Event Reasoning</td>
<td>What does Midge do after Joel falls asleep and before he wakes up again? ←MRS</td>
<td>“Good night, Gracie. Hey. Good morning. - Did the alarm go off? - It sure did. Wow. I didn’t hear it at all. You never do. - Good morning, Jerry. - Good morning, Mrs. Maisel.”</td>
</tr>
<tr>
<td>Failure to Identify Certain Visual Objects</td>
<td>What does Midge suggest in the taxi to improve Joel’s performance? ←MRS</td>
<td>“Maybe you should write a beginning, something that says who you are or something. What do you think? Good evening. What a nice... Good evening, ladies and gentlemen. Thank you for the nice... nice... ‘Nice’ is a bad, bad word. All that applause for me? What am I, putting out after?...”</td>
</tr>
<tr>
<td>Mismatch in Level of Detail Between Query and Context</td>
<td>What did Midge do while her friend Imogene is visiting to make sure she’s still in shape? ←MRS</td>
<td>“She’s going on and on about this miracle treatment she had done in Mexico. It involved goat’s milk and avocados. Right ankle 8, left ankle 8. They smear it on your face, wrap a hot towel around your head...”</td>
</tr>
</tbody>
</table>

Table 7: Impressive and Erroneous Examples of Retrieval Results.
scene description while Liu et al. (2020) study video-language inference, but these are all less challenging than the natural question answering setting our work focuses on.

6 Conclusions and Future Work

This work is the first to study long-distance movie question answering. Through guidance from multimedia source context and structured knowledge retrieval, our proposed model, DEEPMAVEN, is shown to perform well for extractive question answering, as well as abstractive answer generation for the questions in pre-existing MovieQA benchmarks. But improvements are needed for generating answers to the more challenging questions presented by our new DeepMovieQA benchmark, which serve as a better reflection of real-world application settings with more reasoning involved. We aim to kickstart an extractive movie question answering interface for human users, and through this, naturally acquire genuine movie questions for more effectively expanding DeepMovieQA annotation, and later extend this into interactive conversational AI settings.

7 Limitations

Our work focuses on video question answering that aims to be detailed and engaging. It is not interactive in nature like conversation exchanges. Finally, we bear in mind that automatically generated answers may contain the risk of insensitive phrasing, and mitigating language model bias deserves further exploration and efforts.

References


the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pages 539–548.


