

DramaQA: Character-Centered Video Story Understanding with Hierarchical QA

Seongho Choi¹, Kyoung-Woon On¹, Yu-Jung Heo¹, Ahjeong Seo¹, Youwon Jang¹, Seungchan Lee¹, Minsu Lee¹, and Byoung-Tak Zhang^{1,2}

¹ Seoul National University, Seoul 08826, South Korea

² AI Institute (AIIS), Seoul 08826, South Korea

{shchoi,kwon,yjheo,ajseo,ywjang}@bi.snu.ac.kr, dnfdfn1@snu.ac.kr,
{mslee,btzhang}@bi.snu.ac.kr

Abstract. Despite recent progress on computer vision and natural language processing, developing video understanding intelligence is still hard to achieve due to the intrinsic difficulty of story in video. Moreover, there is not a theoretical metric for evaluating the degree of video understanding. In this paper, we propose a novel video question answering (Video QA) task, DramaQA, for a comprehensive understanding of the video story. The DramaQA focused on two perspectives: 1) hierarchical QAs as an evaluation metric based on the cognitive developmental stages of human intelligence. 2) character-centered video annotations to model local coherence of the story. Our dataset is built upon the TV drama “Another Miss Oh”³ and it contains 16,191 QA pairs from 23,928 various length video clips, with each QA pair belonging to one of four difficulty levels. We provide 217,308 annotated images with rich character-centered annotations, including visual bounding boxes, behaviors, and emotions of main characters, and coreference resolved scripts. Additionally, we provide analyses of the dataset as well as Dual Matching Multistream model which effectively learns character-centered representations of video to answer questions about the video. We are planning to release our dataset and model publicly for research purposes and expect that our work will provide a new perspective on video story understanding research.

Keywords: Video Question and Answering, Evaluation Metric for QA, Character-Centered Video Annotation

1 Introduction

A story is a series of events across multiple scenes centered around a succession of character’s actions. Given that humans communicate regularly and naturally through stories, the ability to understand stories is a crucial part of human intelligence that sets humans apart from others [35,38].

Among various kinds of narratives, *dramas*, especially in the form of video, are considered one of the best narrative mediums for developing human-level

³ We received official permission to use these episodes for research purposes from the content provider.

AI algorithms from two points of view. Firstly, they have multiple modalities such as a sequence of images, audio (including dialogue, sound effects, and background music) and text (subtitles or added comments). Secondly, they show cross-sections of everyday life by demonstrating socioculturally appropriate behaviors through the characters. However, understanding video stories is considered to be challenging to current machine learning methods, due to the causal and temporal relationships between events, which can be complex and are often left implicit [27].

One way to enable a machine to understand a video story is to train the machine to answer questions about the video story [30,21]. To train the model, a large number of question-answer pairs for video stories are required. While several recent studies have suggested Video Question Answering (Video QA) datasets [36,13,22,8,16], these datasets do not give sufficiently careful consideration to two aspects of video story understanding. First, the collected QAs in the previous studies are highly-biased and lack variance in question difficulty. However, QAs with hierarchical difficulty levels are crucial, as people with different levels of intelligence will understand the given video story differently [4]. Second, previous works did not provide any consistent annotations for characters to model this coherence. However, focalizing on characters plays an important role in forming local story coherence, because stories are represented by a sequence of actions that characters perform [28,5].

In this work, we propose a new Video QA task, DramaQA, for more comprehensive understanding of video story. 1) We focus on the *understanding* with hierarchical QAs used as a hierarchical evaluation metric based on cognitive developmental stages of human intelligence. We defined the level of understanding in conjunction with Piaget’s theory [4] and collected QAs accordingly. In accordance with [7], we classified questions into one of four hierarchical stages, based on two criteria; memory capacity (MC) and logical complexity (LC). With these hierarchical QAs, we offer a more sophisticated evaluation metric to measure understanding levels of Video QA models. 2) We focus on the *story* with character-centered video annotations. To learn character-centered video representations, the DramaQA provides rich annotations for main characters such as visual bounding boxes, behaviors, and emotions of main characters and also coreference resolved scripts. By sharing character names for all the annotations including QAs, the model can have a coherent view of characters in the video story. 3) We provide a Dual Matching Multistream model to answer questions for video story by utilizing the character-centered video annotations. Using both a context matching module and a character matching module, our model efficiently learns underlying correlations between the video clips, QAs and characters.

In the subsequent sections, related works for previous Video QA datasets and models are reviewed, then the DramaQA dataset is formally introduced. Next, we outline details of the dataset and experimental results from our proposed Dual Matching Multistream model.

2 Related Works

Video understanding area has been actively studied and several tasks including datasets have been proposed such as action recognition [9,34,10,33,2,11], video classification [1], video captioning [39,20,15], and spatio-temporal object segmentation [25]. However, these researches focus on perceiving and recognizing visual elements so that they are not suitable for high-level visual reasoning. To circumvent this limitation, video question answering for short video clips are proposed as a benchmark for high-level video understanding [29,8,22]. These works only dealt with a sequence of images about short video clips, which would not include a meaningful story. Unlike these video question answering, Video Story Question Answering focuses on the story narrative about video. A story is a sequence of events, which means video with meaningful stories contain a relatively long sequence of videos and consists of a series of correlated video events. Video Story Question Answering requires the ability to discriminate what's meaningful in a very long video, and also requires visual processing, natural language, and additional acoustic modeling. Recently, there have been some studies proposing datasets for use in the domain of video story understanding [36,13,16].

MovieQA [36] uses the plot synopses of the movies to create Video QA datasets. In addition to movie clips, the dataset contains various data such as plots, subtitles and DVS. However, since QA pairs are generated based only on the given complicated plot synopses, it is not easy to answer the questions by using an insufficient amount of the dataset. The PororoQA [13] dataset was created by directly watching animation videos (not real-world videos), making their QAs more tightly coupled to these videos. Most of these questions were very similar to subtitles and descriptions. TVQA [16] is a large-scale video QA dataset based on 6 TV shows. TVQA aims to utilize both visual and language information from 60-90 second video clips. They also inform users which specific part of the video is needed for answering questions, which is essential ground truth for localization. Most of their questions are focused on relatively short moments (less than 15 seconds) which is not targeted for story understanding.

In this context, our research suggests Video Story QA dataset with carefully designed evaluation metric and provides a character-centered video annotations to solve it.

3 DramaQA Dataset

3.1 Overview

Drama is a genre of narrative that can be described as a series of events consisting of several main characters. These characteristics of drama make it a suitable target for video story research. We collected the dataset on a popular Korean drama *Another Miss Oh*, which has 18 episodes, 20.5 hours in total. DramaQA dataset consists of sequences of video frames (3 frames per second), character-centered video annotations, and QA pairs with hierarchical difficulty levels. Figure 1 illustrates the DramaQA dataset. We also present a comparison of our dataset to

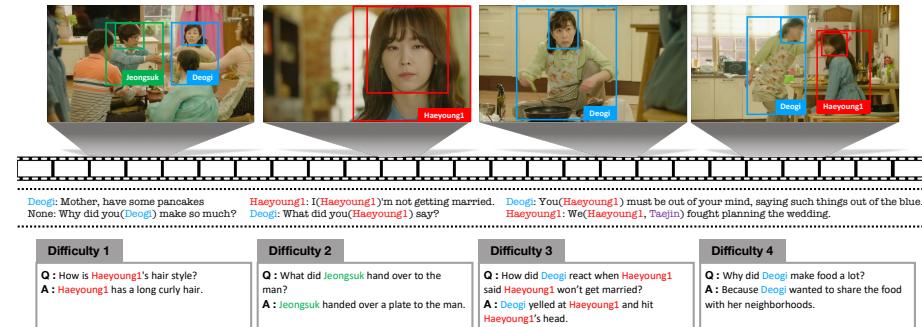


Fig. 1. An example of DramaQA dataset which contains video clips, scripts, and QA pairs with levels of difficulty. A pair of QA corresponds to either a shot or a scene, and each QA is assigned one out of a possible four stages of difficulty (details in Section 3.2). A video clip consists of a sequence of images with visual annotations centering the main characters.

some recently proposed video QA datasets (Table 1). Among the datasets, only the DramaQA provides difficulty levels of the questions and rich information of characters including coreference resolved scripts.

In Section 3.2 and 3.3, we give detailed descriptions about QA hierarchy and character-centered annotations.

3.2 Question-Answer Hierarchy for Levels of Difficulty

To classify question-answer pairs into hierarchical levels of understanding, we propose two criteria: Memory capacity and Logical complexity. Memory capacity (MC) is defined as the required length of the video clip to answer the question, and corresponds to working memory in human cognitive process. Logical complexity (LC) is defined by the number of logical reasoning steps required to answer the question, which is in line with Piaget's developmental stage [26].

Criterion 1: Memory Capacity The capacity of working memory increases gradually over childhood, as does cognitive and reasoning ability required for higher level responses [3,19,23]. From the machine learning perspective, generalizing data gets harder when the search space is increased so that the longer video story to answer a question requires, the harder to reason the answer from the video story is. Here, we consider two levels of memory capacity; shot and scene. Detailed definitions of each level are below:

- **Level 1 (shot):** The questions for this level are based on video clips less than about 10 seconds long, shot from a single camera angle. This set of questions can contain atomic or functional/meaningful action in the video. Most recent datasets which deal with video belong to this level [8,18,22].

Table 1. Comparison between video story QA datasets. Only DramaQA dataset provides hierarchical QAs and character based visual metadata. Average target video length for single QA is divided into shot level and scene level, since we provide multi level QA.

	# Q	# Annotated Images	Avg. Video len. (s)	Textual metadata	Visual metadata	Q. lev
MovieQA[36]	6,462	-	202.7	Plot, DVS, Subtitle	-	-
PororoQA[13]	8,913	-	1.4	Description, Subtitle	-	-
TVQA[16]	152,545	-	76.2	Subtitle	-	-
TVQA+[17]	29,383	148,468	61.49	Subtitle	Obj. Bbox	
DramaQA	16,191	217,308	3.7 ^a 91.3 ^b	Script [*]	Char. Bbox Behavior, Emotion	✓

^a Average video length for shot

^b Average video length for scene

^{*} Coreference resolved script

- **Level 2 (scene):** The questions for this level is based on clips that are about 1-10 minutes long without location change. Videos at this level contain sequences of actions, which augment the shots from Level 1. We consider this level as the “story” level according to our working definition of story. MovieQA [36] and TVQA [16] are the only datasets which belong to this level.

Criterion 2: Logical Complexity Complicated questions often require more (or higher) logical reasoning steps than simple questions. In a similar vein, if a question needs only a single supporting fact with a single relevant datum, we regard this question as having low logical complexity. Here, we define four levels of logical complexity from simple recall to high-level reasoning, similar to hierarchical stages of human development [32].

- **Level 1 (Simple recall on one cue):** The questions at this level can be answered using simple recall; requiring only one supporting fact. Supporting facts are represented as triplets in form of $\{subject\text{-}relationship\text{-}object\}$ such as $\{person\text{-}hold\text{-}cup\}$.
- **Level 2 (Simple analysis on multiple cues):** These questions require recall of multiple supporting facts, which trigger simple inference. For example, two supporting facts $\{tom\text{-}in\text{-}kitchen\}$ and $\{tom\text{-}grab\text{-}tissue\}$ are referenced to answer “Where does Tom grab the tissue?”.
- **Level 3 (Intermediate cognition on dependent multiple cues):** The questions at this level require multiple supporting facts with time factor to answer. Accordingly, the questions at this level cover how situations have changed and subjects have acted.

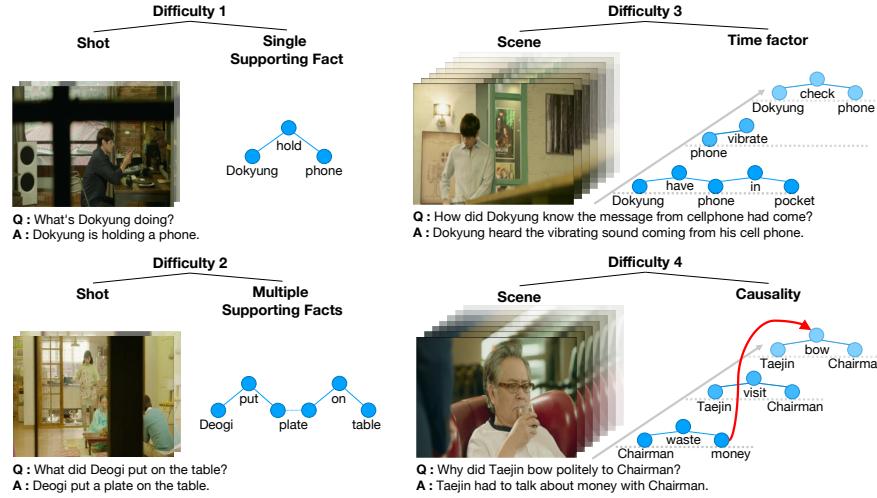


Fig. 2. Four examples of different QA level: Difficulty 1, 2, 3, 4, respectively. **Difficulty 1** and **Difficulty 2** target the length of shot video. **Difficulty 1** requires single supporting fact to answer and **Difficulty 2** requires multiple supporting facts to answer. **Difficulty 3** or **Difficulty 4** requires a time factor to answer and target the length of scene video. Especially, **Difficulty 4** requires causality between supporting facts from a different time.

- **Level 4 (High-level reasoning for causality):** The questions at this level cover reasoning for causality beginning with “Why”. Reasoning for causality is the process of identifying causality, which is the relationship between cause and effect from actions or situations.

Construct difficulties of QA’s Hierarchy with Two Criteria From the two criteria, we define four hierarchical difficulties for QA and these difficulties are consistent with cognitive developmental stages of Piaget’s theory [26,4]. Questions classified to level 1 in MC and LC belong to Difficulty 1 which is available from *Pre-Operational Stage* where a child thinks at a symbolic level, but is not yet using cognitive operations. Questions classified to level 1 in MC and level 2 in LC belong to Difficulty 2 which is also available from *Early Concrete Stage* where a child can utilize a relevant operation between multiple supporting facts. Questions classified to level 2 in MC and level 3 in LC belong to Difficulty 3 which is available from *Middle Concrete Stage* where a child can think by utilizing more than two relevant cognitive operations and utilize dependent multiple supporting facts across time. Questions classified to level 2 in MC and level 4 in LC belong to Difficulty 4 which is available from *Concrete Generalization Stage* where a child can just generalize only from personal and concrete experience and have a higher thought on causality in relation to “why”. Examples for each level are illustrated in Figure 2. We analyze overall distributions of levels per episode and 5W1H question types per difficulty level in the appendix.

Kyungsu : Yes. Yes, that's right. Something came up.
I(Kyungsu)'m sorry. I(Kyungsu)'m really sorry.

Deogi : Are you(Heayoung1) a human? Are
you(Heayoung1) even a human being? Still smiling
after you(Heayoung1) called off the wedding?

Haeyoung1 : It's a hundred times better to not marry instead of
marrying then splitting up later!
We(Heayoung1,Tejin) wouldn't have been able to
live together for a long time anyway!

Deogi : If you(Heayoung1) weren't going to marry
him(Tejin), then why did you(Tejin) bring him
home? How could you(Heayoung1) do this? It's
such an embarrassment!

Haeyoung1 : Seriously! Then should I(Heayoung1) torture
myself and live with him(Tejin)?

Kyungsu : Something came up. Yes. I(Kyungsu)'m sorry. Yes.
Yes. Yes, I(Kyungsu)'m sorry.

Deogi : You(Heayoung1) brat! How could you(Heayoung1)
talk back like this? What are you(Heayoung1)
doing?



(a)

(b)

Fig. 3. Examples of visual metadata and coreference resolved scripts. Our dataset provides visual metadata containing the main characters' bounding box, name, behavior, and emotion for each frame. Providing character-centered visual metadata helps understanding video story with scripts containing the main character's coreference.

3.3 Character-Centered Video Annotations

As the characters are primary components of stories, we provide rich annotations for the main characters in the drama “Another Miss Oh”. As visual metadata, all image frames in the video clips are annotated with main characters’ information. Also, to give a consistent view of the main characters, all coreference of the main characters is resolved in scripts of the video clips. Figure 3 shows the examples of visual metadata and coreference resolved scripts.

Visual Metadata

- **Bounding Box:** In each image frame, bounding boxes of both a face rectangle and a full-body rectangle for the main characters are annotated with their name. In total, 20 main characters are annotated with their unique name. We denoted 20 main characters in the appendix.
- **Behavior & Emotion:** Along with bounding boxes, behaviors and emotions of the characters shown in the image frames are annotated. Including `none` behavior, total 28 behavioral verbs, such as `drink`, `hold`, `cook`, is used for behavior expression. We denoted details of 28 behavioral verbs in the appendix. Also, we present characters’ emotion with 7 emotional adjectives; `anger`, `disgust`, `fear`, `happiness`, `sadness`, `surprise`, and `neutral`.

In Figure 4, distributions of main character and their behavior and emotion in visual metadata is visualized. As shown in Figure 4(a), Haeyoung1 and Dokyung appear the most frequently among all characters. For Figure 4(b) and (c), note that various behaviors and emotions are represented except the situations when cannot express much information due to its own trait like `none` behavior or neutral emotion.

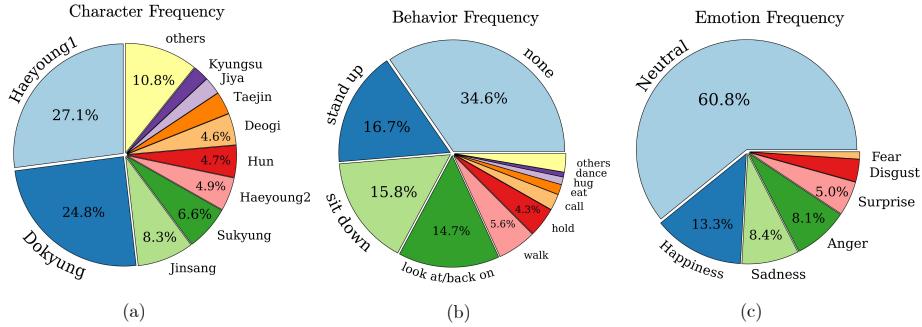


Fig. 4. (a) The percentage of each character’s frequency in visual metadata. *Haeyoung1* and *Dokyung* is two main characters of drama *AnotherMissOh*. *Haeyoung2* is the person who has same name with *Haeyoung1*, but we divided their name with numbers to get rid of confusion. (b) The percentage of each behavior frequency in the visual metadata. *none* behavior occupies a lot because there are many frames with only character’s face. (c) The percentage of each emotion frequency in the visual metadata.

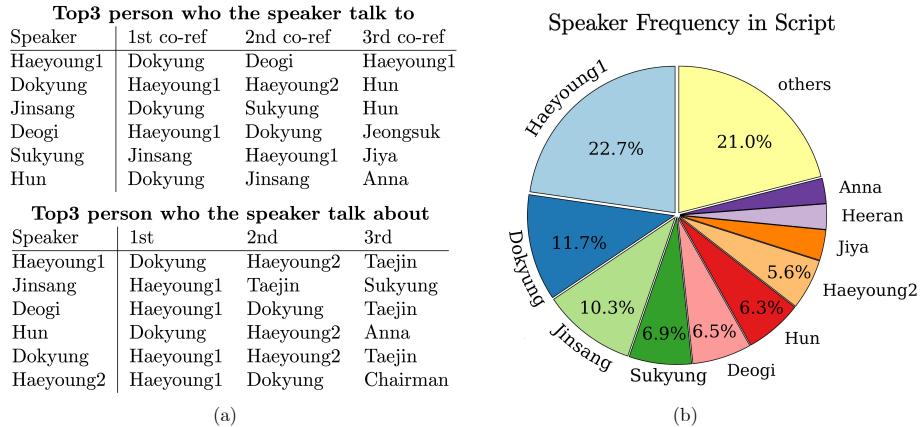


Fig. 5. (a) **Top:** Top-3 most frequently the person who the speaker talk to, for each top 6 most spoken person. **Bottom:** Top-3 most frequently the person who the speaker talks about, for each top 6 most spoken person. (b) The percentage of each person’s utterance in the script.

Coreference Resolved Scripts To understand video stories, especially drama, it is crucial to understand the dialogue between the characters. Notably, the information such as “Who is talking to whom about who did what?” is significant in order to understand whole stories. In DramaQA, we provide this information by resolving the coreferences for main characters in scripts. As shown in Figure 3, we annotate the characters’ names to all personal pronouns for characters, such as I, you, we, him, etc. By doing so, characters in scripts can be matched with those in visual metadata and QAs. We show analyses of each person’s utterances

in the scripts (Figure 5). First of all, we analyze who was the most frequent person talking with the main character and mentioned in their dialogue. As shown in Figure 5(a), we can see that two protagonists of the drama, *Dokyung* and *Haeyoung1*, appeared most often in their dialogue. Also, it indirectly shows the relationship between the main characters. *Hun*, *Dokyung*'s brother, is familiar to *Dokyung* but a stranger to *Haeyoung1*. Figure 5(b) shows the percentage of each character's utterance from whole episodes.

4 Model

We propose Dual Matching Multistream which grounds evidence in coherent characters to answer questions about the video. Our main goal is to build a QA model that understands the story, by utilizing the three modalities of the main character-centric annotations: speaker annotated scripts, visual metadata, and visual bounding boxes. We use a context matching module to get a QA-aware sequence for each stream, and a character matching module to focus on parts directly related to characters in QA. Outputs of these two modules are converted to a score for each answer candidate to select the most appropriate answer. Figure 6 shows our network architecture. In the following sections, we give detailed descriptions about feature extraction in Section 4.1, context matching module in Section 4.2, character matching module in Section 4.3, and answer selection in Section 4.4.

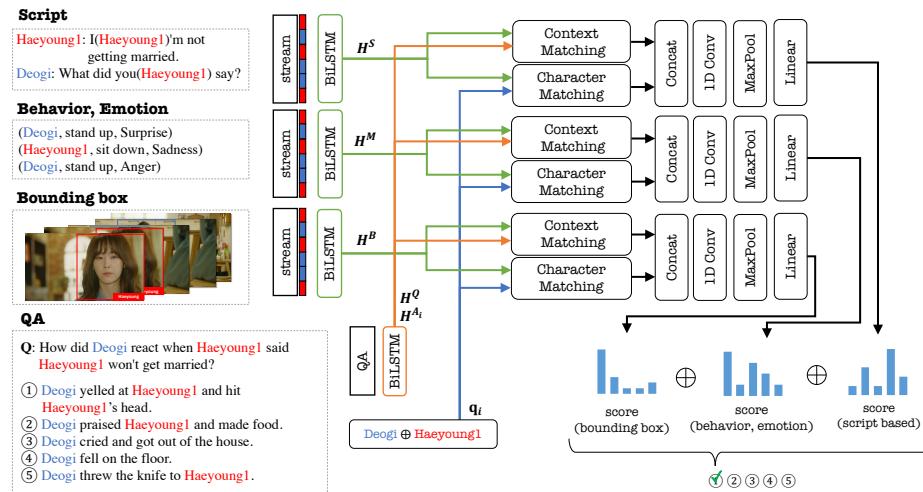


Fig. 6. Our Dual Matching Multistream model, which learns underlying correlations between the video clips, QAs and characters using both a context matching module and a character matching module. Final scores for answer selection is sum of each stream's output score.

4.1 Feature Extraction

An input into our model consists of a question, a set of five candidate answers, and three types of streams related to video context: speaker annotated scripts, visual metadata (behavior and emotion), and visual bounding boxes. An input stream from the script concatenates all words that appear. We denote the script $\mathbf{S} \in \mathbb{R}^{T_S \times D^W}$ where T_S is the number of words in the target video and D^W is the word embedding dimension of each word. An input stream from visual metadata concatenates all $\{\text{behavior}, \text{emotion}\}$ pairs that appear in the target video. We denote the visual metadata stream $\mathbf{M} \in \mathbb{R}^{2T_V \times D^W}$ where T_V is the number of total visual samples. The visual bounding box input stream concatenates all of the visual features of a person’s full body bounding box. We denote the visual bounding box feature stream as $\mathbf{B} \in \mathbb{R}^{T_V \times D^V}$ where D^V is the feature dimension of each bounding box. Each question and its five corresponding answer candidates are also preprocessed using the same method as the script stream. We denote a question as $\mathbf{Q} \in \mathbb{R}^{T_Q \times D^W}$ and i -th answer candidates as $\mathbf{A}_i \in \mathbb{R}^{T_{A_i} \times D^W}$, where T_Q and T_{A_i} is the length of each sentence.

In order to better understand the context, we also use character information from a speaker of script and a character which is annotated in visual metadata. Both pieces of character information are converted to one-hot vector and concatenated to input streams embedding dimension respectively. Then, we use bi-directional LSTM to get streams with temporal context from input streams. We concatenate the hidden states from both forward and backward directions at each timestep and we denote that as $H^S \in \mathbb{R}^{T_S \times 2h}$, $H^M \in \mathbb{R}^{T_M \times 2h}$, and $H^B \in \mathbb{R}^{T_B \times 2h}$ where h is hidden dimension of BiLSTM. Similarly, we can get hidden states of question H^Q and i -th answer candidate sentence H^{A_i} . In Figure 6, green lines are indicating H^S , H^M , and H^B respectively, and orange line is indicating H^Q and H^{A_i} as query.

4.2 Context Matching Module

The context matching module converts each input sequence to a query-aware context by using the question and answers as a query. This approach was taken from [31]. Context vectors are updated with a weighted sum of query sequences based on the dot product similarity between each query timestep and its corresponding context vector.

$$C^{S,Q} = (H^S(H^Q)^T)H^Q \in \mathbb{R}^{T_S \times 2h} \quad (1)$$

$$C^{S,A_i} = (H^S(H^{A_i})^T)H^{A_i} \in \mathbb{R}^{T_S \times 2h} \quad (2)$$

where H^Q is the query sequence from the given question and H^{A_i} is query sequence from the i -th candidate answer. We can get $C^{M,Q}$, C^{M,A_i} , $C^{B,Q}$, and C^{B,A_i} in the same manner.

4.3 Character Matching Module

Under the assumption that there is background knowledge that covers the entire video clip, such as the characteristics of each of the main characters, we have global representations for each character name $\mathbf{M}_C \in \mathbb{R}^{N_C \times d}$, where N_C is the number of main characters and d is a dimension of each character representation. In our case d is same with $2h$ for multi-head attention. We use a multi-hot vector $\mathbf{c}_i \in \mathbb{R}^{N_C}$ to encode whether a character's name appears in each QA pair $\{\mathbf{Q}, \mathbf{A}_i\}$. From these, we get character query $\mathbf{q}_i = \mathbf{c}_i^T \mathbf{M}_C$ from each QA pair, which is equivalent to the sum of representation of characters who appear in \mathbf{Q} or \mathbf{A}_i . In Figure 6, blue line is the query \mathbf{q}_i .

Using this \mathbf{q}_i as query, we used a variation of multi-head attention on input stream H^S , H^M , and H^B to find parts directly related to the character appearing in the question and answer. This idea is borrowed from the previous work [37,12]. This module takes input stream H^S , H^M , and H^B as key K and \mathbf{q}_i as query q . We project the query and each input stream timestep respectively to h hidden projections of d_k dimension, with h different parameter matrices. Then dot product attention is calculated between each input projection qW_i^q and query projection KW_i^K .

$$\mathbf{a}_i = \text{DotProd}(qW_i^q, KW_i^K) \in \mathbb{R}^T \quad (3)$$

$$\text{DotProd}(x, Y) = \text{softmax}(xY^T / \sqrt{d_k}) \quad (4)$$

where \mathbf{a}_i is attention score for each timestep of input stream at i -th projection, and W_i^q and W_i^K are i -th weight matrices.

After we get dot product attention, we expand attention scores and multiply to each projection vector.

$$\text{head}_i = (\mathbf{a}_i \times \mathbf{1}^T) \odot (KW_i^V) \quad (5)$$

where $\mathbf{1}$ is a vector of all ones with d_k dimension, \times is the outer product of two vectors, W_i^V is i -th projection matrix and \odot is element-wise multiplication.

After concatenating all heads in the second dimension, a linear layer is applied to ensure that shape matches the input stream.

$$\text{MultiHeadAttn}(H, \mathbf{q}_i) = [\text{head}_1; \dots; \text{head}_h]W_o \quad (6)$$

where W_o is a linear layer with $\mathbb{R}^{hd_k \times d}$.

As an output of multi-head attention has the same shape as the input stream, we add the input and the output with normalization.

$$H' = \text{Norm}(H + \text{MultiHeadAttn}(H, \mathbf{q}_i)) \quad (7)$$

where H' is C^{S, \mathbf{q}_i} , C^{M, \mathbf{q}_i} , or C^{B, \mathbf{q}_i} for H^S , H^M , or H^B respectively. This output is the context of the story directly related to the person in the QA pair.

4.4 Answer Selection

We concatenate H^S from Section 4.1, $C^{S,Q}$ and C^{S,A_i} from Section 4.2, and C^{S,\mathbf{q}_i} from Section 4.3. We also concatenate boolean flag f which is TRUE when the speaker or the person in visual metadata appears in the question and answer pair.

$$H^{S_{all},Q,A_i} = [H^S; C^{S,Q}; C^{S,A_i}; C^{S,\mathbf{q}_i}; f], \quad (8)$$

where we can get H^{M_{all},Q,A_i} and H^{B_{all},Q,A_i} with the same concatenation process.

For each concatenated stream H^{S_{all},Q,A_i} , we apply 1-D convolution filters with various kernel sizes and concatenate them to get final representation:

$$\mathbf{o}^S[i] = \text{maxpool}([\text{Conv}_{i1}; \text{Conv}_{i2}; \text{Conv}_{i3}; \text{Conv}_{i4}]) \quad (9)$$

$$\text{Conv}_{ij} = \text{ReLU}(\text{Conv}(H^{S_{all},q,a_i}, w_c^{\mathbf{s}_j})) \quad (10)$$

where $\mathbf{o}^S[i]$ is the final representation for i -th answer candidate from the script, and $w_c^{\mathbf{s}_j} \in \mathbb{R}^{j \times (4d+1) \times d/2}$ is a weight parameter of each convolution with kernel size j . Applying max-pool over time and linear layer after $\mathbf{o}^S[i]$, we calculate scalar score for each candidate answer. We can get \mathbf{o}^M and \mathbf{o}^B using the same approach. The final output score is simply the sum of output scores from the three different streams, and the model selects the answer candidate with the largest final output score as the correct answer.

5 Results

5.1 Settings

DramaQA dataset has 16,191 QA pairs which we separated into 10,098 for train set, 3,071 for validation set, and 3,022 for test set, each of which doesn't have any QA which targets video existing in other set. We use pretrained GloVe features [24] to initialize and make each word embedding trainable. We also initialize the main character words with random vectors to represent character level features. To initialize and fine-tune words of text input, we use 300 dimensions. For visual bounding box features, we used ResNet-18 [6] pretrained feature extraction. To get streams with temporal context from input streams, we use 150 hidden dimensions for BiLSTM, which means the output of Section 4.1 has 300 dimensions for each timestep. We use 4 heads and 75 dimensions for d_k at multi-head attention layer at Section 4.3. We limited our script input length to 300 words and our visual metadata samples to 100 samples for speed and memory usage, which covers over 97% of textual data and 82% of visual samples. The batch size is 16 and cross-entropy loss is used. For optimization, Adam [14] is used with 10^{-4} learning rate and 10^{-5} weight decay.

Table 2. Quantitative result about ablation study of our model on the DramaQA test set. We divided test set by difficulty level and get performance of each set. Last and second last columns respectively show the average of performance of each set and performance of overall test set, respectively. **Our- $\{S, M, B\}$** indicates our model without input script, behavior&emotion, and visual bounding box respectively. **Our-Char.Match** is our model without character match module and **Our-Con.Match** is our model without context matching module.

Model	Diff. 1	Diff. 2	Diff. 3	Diff. 4	Overall	Diff. Avg.
QA+DotProd	51.46	47.66	41.46	50.38	48.98	47.74
QA+MLP	30.29	27.73	28.18	27.43	28.95	28.41
Our-M	74.84	70.22	54.77	57.43	68.64	64.31
Our-B	74.48	69.20	56.03	57.93	68.44	64.41
Our-S	75.41	69.46	51.76	56.17	68.14	63.20
Our-Con.Match	60.77	58.81	32.41	33.75	52.89	46.43
Our-Char.Match	76.05	70.47	54.27	54.16	68.77	63.74
Our	75.69	71.10	56.03	55.92	69.24	64.69

5.2 Quantitative Results

Table 2 shows our quantitative results on our dataset. **QA+DotProd** model is designed to choose the highest score on the cosine similarity between the average of question’s word embeddings and the average of corresponding candidate answer’s word embeddings. **QA+MLP** model concatenates average of question’s word embeddings and the average of corresponding candidate answer’s word embeddings and gets scores with softmax function after simple MLP network with 600 dimensions for input, 50 dimensions for the hidden layer, and 1 dimension for output. **Our- $\{S, M, B\}$** indicates our model without input script, behavior&emotion, and visual bounding box respectively. Note that three input streams are helpful to infer a correct answer. After all, most of QAs in shot level inquire visual information and most of QAs in scene level need joint understanding of visual and textual information. **Our-Char.Match** is our model without character match module and **Our-Con.Match** is our model without context matching module. Note that **Con.Match** module is more dominant to get correct answer than **Char.Match** module. **Char.Match** module only covers QA when different characters appear for each answer candidate, so it shows lower accuracy.

5.3 Qualitative Results

In this section, we visualize our qualitative results from Dual Matching Multi-stream model. As shown in Figure 7, our model predicts an answer successfully by matching characters from candidate answers with their information from each input source. For example, given emotions & behaviors, the model can infer *Deogi* yelled at *Haeyoung1* as *Deogi* is angry at *Haeyoung1*. The given bounding boxes persist the model to notice *Deogi* hit *Hayoung1*’s head. Moreover, the

given scripts show the situation where *Deogi* is yelling at *Haeyoung1*. By collecting these clues, the model solves the QA task correctly. More examples including failures are provided in the appendix.

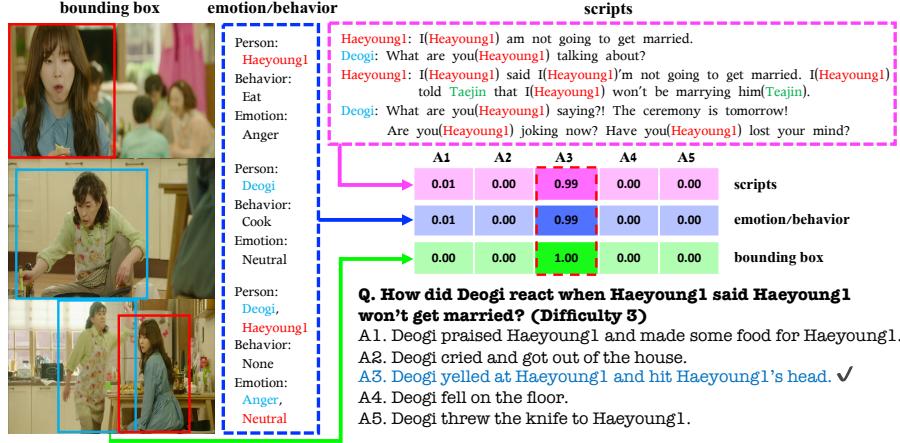


Fig. 7. Examples of our model’s correct prediction. Each answer score of inputs are shown in colored tables. The model predictions are indicated by checkmark, and ground truth answers are in blue.

6 Conclusion

To develop video story understanding intelligence, we propose DramaQA dataset with hierarchical evaluation metric and character-centered video annotations. Our dataset has cognitive-based difficulty levels for QA and hierarchical evaluation metric. Also, it provides coreference resolved script and rich visual metadata for character-centered video. Utilizing character-centered annotations, we suggest a Dual Matching Multistream model with a context matching module and a character matching module. Using both a context matching module and a character matching module, our model efficiently learns underlying correlations between the video clips, QAs and characters. For further work, we will provide hierarchical character-centered description objects, related objects, and places.

Acknowledgement

This work was partly supported by the Institute for Information & Communications Technology Promotion (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind) and Korea Institute for Advancement Technology (P0006720-GENKO) grant funded by the Korea government.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
3. Case, R.: Implications of neo-piagetian theory for improving the design of instruction. *Cognition, development, and instruction* pp. 161–186 (1980)
4. Collis, K.F.: *A Study of Concrete and Formal Operations in School Mathematics: A Piagetian Viewpoint*. Hawthorn Vic : Australian Council for Educational Research (1975)
5. Grosz, B.J., Weinstein, S., Joshi, A.K.: Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* **21**(2), 203–225 (1995)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 770–778 (2015)
7. Heo, Y., On, K., Choi, S., Lim, J., Kim, J., Ryu, J., Bae, B., Zhang, B.: Constructing hierarchical q&a datasets for video story understanding. CoRR **abs/1904.00623** (2019), <http://arxiv.org/abs/1904.00623>
8. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: CVPR (2017)
9. Jhuang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T.: A large video database for human motion recognition. In: Proc. of IEEE International Conference on Computer Vision. vol. 4, p. 6 (2011)
10. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
11. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
12. Kim, K., Choi, S., Kim, J., Zhang, B.: Multimodal dual attention memory for video story question answering. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*. Lecture Notes in Computer Science, vol. 11219, pp. 698–713. Springer (2018). https://doi.org/10.1007/978-3-030-01267-0_41, https://doi.org/10.1007/978-3-030-01267-0_41
13. Kim, K.M., Heo, M.O., Choi, S.H., Zhang, B.T.: Deepstory: Video story QA by deep embedded memory networks. IJCAI International Joint Conference on Artificial Intelligence pp. 2016–2022 (2017)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), <http://arxiv.org/abs/1412.6980>
15. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)

16. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018)
17. Lei, J., Yu, L., Berg, T.L., Bansal, M.: TVQA+: spatio-temporal grounding for video question answering. CoRR **abs/1904.11574** (2019), <http://arxiv.org/abs/1904.11574>
18. Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., Pal, C.: A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
19. McLaughlin, G.H.: Psycho-logic: A possible alternative to piaget's formulation. *British Journal of Educational Psychology* **33**(1), 61–67 (1963)
20. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2630–2640 (2019)
21. Mueller, E.T.: Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research* **5**(4), 307–340 (2004)
22. Mun, J., Seo, P.H., Jung, I., Han, B.: Marioqa: Answering questions by watching gameplay videos. In: ICCV (2017)
23. Pascual-Leone, J.: Cognitive development and cognitive style: A general psychological integration (1969)
24. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). <https://doi.org/10.3115/v1/D14-1162>, <https://www.aclweb.org/anthology/D14-1162>
25. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 724–732 (2016)
26. Piaget, J.: Intellectual evolution from adolescence to adulthood. *Human development* **15**(1), 1–12 (1972)
27. Riedl, M.O.: Computational narrative intelligence: A human-centered goal for artificial intelligence. arXiv preprint arXiv:1602.06484 (2016)
28. Riedl, M.O., Young, R.M.: Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* **39**, 217–268 (2010)
29. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. *International Journal of Computer Vision* **123**(1), 94–120 (2017)
30. Schank, R.C., Abelson, R.P.: Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Psychology Press (2013)
31. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. ArXiv **abs/1611.01603** (2016)
32. Seol, S., Sharp, A., Kim, P.: Stanford mobile inquiry-based learning environment (smile): using mobile phones to promote student inquiries in the elementary classroom. In: Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS). p. 1. The Steering Committee of The World Congress in Computer Science, Computer ... (2011)
33. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016)

34. Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* **2** (2012)
35. Szilas, N.: Interactive drama on computer: beyond linear narrative. In: *Proceedings of the AAAI fall symposium on narrative intelligence*. pp. 150–156 (1999)
36. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: Understanding Stories in Movies through Question-Answering. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
38. Winston, P.H.: The strong story hypothesis and the directed perception hypothesis. In: *2011 AAAI Fall Symposium Series* (2011)
39. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5288–5296 (2016)

APPENDIX

This appendix provides additional information not described in the main text. It contains analyses about the dataset in Section A.1, dataset collection methods in Section A.2 and qualitative results of our model in Section B.

A Dataset

A.1 Dataset Analysis

Hierarchical QAs DramaQA has four hierarchical difficulty levels for each questions, which are used as an evaluation metric. Figure 8(a) shows overall distributions of levels along with the episodes. Because a single *scene* of a video has multiple *shots*, the number of questions for Difficulty 1 and 2 is naturally larger than that for Difficulty 3 and 4. We also visualize 5W1H question types per difficulty level in Figure 8(b). In Difficulty 1 and 2, **Who** and **What** questions are the majority. In case of Difficulty 3, **How** and **What** types are the top-2 questions. In Difficulty 4, Most of questions are start from **Why**.

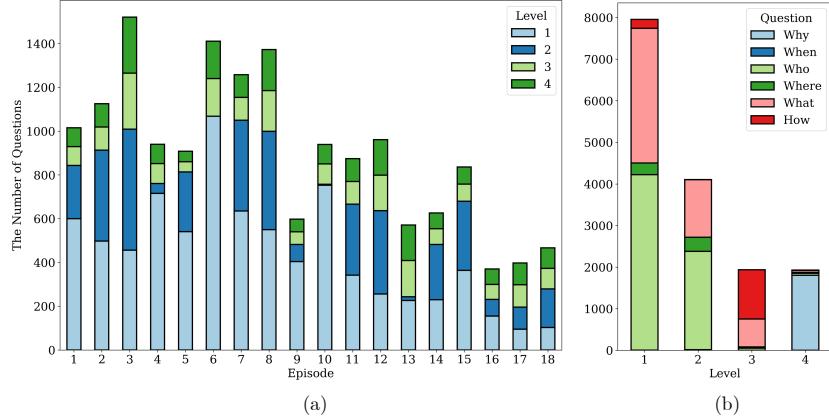


Fig. 8. (a) The number of QA pairs per episode and difficulty level. Given that the scene length is tens of times than the size of the shot, the variation between levels is small compared to the number of videos. (b) The number of 5W1H question types per difficulty level.

A.2 Data Collection

For the data collection process, we hired fluent English speakers without using a crowdsourcing service because of a copyright issue on the distribution of the drama video as well as using the characteristics of the drama annotation-task.

Since annotating visual information requires not only knowledge of the characters but also the stories going on the drama, utilizing a crowdsourcing service with a considerable number of part-time workers might decrease the quality of the resulting annotation dataset. Therefore, with automated visual annotation tagging tools, all annotation tasks were carried out by a small group of dedicated workers who are aware of the whole drama story-line and the individual characters involved in the story.

For visual metadata annotation, the visual bounding boxes were created using an automated tagging tool, and workers manually annotated the main characters' names, behaviors, and emotions. We predefined main characters, emotions, and behaviors as follows:

- Main character: Anna, Chairman, Deogi, Dokyung, Gitae, Haeyoung1, Haeyoung2, Heeran, Hun, Jeongsuk, Jinsang, Jiya, Kyungsu, Sangseok, Seohee, Soontack, Sukyung, Sungjin, Taejin, Yijoon
- Emotion: anger, disgust, fear, happiness, sadness, surprise, neutral
- Behavior: drink, hold, point out, put arms around each other's shoulder, clean, cook, cut, dance, destroy, eat, look for, high-five, hug, kiss, look at/back on, nod, open, call, play instruments, push away, shake hands, sing, sit down, smoke, stand up, walk, watch, wave hands, write

The QA pairs were created with the following rules: 1) Workers must use the main characters' names (i.e. *Haeyoung1*, *Dokyung*, *Deogi*, etc.) instead of pronouns (i.e. *They*, *He*, *She*, etc.). 2) Questions and answers should be complete sentences. 3) All sentences should be case-sensitive. For hierarchical difficulty of QAs, different rules were applied for each level:

- Difficulty 1
 - The Question-Answering (QA) set should be based on only one supporting fact (*A triplet form of subject-relationship-object*) from the video.
 - Question can (should) be started with *Who*, *Where*, and *What*.
- Difficulty 2
 - The Question-Answering (QA) set should be based on multiple supporting facts from the video.
 - Question can (should) be started with *Who*, *Where*, and *What*.
- Difficulty 3
 - The Question-Answering (QA) set should be based on multiple situations/actions with sequential information. To answer the question at this difficulty, temporal connection of multiple supporting facts should be considered, differently from Difficulty 2 questions.
 - Question can (should) be started with *How* (recommended) and *What*.
- Difficulty 4
 - The Question-Answering (QA) set should be based on reasoning for causality.
 - Question can (should) be started with *Why*.

B Qualitative Results

Figure 9 and 10 are the examples of our model’s prediction. Each answer scores from scripts, emotion/behavior, and bounding boxes are shown in colored tables. As these answer scores are obtained using softmax, the sum of these scores is 1. A checkmark indicates the model predictions, and ground truth answer is in blue.

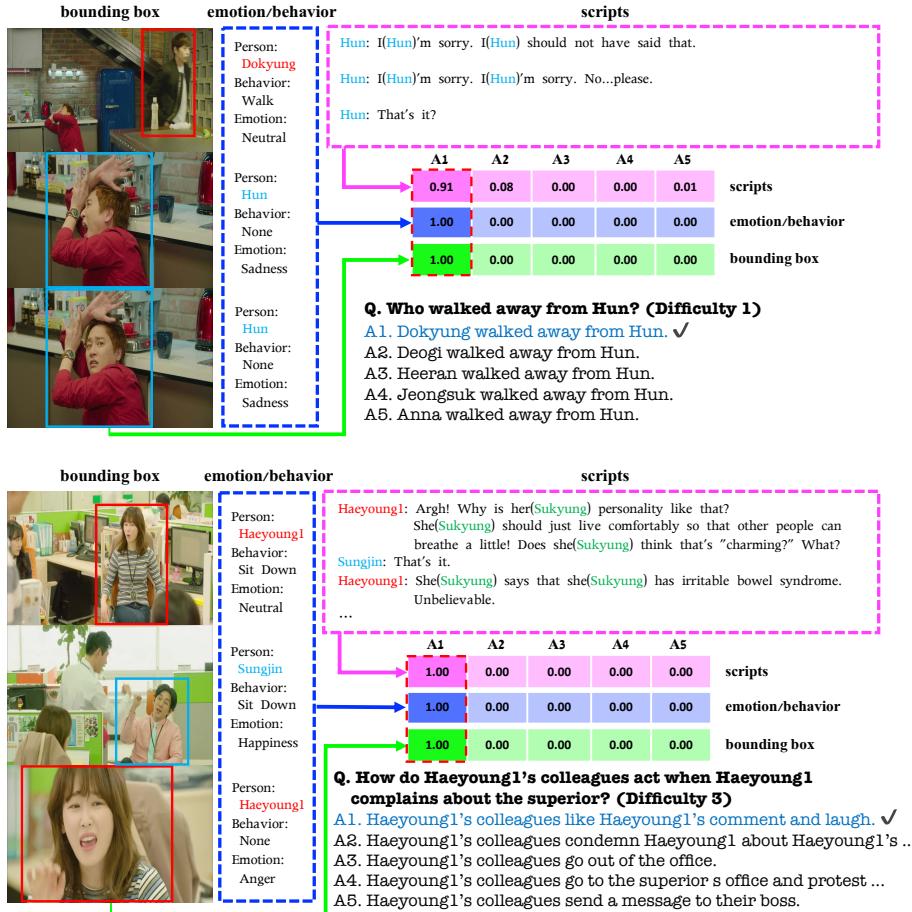


Fig. 9. Examples of our model’s correct prediction. In the case of the example above, we can find evidence for the correct answer from visual sources, so scores from bounding box and emotion&behavior are high. Although it doesn’t have any clue from script, it has high score from script using the relationship between *Dokyung* and *Hun*. In the case of the example below, we can find evidence for the correct answer from all input sources, so it correctly choose the answer sentence.



Fig. 10. Examples of our model’s prediction. In the case of the above example, our model predicts the first candidate answer, which is incorrect. Since the correct answer cannot be obtained from any input sources, the wrong answer is chosen. In the case of the example below, we can find evidence for the correct answer from all input sources.