# Constructing Hierarchical Q&A Datasets for Video Story Understanding

## Abstract

Video understanding is emerging as a new paradigm for studying human-like AI. Question-and-answering (Q&A) is used as a general benchmark to measure the level of intelligence for video understanding. While several previous studies have suggested datasets for video Q&A tasks, they did not really incorporate story-level understanding, resulting in highly-biased and lack of variance in degree of question difficulty. In this paper, we propose a hierarchical method for building Q&A datasets, i.e. hierarchical difficulty levels. We introduce three criteria for video story understanding, i.e. memory capacity, logical complexity, and DIKW (Data-Information-Knowledge-Wisdom) pyramid. We discuss how three-dimensional map constructed from these criteria can be used as a metric for evaluating the levels of intelligence relating to video story understanding.

## Introduction

In narratology, story is often differentiated from discourse, where story refers to content (i.e., what to tell) and discourse denotes expression or representation (i.e., how to tell it)(Chatman 1978; Genette 1980). While the representation can vary from text and oral storytelling to films, dramas, and virtual environments including virtual reality (VR), understanding of the given story shares some common key aspects regardless of the represented media.

According to computational linguists, narrative theorists, and cognitive scientists, narrative understanding is somehow linked with the measurement of reader's intelligence. For example, readers can understand story as a way of problem solving in which they keep focusing on how main characters overcome coming obstacles throughout story (Black and Bower 1980). Thus readers, while reading, make inferences both in prospect and in retrospect about what events will occur and how these events could occur, considering the causal relationships between different events in the story (Trabasso and Van Den Broek 1985; McKoon and Ratcliff 1992; Graesser, Singer, and Trabasso 1994). Inferring causal relationships between events, is a key element for the reader to reconstruct a given narrative as a mental model in the reader's mind (Zwaan 1999; Zwaan, Langston, and Graesser 1995). Humans have the natural capability of "organizing our experience into narrative form" as narrative intelligence (Blair and Meyer 1997; Mateas and Sengers 1999).

Recently, video story serves as a testbed of real-world data to construct human-level AI from two points of view. First, video data has various modalities such as sequence of images, audios (including dialogue, sound effects, background music), and often texts (subtitles or added comments). Second, video shows a cross-section of everyday life. Understanding video story involves analyzing and simulating human vision, language, thinking, and behavior, which is a significant challenge to current machine learning technology.

To measure human-level machine intelligence, we apply video Question-and-Answering (video Q&A) task as a proxy of video story understanding. The task can be regarded as a Turing Test for video story understanding(Turing 1950). While several previous studies have suggested various datasets for the video Q&A task(Tapaswi et al. 2015; Kim et al. 2017; Mun et al. 2017; Jang et al. 2017; Lei et al. 2018), they are built without careful consideration of "understanding of video story". For such reason, the previously released video Q&A datasets are highly-biased and lack of variance in question difficulty. The construction of Q&A dataset with hierarchical difficulty levels in terms of story understanding is crucial, as people with different perspectives (or different intelligence levels) will understand the given video story differently.

In this paper we propose three criteria such as memory capacity, logical complexity, and DIKW (Data-Information-Knowledge-Wisdom) hierarchy for video story understanding and construct a three-dimensional hierarchical map of video story understanding using the criteria. The constructed three-dimensional map can leverage the understanding of developmental stages of human intelligence. We expect that the proposed hierarchical criteria can be utilized later as a metric that can help evaluate the levels of intelligence relating to video story understanding. Our main contributions are two fold. First, we suggest three criteria for constructing hierarchical video Q&A datasets. The criteria can be used to analyze the quality of video Q&A dataset in terms of bias and variance for dataset difficulty. Second, we interlink proposed three criteria to neo-piagetian's theory, which can help interpreting our story-enabled intelligence to cognitive development stage of human.

## Related Works

While video understanding is still in its early stage, researchers proposed video Question-and-Answering (Q&A) dataset as a general benchmark to measure video understanding intelligence. The most notable datasets proposed so far are MovieQA(Tapaswi et al. 2015), PororoQA(Kim et al. 2017), MarioQA(Mun et al. 2017), TGIF-QA(Jang et al. 2017), and TVQA(Lei et al. 2018). Here, we review above video Q&A datasets and present our contributions.

MovieQA aims to evaluate story understanding of video and text in movie. For the MovieQA, question and answer pairs are collected by annotators who read plot synopses of movies instead of the entire movies. PororoQA is comprised of targeted animation videos, which makes its content easier to understand than MovieQA dataset. MarioQA dataset is also based on synthetic videos constructed automatically from the popular Mario game playing videos. The dataset focuses on understanding of temporal relationship between multiple events. When the dataset was generated, template-based question and answer generation methods were used from extracted events. TGIF-QA dataset focuses on only visual information in the GIF-format images. TGIF-QA limited the question type to three types: repetition count, repeating action, and state transition. Those types of questions are required spatio-temporal reasoning from videos. TVQA question is a large-scale video QA dataset based on 6 popular TV shows about sitcoms, medical and crime TV programs. For the TVQA, all questions and answers are attached to 60-90 seconds video clips. It requires comprehension for subtitles-based dialogue and recognition of relevant visual concepts to answer the questions in the dataset.

Our work contributes to this line of research, but instead of introducing new datasets, here we propose new criteria for constructing video Q&A dataset on careful consideration of video story understanding.

## Three-dimensional Video Q&A Hierarchy

This section describes three criteria as measures of video story understanding. The three criteria are as follows: memory capacity, logical complexity, and DIKW hierarchy. These three criteria are combined to construct a three-dimensional video understanding map. Every question in the Q&A dataset is classified in each level by each criterion respectively, and then represented as a point on the three-dimensional map which has three dimensions corresponding to the three criteria. Finally, every point in the map is assigned to the cognitive development stage according to Piaget's theory(Piaget 1972; Collis 1975b). We explain this process in detail in the following subsections.

### Criterion 1: Memory Capacity

When determining the difficulty of questions collected for the video, the length of the video is crucial for reasoning and finding the correct answer in machine learning perspective. If the length of the video required for answering the question is longer, the question can be classified as more difficult, and vice versa. For example, a question targeted to short video is a lot more difficult than a question targeted to

an image frame, and a question targeted to entire video is a lot more difficult than a question targeted to one segment video. This criterion also can be interpreted as memory capacity of humans. We in this paper define *Memory Capacity* as the length of the target video which has to be considered to answer given question. The classification results are as follows.

- Level 1 (frame): The questions for this level are based on a video's frame. This level has the same difficulty as that of a kind of Visual Question Answering (VQA) dataset(Malinowski, Rohrbach, and Fritz 2015; Ren, Kiros, and Zemel 2015; Agrawal et al. 2017; Zhu et al. 2016; Johnson et al. 2017; Wang et al. 2018).

- Level 2 (shot): The questions for this level are based on a video length less than about 10 seconds without change of viewpoint. This set of questions can contain atomic action or functional/meaningful action in the video. Most recent datasets which deal with video belong to this level(Jang et al. 2017; Maharaj et al. 2017; Mun et al. 2017). At this level, both atomic action and meaningful action can appear and their boundary is vague. For example, shaking hands (atomic action) and a gesture to go away (meaningful action) have a similar action, However, their meaning is different depending on the situation, not depending on video length.

- Level 3 (scene): The question set for this level is based on 1-3 minutes long clips without place change. Videos at this level contain sequences of actions, which augments the level of difficulty from level 2. We consider this level as the "story" level according to our working definition of story. MovieQA(Tapaswi et al. 2015) and TVQA(Lei et al. 2018) are the only datasets which belong to this level. For example, the popular TV sitcom *Friends* has 13 scenes per episode on average, and a movie has 120 scenes on average.

- Level 4 (sequence): The set of questions at this level relate to more than two scenes. Sequence has video length less than entire movie. To the best of our knowledge, there are no datasets dealing with the video at this level.

- Level 5 (entire): The question set for this level is based on an entire story from beginning to end. Questions at this level attached to whole video like an entire movie or an episode of a drama.

### Criteria 2: Logical Complexity

Complicated questions often require more (or higher) logical reasoning steps than simple questions. In other words, if a question requires multiple supporting facts which have interrelations to answer, we regard that the question has high logical complexity. For story-enabled intelligence, it is required to trace several logical reasoning steps by combining multiple supporting facts to give a correct answer to a given question. In a similar vein, if a question needs only a single supporting fact with a single relevant data, we regard that it has low logical complexity. It may need only one reasoning step or one perception step to answer the question.

This subsection describes the second criterion *logical complexity* to define the level of difficulties for questions. We define five logical complexity levels based on the Stanford Mobile Inquiry-based Learning Environment(SMILE)(Seol, Sharp, and Kim 2011). In the SMILE project, students learn online lectures or documents via a mobile platform and generate relevant questions based on what they have learned. Each question made by a student is classified into five logical complexity levels as follows:

- Level 1 (Simple recall on one cue) : The question set at this level can be responded with minimum cognitive effort, involving simple recall or simple arithmetic calculations. The questions at this level requires only one supporting fact such as {*boy-hold-cup*}. As the questions at this level are too simple, they may not trigger much interaction.

- Level 2 (Simple analysis on multiple cues) : The question set at this level can be responded with simple analysis of the question types or problems with simple reasoning. The questions at this level asks for factual information involving recall of independent multiple supporting facts, which trigger simple inference or quick interpretation. This question set begins with simple question types starting with "Who", "What", "When", "Where", "How many", and so on. Responses come from a range of clearly defined scope with little room for dispute.

- Level 3 (Intermediate cognition on dependent multiple cues) : The question set can be responded with intermediate level of cognition and analysis. The questions at this level require multiple cues with dependent data across time, and cover "how-to" questions which deals with sequential information. It requires comparison, classification, or categorization in responding to given questions at this level.

- Level 4 (High-level reasoning on causality) : The question set at this level can be responded with higher-level of analysis and reasoning rather than a lower-level thinking question. The questions at this level require more than simple recall or simple arithmetic tasks to answer. Even simple analysis or reasoning may not be enough to get an answer. The question set covers causality beginning with "Why". It requires own interpretation or synthesis in responding to given questions at this level.

- Level 5 (Creative thinking) : The question set at this level can be responded by requiring imagination and creation of new theory or hypothesis with supporting rationale. The question at this level covers creative thinking and reasoning that may help defining a new solution or concept that has not existed previously. For example, the questions (#19 and #20) in Table 1 draw an unique solution by formulating own rational equations about not occured situations.

## Criterion 3: DIKW Hierarchy

The DIKW (Data, Information, Knowledge, and Wisdom) hierarchy is widely accepted as a way of representing different levels of what we see and what we know (Schumaker 2011). In terms of video Question-and-answering, a level of understanding can also be identified by answering questions based on different levels ranging from data, information, knowledge, and wisdom. In this section, details of the four levels of video understanding are discussed.

- Level 1 (Data level): Data are the observations of the physical world (Schumaker 2011; Carlisle 2006) and are symbolic representation of things, events and activities(Ackoff 1989; Rowley 2007). In terms of video Q&A, the data level covers the questions for characters, characters' lines, objects, sounds, locations and simple behaviors which has no subjective meaning or goal for the environment (such as standing, walking, calling, and etc.).

- Level 2 (Information level) : Information refers to the data that have been shaped into a meaningful and useful form (Rowley 2007). Specifically, it includes the addition of relationships between data (Barlas, Ginart, and Dorrity 2005). In terms of video Q&A, information-level questions focus on the interaction between characters and objects such as actions, emotions, and changing circumstances which can be obtained from the scene of the video.

- Level 3 (Knowledge level) : Knowledge refers to the aggregation of related information that provides a clearer understanding of information (Barlas, Ginart, and Dorrity 2005; Schumaker 2011). Knowledge also involves the synthesis of multiple sources of information over time (Rowley 2007; Despres and Chauvel 2012). In terms of video Q&A, knowledge-level questions can be answered only with accumulated information of multiple scenes of the video including knowledge from fictional universe of contents.

- Level 4 (Wisdom level): Wisdom refers to accumulated knowledge, with which it is possible to apply understood concepts from one domain to new situations or problems (Rowley 2007). In terms of video Q&A, the wisdom-level questions can be answered by utilizing useful meta relationship of knowledge including nonsense and humor. For example, the question #28 in Table 1 needs to understand the character "Chandler" in terms of a sense of humor.

## Interpretation as Cognitive Development Stage

In the following section, we interpret proposed three criteria *(i.e., memory capacity, logical complexity, and DIKW pyramid)* from the viewpoint of cognitive development of human intelligence. The detailed cognitive development defined by Piaget is introduced and then we apply the cognitive development stage to criteria of three-dimensional video Q&A hierarchy.

**Piaget's theory of cognitive development** In this section, we explain cognitive development of human based on one of Neo-Piagetian theory(Collis 1975b) recasting of Piaget's theory of developmental stages(Piaget 1972). Piaget's theory explains in detail the process by which human cognitive ability develops, in conjunction with information processing models. In order to justify three criteria proposed in this paper in terms of human intelligence development, we examine the details of the developmental stages of Piagets theory.

Table 1: Examples of Question-and-Answering datasets for each criterion based on TV sitcom Friends

| | |
|---|---|
| **Criterion 1: Memory capacity** | |
| Level 1: Frame | (1) Q: What does Ross have on his shoulder?<br>A: A monkey is on his shoulder.<br>(2) Q: What is the hat made out of?<br>A:The hat is made out of a cup. |
| Level 2: Shot | (3) Q: What did Chandler swallow?<br>A: The button off his shirt.<br>(4) Q: How many times did the Ross knock the door?<br>A: Five times. |
| Level 3: Scene | (5) Q: What did Rachel do wrong when she did her first load of laundry?<br>A: Rachel left a red sock in her load of white clothes which made them pink.<br>(6) Q: Why does Rachel's boyfriend visit Phoebe at work?<br>A: To get a massage. |
| Level 4: Sequence | (7) Q: Why does Rachel not want to drink on her date?<br>A: Because she doesn't want to end up in his bed on the first date.<br>(8) Q: Why is Ross having the locks replaced?<br>A: He lost all of his keys. |
| Level 5: Entire | (9) Q: What is the main topic of this movie?<br>A: The story about true love.<br>(10) Q: How does the main character's personality change from the beginning to end?<br>A: At first he was fainthearted, but later he became brave. |
| **Criterion 2: Logical Complexity** | |
| Level 1: Simple recall | (11) Q: Who is drinking water?<br>A: Monica is drinking water.<br>(12) Q: What is it Ross is holding?<br>A: Ross is holding a phone. |
| Level 2: Simple analysis | (13) Q: What does Joey say to Chandler about the woman?<br>A: Joey says "woman love babies and guys who love babies".<br>(14) Q: What did Joey and Chandler do?<br>A: drinking and dancing. |
| Level 3: Intermediate cognition | (15) Q: How did Rachel regain the cart from the woman?<br>A: Rachel get in a cart.<br>(16) Q: How did Rachel figure out the truth of prom in high school?<br>A: Rachel saw the video which shoot at the day. |
| Level 4: High-level reasoning | (17) Q: Why does Rachel storm out of the office?<br>A: She thinks the interviewer is trying to sleep with her.<br>(18) Q: Why is Monica visiting Chandler at work?<br>A: They are going to look at houses together. |
| Level 5: Creative thinking | (19) Q: If Ross did not break up with Rachel, what could have been the consequence?<br>A: They will love each other forever.<br>(20) Q: If Rachel's parents didn't divorce, what could be different for her birthday party?<br>A: Rachel's friends will organize the only one birthday party, not two separate parties. |
| **Criterion 3: DIKW hierarchy** | |
| Level 1: Data-level | (21) Q: What does Ross do?<br>A: Ross walks to the customer.<br>(22) Q: Where does Phoebe grab the tissue?<br>A: Phoebe grabs from the kitchen. |
| Level 2: Information-level | (23) Q: What is Rachel's emotion to Monica's situation?<br>A: Rachel is Horrified.<br>(24) Q: What was Rachel doing before she comes into the room?<br>A: She was dancing. |
| Level 3: Knowledge-level | (25) Q: How do Joey and Chandler decide which baby to choose?<br>A: Asking the transit authority employee.<br>(26) Q: Why are Joey and Chandler at the Transit Authority<br>A: To apply for a job. |
| Level 4: Wisdom-level | (27) Q: Why now Rachel is telling Joey that her gynecologist tried to kill her?<br>A: To see if Joey could say a few words to her gynecologist too.<br>(28) Q: If Chandler suffers from jet lag now, what he could have done?<br>A: He calls 911. |

| | | Pre-Operational Stage (Stage 1) | Early Concrete Stage(Stage 2) | Middle Concrete Stage(Stage 3) | Concrete Generalization Stage(Stage 4) | Formal Stage(Stage 5) |
|---|---|---|---|---|---|---|
| **Memory Capacity** | Level1: Frame | ▬ | ▬ | ▬ | ▬ | ▬ |
| | Level2: Shot | ▬ | ▬ | ▬ | ▬ | ▬ |
| | Level3: Scene | | ▬ | ▬ | ▬ | ▬ |
| | Level4: Sequence | | | ▬ | ▬ | ▬ |
| | Level5: Entire | | | | ▬ | ▬ |
| **Logical Complexity** | Level1: Simple recall on one cue | ▬ | ▬ | ▬ | ▬ | ▬ |
| | Level2: Simple analysis on multiple cues | ▬ | ▬ | ▬ | ▬ | ▬ |
| | Level3: Intermediate cognition on dependent multiple cues | | | ▬ | ▬ | ▬ |
| | Level4: High-level reasoning on causality | | | | ▬ | ▬ |
| | Level5: Creative thinking | | | | | ▬ |
| **DIKW Hierarchy** | Level1: Data | ▬ | ▬ | ▬ | ▬ | ▬ |
| | Level2: Information | | | ▬ | ▬ | ▬ |
| | Level3: Knowledge | | | | ▬ | ▬ |
| | Level4: Wisdom | | | | | ▬ |

Figure 1: Interpretation of the proposed three criteria*(i.e., Memory capacity, Logical complexity, and DIKW Hierarchy)* as cognitive development stage proposed by Piaget and recasted by Collis. The highlighted bar means the possibility to apply cognitive operations for answering given question on each level of a criterion from each cognitive developmental stage.

Piaget's original model suggests a sensory-motor stage that occurs from birth, however, that stage involves only representations related to sensory-motor activity. Thus, we focus on the later stages that follow the pre-operational stage in which a child shows understanding behavior(Collis 1975a).

- Stage 1 (Pre-Operational Stage; 4 to 6 years) : At this stage, a child is thinking at a symbolic level, but is not yet using cognitive operations. The child can not transform, combine or separate ideas. Thinking at this stage is not logical and often unreasonable. Associations are made on the basis of emotion, preference at this stage, and it has a very egocentric sight of one's own world.

- Stage 2 (Early Concrete Stage; 7 to 9 years) : At this stage, a child can utilize only one relevant operation. Thinking at this stage has become detached from instant impressions and is structured around a single mental operation, which is a first step towards logical thinking.

- Stage 3 (Middle Concrete Stage; 10 to 12 years) : At this stage, a child can think by utilizing more than two relevant cognitive operations and acquire the facts of dialogues. This is regarded as the foundation of proper logical functioning. However, a child at this stage lacks own ability to identify general fact that integrates relevant facts into coherent one. Moreover, thinking at this stage are still concrete, not abstract.

- Stage 4 (Concrete Generalization Stage; 13 to 15 years) : Piaget referred to this stage as the early formal stage, particularly for abstract thinking. A Child at this stage, however, can just generalize only from personal and concrete experiences. The child do not have own ability to

hypothesize possible concepts or knowledge that is quite abstract.

- Stage 5 (Formal Stage; 16 years onward) : This stage is characterized purely by abstract thought. Rules can be integrated to obtain novel results that are beyond the individuals own personal experiences. However, this is not a stage that every person can reach.

**Applying Piaget's human developmental stage to criteria of three-dimensional hierarchy** Human understanding, as Piaget stated, can be classified into different stages. We propose that the concepts from Piaget's theory of development correspond to the three-dimensional video Q&A Hierarchy criteria.

First, the development stage can be explained from the perspective of the memory capacity criterion. Memory capacity corresponds to working memory of the cognitive process model. (Case 1980a) suggested that the working memory available for problems increases with age, as does the space required for higher level responses. This relationship between working memory and age leads to the proposition that cognitive developmental stages can be explained by increasing attention span, or working memory capacity(Case 1980b; Mclaughlin 1963; Pascual-Leone 1969). Thus, we assume that Piaget's theory of development of human intelligence with age can be in accordance with the memory capacity criterion. For example, understanding a static image can be understood to be from Stage 1 (Pre-operational). Also, understanding video within 10 seconds is possible from Stage 1. However, beyond minutes*(e.g., understanding a scene within 3 minutes)*, is possible from Stage 2 (Early
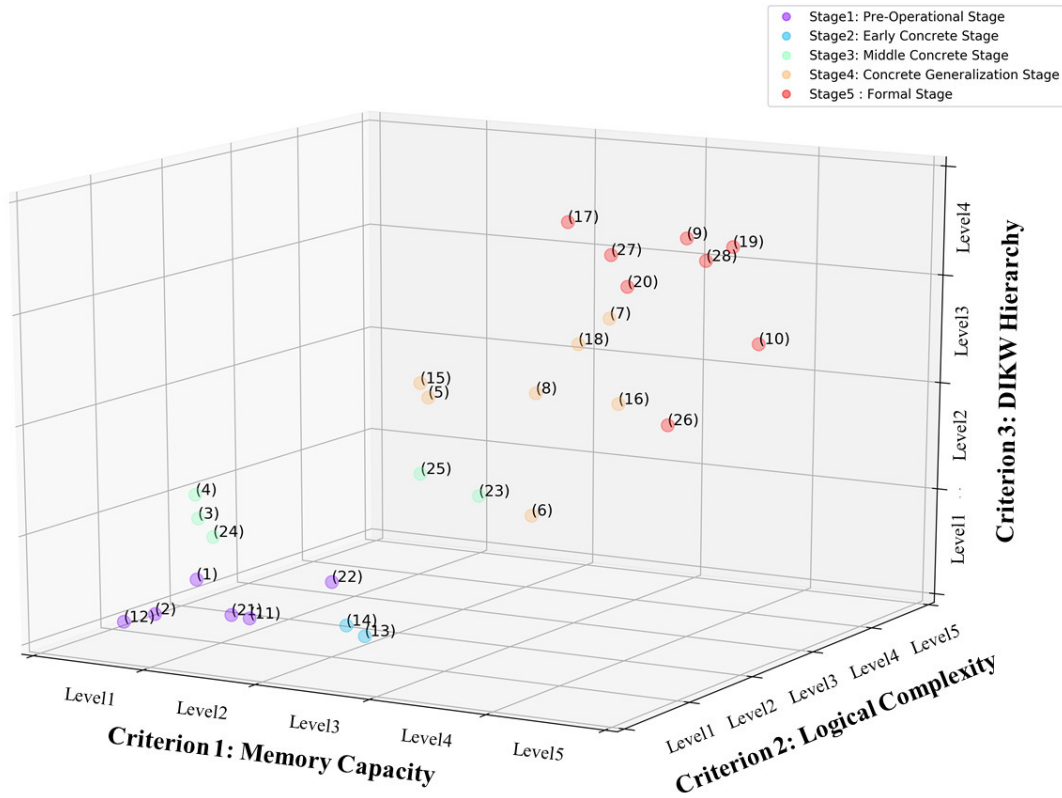
Figure 2: Example of three-dimensional video Question-and-Answering (video Q&A) hierarchy. Each point represents each question in Table 1. Three coordinates of each point is assigned by the definition of level of three criteria, and matched to one developmental stage which is the highest stage among derived three stages following the interpretation of Figure 1.

concrete). Beyond this, understanding two or more scenes *(e.g., understanding sequences changing time and place)* is possible from Stage 3 (Middle concrete). Finally, it is possible from Stage 4 (Concrete generalization) to know and understand whole video entirely. While this mapping is not exactly distinct, there exists a clear hierarchy as shown in Figure 1.

Piaget's developmental stages are also consistent with the logical complexity criterion. As the SMILE project proposes, from simple recall to assumption-based reasoning, methods have a kind of hierarchy that is closely related to a person's stage of development. For example, level 1 and level 2 is available from Stage 1 (Pre-operational) in that it needs a simple call. Specifically, level 1 requires one supporting fact *(e.g., {jacket-is-black})*, on the other hand, level 2 requires independent multiple supporting facts. Level 3 is available from Stage 3 (Middle concrete), in that this level can be understood using dependent multiple supporting facts across time. Level 4 is available from Stage 4 (Concrete generalization), because this level requires a higher thought on causality in relation to "Why". Finally, level 5 is available from Stage 5 (Formal Stage), as it requires creativity and abstract thinking about new ideas. As such, each phase of SMILE can be expressed as roughly equivalent to the hu-

man developmental stage postulated by Piaget(Collis 1972).

Piaget's human developmental stages are not exactly consistent with the DIKW hierarchy criterion. However, data level is possible from the Stage 1 (Pre-operational) in terms of providing simple factual data. Information level is possible from Stage 3 (Middle concrete), because it identify a relationship between some real world entities. Knowledge level is roughly equivalent from Stage 4 (Concrete generalization), as both are incapable of inferring information from abstract variables. Finally, wisdom level in DIKW hierarchy criterion is possible from Stage 5 (Formal stage), in that it can be inferred and applied to new situations like knowledge transferring(Collis 1975a). Figure 2 shows three-dimensional hierarchical map for each question represented in Table 1. Three coordinates of each point are assigned by the definition of level of three criteria, and matched to one developmental stage which is the highest stage among derived three stages following the interpretation of Figure 1. For example, a question "What did Joey and Chandler do at Ross house?" is set {3, 2, 1} level for three criteria. Each level can be developed from {2, 1, 1} stage, so that the question is mapped Stage 2 which is the highest stage among {2, 1, 1} stage.

## Discussion and Future Works

In this paper, we proposed a theoretical framework with three criteria (i.e., memory capacity, logical complexity, and DIKW hierarchy) to construct a hierarchical Q&A dataset for video story understanding. A key contribution of our work is to suggest an approach to classify the difficulties of questions for video story understanding in accordance with three criteria. Interestingly, the three criteria can be mapped with the five stages of human development in Piaget's theory, which can serve as a basis for Q&A systems to evaluate video story understanding.

While suggested three criteria can be linked to human developmental stages, it is hard to define separately and exactly. This is due to the limitation of Piaget's theory, where the human development stage cannot be discontinuous. That is, each developmental stage is not always precisely differentiated by age and cognitive abilities. Especially, the agreement for development stage of the high-level cognition is still controversial. Nevertheless, Collis' stages are the appropriate attempt to classify understanding according to human development stages. Furthermore, the connection from the three criteria to five developmental stages presents the possibility that our story-enabled intelligence can be associated to cognitive development stages of human. Applying knowledge about human cognitive development will help to set the direction of human-level AI research in detail.

Moreover, comparing with human development stages, machine learning approach requires clear and explicit specification. For example, in Collis' classification, each stage has an approximate two-year interval, while the machine needs to be viewed with a more detailed and specific classification criterion - such as either every month or every season.

As future work, we plan to extend the proposed criteria to reflect some viewpoints from cognitive narratology (e.g., Zwaan's five index model of narrative understanding including space, time, characters, goals, and causation(Zwaan 1999)). We also plan to construct a carefully designed hierarchical dataset, using the proposed criteria as a guideline for video story understanding. The proposed video Q&A hierarchy can be used as a metric for the developmental level of machine intelligence, and as a guidance to what dataset should be collected to study the desired level of machine intelligence.

## References

Ackoff, R. L. 1989. From data to wisdom. *Journal of applied systems analysis* 16(1):3–9.

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. VQA: Visual Question Answering: www.visualqa.org. *International Journal of Computer Vision* 123(1):4–31.

Barlas, I.; Ginart, A.; and Dorrity, J. L. 2005. Self-evolution in knowledge bases. In *Autotestcon, 2005. IEEE*, 325–331. IEEE.

Black, J., and Bower, G. 1980. Story understanding as problem solving. *Poetics* 9:223–250.

Blair, D., and Meyer, T. 1997. *Tools for an interactive virtual cinema*. Berlin, Heidelberg: Springer Berlin Heidelberg. 83–91.

Carlisle, J. P. 2006. Escaping the veil of maya: Wisdom and the organization. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 7, 162a–162a.

Case, R. 1980a. Implications of neo-piagetian theory for improving the design of instruction. *Cognition, development, and instruction* 161–186.

Case, R. 1980b. The underlying mechanism of intellectual development. *Cognition, development, and instruction*.

Chatman, S. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell paperbacks. Cornell University Press.

Collis, K. 1972. Concrete to abstract–a new viewpoint. *Australian Mathematics Teacher* 28(3):113–118.

Collis, K. F. 1975a. *The development of formal reasoning*. University of Newcastle.

Collis, K. F. 1975b. *A Study of Concrete and Formal Operations in School Mathematics: A Piagetian Viewpoint*. Hawthorn Vic : Australian Council for Educational Research.

Despres, C., and Chauvel, D. 2012. *Knowledge horizons*. Routledge.

Genette, G. 1980. *Narrative Discourse: An Essay in Method*. Cornell paperbacks. Cornell University Press.

Graesser, A. C.; Singer, M.; and Trabasso, T. 1994. Constructing inferences during narrative text comprehension. *Psychological Review* 101:371–395.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 1988–1997. IEEE.

Kim, K.-M.; Heo, M.-O.; Choi, S.-H.; and Zhang, B.-T. 2017. Deepstory: Video story QA by deep embedded memory networks. In *IJCAI International Joint Conference on Artificial Intelligence*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Maharaj, T.; Ballas, N.; Rohrbach, A.; Courville, A.; and Pal, C. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January:7359–7368.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, 1–9.

Mateas, M., and Sengers, P. 1999. Narrative intelligence. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, 1–10.

McKoon, G., and Ratcliff, R. 1992. Inference during reading. *Psychological Review* 99(3):440–466.

Mclaughlin, G. H. 1963. Psycho-logic: A possible alternative to piaget's formulation. *British Journal of Educational Psychology* 33(1):61–67.

Mun, J.; Seo, P. H.; Jung, I.; and Han, B. 2017. Marioqa: Answering questions by watching gameplay videos. In *ICCV*.

Pascual-Leone, J. 1969. *Cognitive development and cognitive style: A general psychological integration*.

Piaget, J. 1972. Intellectual evolution from adolescence to adulthood. *Human development* 15(1):1–12.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*, 2953–2961.

Rowley, J. 2007. The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science* 33(2):163–180.

Schumaker, R. P. 2011. From data to wisdom: the progression of computational learning in text mining. *Communications of the IIMA* 11(1):4.

Seol, S.; Sharp, A.; and Kim, P. 2011. Stanford mobile inquiry-based learning environment (smile): using mobile phones to promote student inquires in the elementary classroom. In *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*, 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. MovieQA: Understanding Stories in Movies through Question-Answering.

Trabasso, T., and Van Den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language* 24(5):612–630.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and van den Hengel, A. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40(10):2413–2427.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4995–5004.

Zwaan, R. A.; Langston, M. C.; and Graesser, A. C. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science* 6(5):292–297.

Zwaan, R. A. 1999. Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science* 8(1):15–18.