MovieQA: Understanding Stories in Movies through Question-Answering

Kunsu OH

Visual Question Answering at ModuLab 190327

Contents

Abstract

We introduce the MovieQA dataset which aims to evaluate automatic story comprehension from both video and text. The dataset consists of 14,944 questions about 408 movies with high semantic diversity. The questions range from simpler "Who" did "What" to "Whom", to "Why" and "How" certain events occurred. Each question comes with a set of five possible answers; a correct one and four deceiving answers provided by human annotators. Our dataset is unique in that it contains multiple sources of information – video clips, plots, subtitles, scripts, and DVS [32]. We analyze our data through various statistics and methods. We further extend existing QA techniques to show that question-answering with such open-ended semantics is hard. We make this data set public along with an evaluation benchmark to encourage inspiring work in this challenging domain.

It is new dataset for Movie Question and Answering…

Many Q and As

Hard works by annotators…

Videos and texts…

Main idea of MovieQA

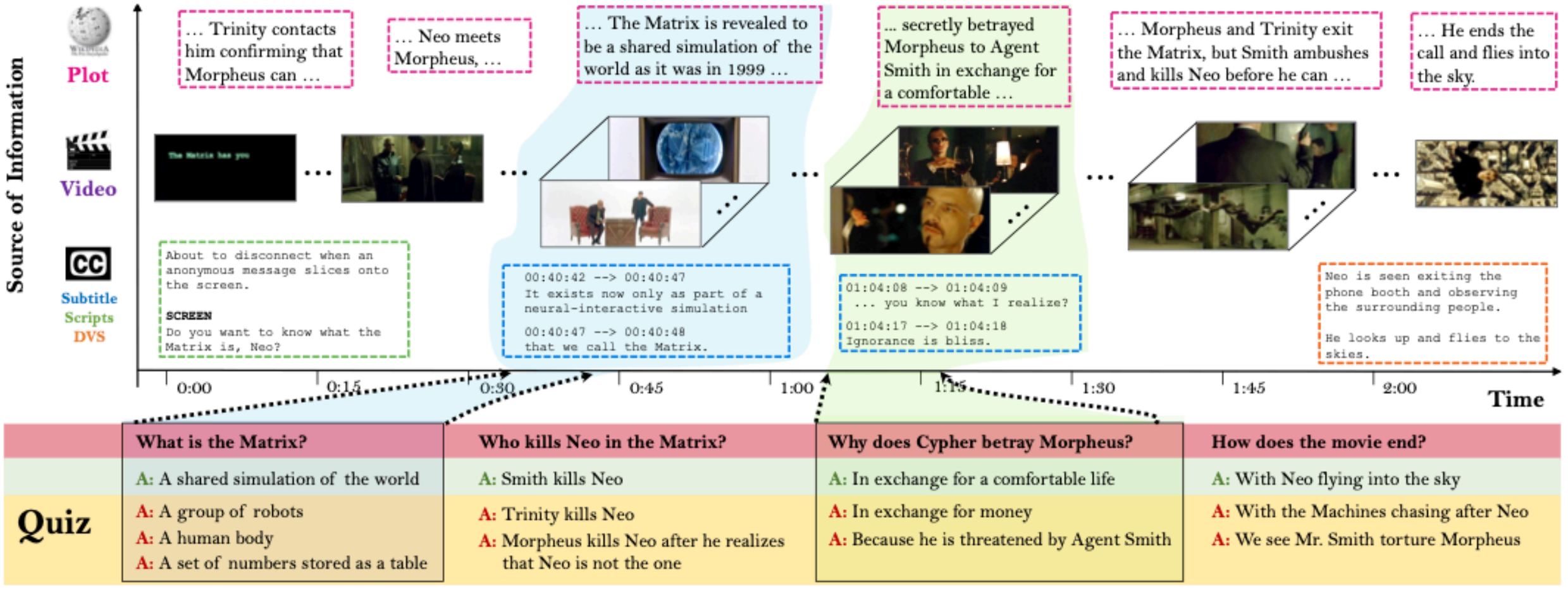

Figure 1: Our MovieQA dataset contains 14,944 questions about 408 movies. It contains multiple sources of information: plots, subtitles, video clips, scripts, and DVS transcriptions. In this figure we show example QAs from *The Matrix* and localize them in the timeline.

- MovieQA dataset includes videos, QAs, and some texts
  — Plot, subtitles, scripts, DVS.

Distributions of data

	TRAIN	VAL	TEST	TOTAL
Movies with Plots and Subtitles				
#Movies	269	56	83	408
#QA	9848	1958	3138	14944
Q #words	9.3	9.3	9.5	9.3 ± 3.5
CA. #words	5.7	5.4	5.4	5.6 ± 4.1
WA. #words	5.2	5.0	5.1	5.1 ± 3.9
Movies with Video Clips				
#Movies	93	21	26	140
#QA	4318	886	1258	6462
#Video clips	4385	1098	1288	6771
Mean clip dur. (s)	201.0	198.5	211.4	202.7 ± 216.2
Mean QA #shots	45.6	49.0	46.6	46.3 ± 57.1

Table 1: MovieQA dataset stats. Our dataset supports two modes of answering: text and video. We present the split into train, val, and test splits for the number of movies and questions. We also present mean counts with standard deviations in the total column.

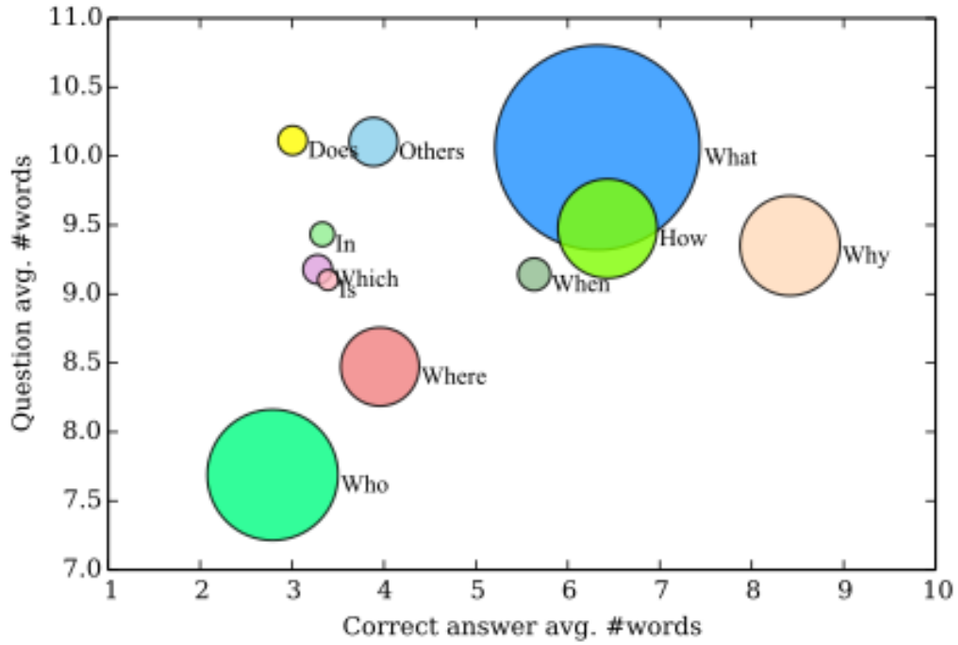

Figure 3: Average number of words in MovieQA dataset based on the first word in the question. Area of a bubble indicates #QA.

- Brablabalalaabalbalbal…

Distributions of data

	Txt	Img	Vid	Goal	Data source	AType	#Q	AW
MCTest [31]	✓	-	-	reading comprehension	Children stories	MC (4)	2,640	3.40
bAbI [44]	✓	-	-	reasoning for toy tasks	Synthetic	Word	20×2,000	1.0
CNN+DailyMail [12]	✓	-	-	information abstraction	News articles	Word	1,000,000*	1*
DAQUAR [23]	-	✓	-	visual: counts, colors, objects	NYU-RGBD	Word/List	12,468	1.15
Visual Madlibs [47]	-	✓	-	visual: scene, objects, person, ...	COCO+Prompts	FITB/MC (4)	2×75,208*	2.59
VQA (v1) [1]	-	✓	-	visual understanding	COCO+Abstract	Open/MC (18)	764,163	1.24
MovieQA	✓	✓	✓	text+visual story comprehension	Movie stories	MC (5)	14,944	5.29

Table 2: A comparison of various QA datasets. First three columns depict the modality in which the story is presented. AType: answer type; AW: average # of words in answer(s); MC (N): multiple choice with N answers; FITB: fill in the blanks; *estimated information.

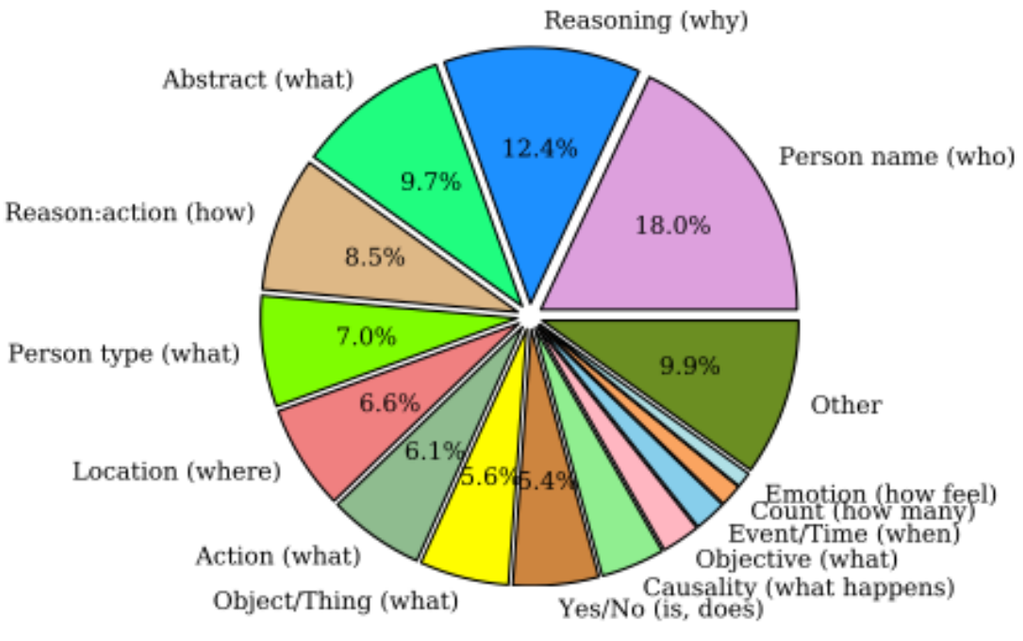

Figure 4: Stats about MovieQA questions based on answer types. Note how questions beginning with the same word may cover a variety of answer types: *Causality*: What happens ... ?; *Action*: What did X do? *Person name*: What is the killer's name?; *etc*.

Text type	# Movies	# Sent. / Mov.	# Words in Sent.
Plot	408	35.2	20.3
Subtitle	408	1558.3	6.2
Script	199	2876.8	8.3
DVS	60	636.3	9.3

Table 3: Statistics for the various text sources used for answering.

Tests…

Method	Plot	DVS	Subtitle	Script
Cosine TFIDF	47.6	24.5	24.5	24.6
Cosine SkipThought	31.0	19.9	21.3	21.2
Cosine Word2Vec	46.4	26.6	24.5	23.4
SSCB TFIDF	48.5	24.5	27.6	26.1
SSCB SkipThought	28.3	24.5	20.8	21.0
SSCB Word2Vec	45.1	24.8	24.8	25.0
SSCB Fusion	**56.7**	24.8	27.7	28.7
MemN2N (w2v, linproj)	40.6	**33.0**	**38.0**	**42.3**

Table 5: Accuracy for Text-based QA. **Top**: results for the Searching student with cosine similarity; **Middle**: Convnet SSCB; and **Bottom**: the modified Memory Network.

Method	Video	Subtitle	Video+Subtitle
SSCB all clips	21.6	22.3	21.9
MemN2N all clips	**23.1**	**38.0**	**34.2**

Table 6: Accuracy for Video-based QA and late fusion of Subtitle and Video scores.

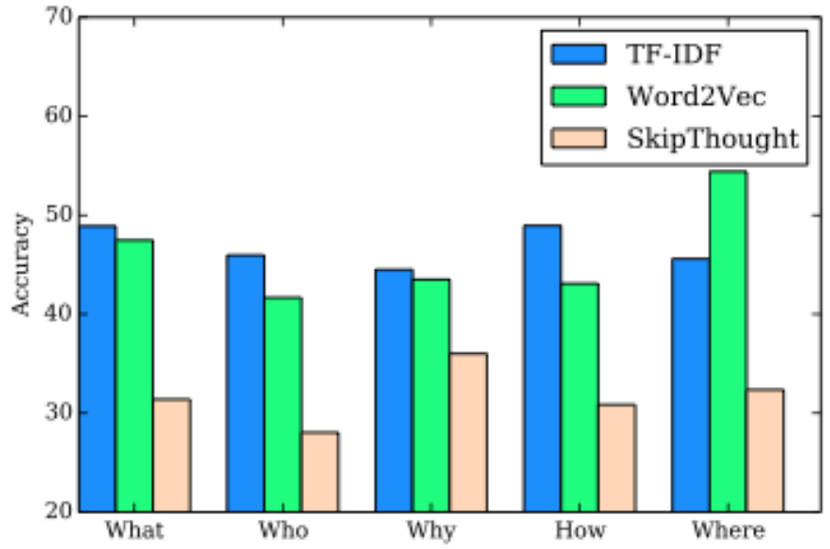

Figure 6: Accuracy for different feature representations of plot sentences with respect to the first word of the question.

- Simple tests said videoQA is not good compared with text only… hm.. and 3 years later…

http://movieqa.cs.toronto.edu/leaderboard/

Recent tests…

W Plot Synopses only

Date	Team Name	Affiliation	Accuracy	Notes	Paper	Code
04 Aug 2018, 13:30	IMS	Institute for Natural Language Processing, University of Stuttgart	85.12	Attention-Based Matching Network (LSTM)		

Scripts only

Date	Team Name	Affiliation	Accuracy	Notes	Paper	Code
08 Nov 2018, 22:15	Anonymous	Anonymous	45.49	Universal QA model which works with any story	-	-

DVS only

Date	Team Name	Affiliation	Accuracy	Notes	Paper	Code
08 Nov 2018, 22:15	Anonymous	Anonymous	49.65	Universal QA model which works with any story	-	-

Subtitles only

Date	Team Name	Affiliation	Accuracy	Notes	Paper	Code
08 Nov 2018, 22:15	Anonymous	Anonymous	44.01	Universal QA model which works with any story	-	-

Movie: Video+Subtitles

Date	Team Name	Affiliation	Accuracy	Notes	Paper	Code
08 Nov 2018, 22:15	Anonymous	Anonymous	46.98	Universal QA model which works with any story	-	-
20 Jan 2019, 23:26	Anonymous	Anonymous	45.31	new method to optimize all MEM network	-	-
12 Feb 2019, 08:37	Anonymous	Anonymous	44.52	Video clip feature and Text feature fusion	-	-
23 Jan 2019, 09:02	Anonymous	Anonymous	44.12	come up with a new method for QA	-	-
27 Jan 2019, 21:42	Anonymous	Anonymous	43.56	A new method to optimize MemNetwork	-	-