On the hidden treasure of dialog in video question answering

Deniz Engin[1,2] François Schnitzler[2] Ngoc Q. K. Duong[2] Yannis Avrithis[1]

[1] Inria, Univ Rennes, CNRS, IRISA [2]InterDigital

Introduction

Goal: Knowledge-based video question answering on TV shows without using human-annotated knowledge

Problem: Question not answerable only from the given scene → high-level understanding of whole episode/show required

Recent Works: Rely on human-annotated knowledge from dataset or plot summaries

Leonard: Come on. Is that really necessary?
Sheldon: Leonard, I believe it is. This is trash talk. Trash talk is a traditional component in all sporting events.
Sheldon: Kripke your robot is inferior and it will be defeated by ours because ours exceeds yours in both design and execution.

What did the guys name their robot?

A) Killer Robot B) Terminator
C) Monte D) Crippler

Motivation

- Substituting human-generated knowledge by automatically generated summaries from the raw dialog

Episode A - Scene 1

Leonard: Come on. Is that really necessary?
Sheldon: Leonard, I believe it is. This is trash talk. Trash talk is a traditional component in all sporting events.
Sheldon: Kripke your robot is inferior and it will be defeated by ours because ours exceeds yours in both design and execution.

Scene Dialog Summary

Kripke is going to name his robot Scrap Metal. Sheldon and Leonard are going to defeat Kripke's robot because theirs is better in design and execution.

Episode A - Scene 2

Kripke: Word around the plasma lab is you built a robot?
Leonard: Yes, we did, Kripke.
Sheldon: His name is Monte.

Scene Dialog Summary

Leonard and Raj have built a robot called Monte. Kripke is going to enter him in the Southern California Robot Fighting League Round Robin Invitational.

QA (Episode A - Scene 1)

What did the guys name their robot?

A) Killer Robot B) Terminator
C) Monte D) Crippler

Episode Dialog Summary

... Leonard and Raj have built a robot called Monte. Kripke is going to enter him in the Southern California Robot Fighting League Round Robin Invitational. ...

Dialog Summarization: Character Names Matter

Following a dialog summarization method considering char names [1]

- Extracting embeddings of utterances by Sentence-BERT
- Segmentation of utterances according to topic and stage views
- Encoding each segmented conversation and generating summaries via multi-view encoder

Dialog

Bernadette: Knock, knock.
Howard: Oh, great, you made it. Come on in
Howard: I invited her.
Bernadette: So where's the telescope?

Episode Dialog Summary

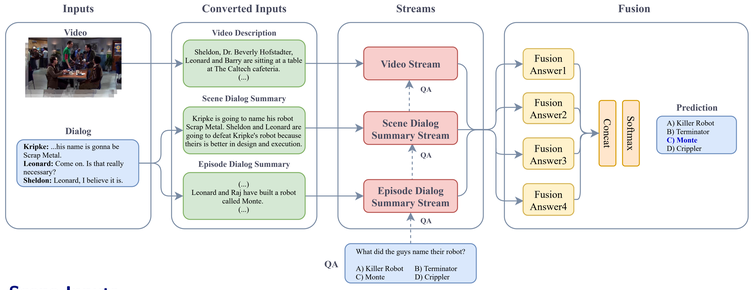
(...)
It's a Romulan battle bagel, not a starship. Howard invited Bernadette in. The telescope is in Hawaii, but Raj controls it from here.
(...)

QA

Who did Howard invite to join him and Raj in Raj's lab?

A) Bernadette B) Leonard
C) Penny D) Amy

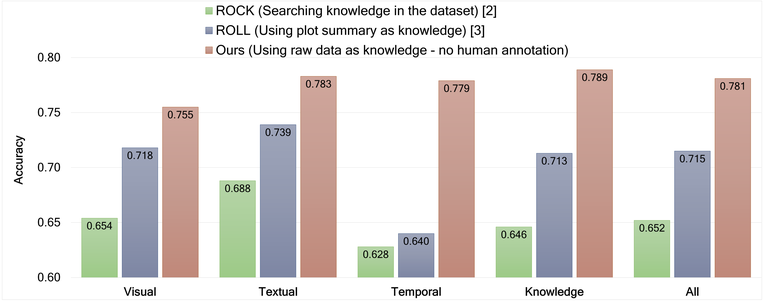
Our Method



- **Scene Inputs**
  - **Video:** Converting video into text description by applying a set of rules on generated scene graph via visual recognition pipelines by following [3]
  - **Scene Dialog Summary:** Dialog summarization
- **Episode Inputs**
  - **Episode Dialog Summary:** Concatenating scene dialog summaries for all scenes of an episode
- **Single-Stream QA:** Training BERT
  - Splitting episode inputs into parts for training → weakly supervised localization over parts required
- **Multi-Stream QA:** Fusing extracted features from each single stream

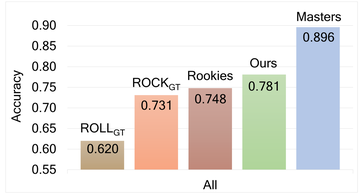
Comparison with the state-of-the-art

- Our method outperforms the state-of-the-art on the KnowIT VQA dataset by a large margin **without** using question-specific human annotation or human-made plot summaries



Comparison with human evaluators

- Our method outperforms **human evaluators** who have never watched any whole episode before, and recent works which uses **GT knowledge**



Rookies: Human evaluators never watched any episode [2]
Masters: Human evaluators watched the show [2]
ROCK$_{GT}$: Using GT knowledge from the dataset [2]
ROLL$_{GT}$: Using GT knowledge from the dataset [1]
Ours: Using raw data as knowledge - no human annotation

Qualitative Results: Knowledge-based QA

Video Description

(...)
Sheldon sitting on chair. Curtain and building behind Sheldon.
(...)

Scene Dialog Summary

Sheldon will send him an email when they get back. He needs to read it.
(...)

Episode Dialog Summary

(...)
Sheldon forgot his flash drive, so he has to go back and get it.
(...)

QA

What has Sheldon forgotten here?

A) His flash drive B) His thesis
C) His suitcase D) His laptop

Attention

Video Description ■ Scene Dialog Summary ■ Episode Dialog Summary

Contributions

- Building **a knowledge-base VideoQA system** without extra human annotation
- Applying **dialog summarization** to VideoQA
- Introducing **a weakly-supervised soft temporal attention approach** for localization by a linear combination of most confident parts
- Introducing **a simple fusion method** by applying multi-stream attention over each input stream

References

1. Jiaao Chen and Diyi Yang. Multi-view sequence-tosequence models with conversational structure for abstractive dialogue summarization. *In Proc. EMNLP*, 2020.
2. Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering knowledge-based questions about videos. *In Proc. AAAI*, 2020.
3. Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. *In Proc. ECCV*, 2020.

Project Page