MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding

Revanth Gangi Reddy[1], Xilin Rui[2], Manling Li[1], Xudong Lin[3], Haoyang Wen[1], Jaemin Cho[4], Lifu Huang[5], Mohit Bansal[4], Avirup Sil[6], Shih-Fu Chang[3], Alexander Schwing[1], Heng Ji[1]

[1]University of Illinois at Urbana-Champaign [2]Tsinghua University [3]Columbia University
[4]University of North Carolina at Chapel Hill [5]Virginia Tech [6]IBM Research AI

MuMuQA: Multimedia Multi-hop QA

- **Motivation:** To answer questions about news articles, humans seamlessly combine context from multiple modalities, such as images and text. Images in the real world, especially in news, have objects that are co-referential in the text.
- **New QA Task:** Given an image-caption pair and its associated news body text, a question is answered by extracting a short span from the body text.
- Answering the questions requires ***multi-hop reasoning***:
  - The first hop requires cross-media grounding between image and caption to get the ***bridge item***.
  - The second hop requires reasoning over body text using the bridge item to extract the final answer.

Image - Caption	Body Text
 Israeli Prime Minister Benjamin Netanyahu (R) speaks with Finance Minister Moshe Kahlon during the weekly cabinet meeting in Jerusalem	A dispute between Israeli Prime Minister Benjamin Netanyahu and his finance minister over broadcast regulation sparked speculation on Sunday that Netanyahu could seek an election two years ahead of schedule. ... The Israeli media quoted Netanyahu as telling ministers from his Likud party that he would dissolve the government if Kahlon didn't fall into line. Kahlon heads the Kulanu party, a centre-right partner in Netanyahu's ...

Question: What party does the person with the blue tie in the image belong to?
Answer: Likud
Bridge item: Benjamin Netanyahu

Image - Caption	Body Text
 Opposition supporters clash with security forces during a rally against Venezuela's President Nicolas Maduro in Caracas, Venezuela, April 26, 2017.	Venezuelan security forces fired scores of tear gas volleys and turned water cannons on rock-throwing protestors on a bridge in Caracas on Wednesday as the death toll from this month's anti-government unrest hit at least 29. ... Red-shirted supporters of Maduro, the 54-year-old former bus driver who succeeded Hugo Chavez in 2013, also rallied on the streets of the capital, punching their fists in the air and denouncing opposition "terrorists".

Question: What are the people in the image accused of behaving like?
Answer: terrorists
Bridge item: Opposition supporters

Figure: Two examples from our evaluation benchmark with the question-answer pairs and their corresponding news articles.

Contributions

- We release a new QA evaluation benchmark, **MuMuQA**, based on multi-hop reasoning and cross-media grounding of information present in news articles.
- We introduce a novel pipeline to automatically generate silver-standard training data for this task.
- We provide competitive baselines that leverage different modalities and demonstrate the benefit of using multimodal information.

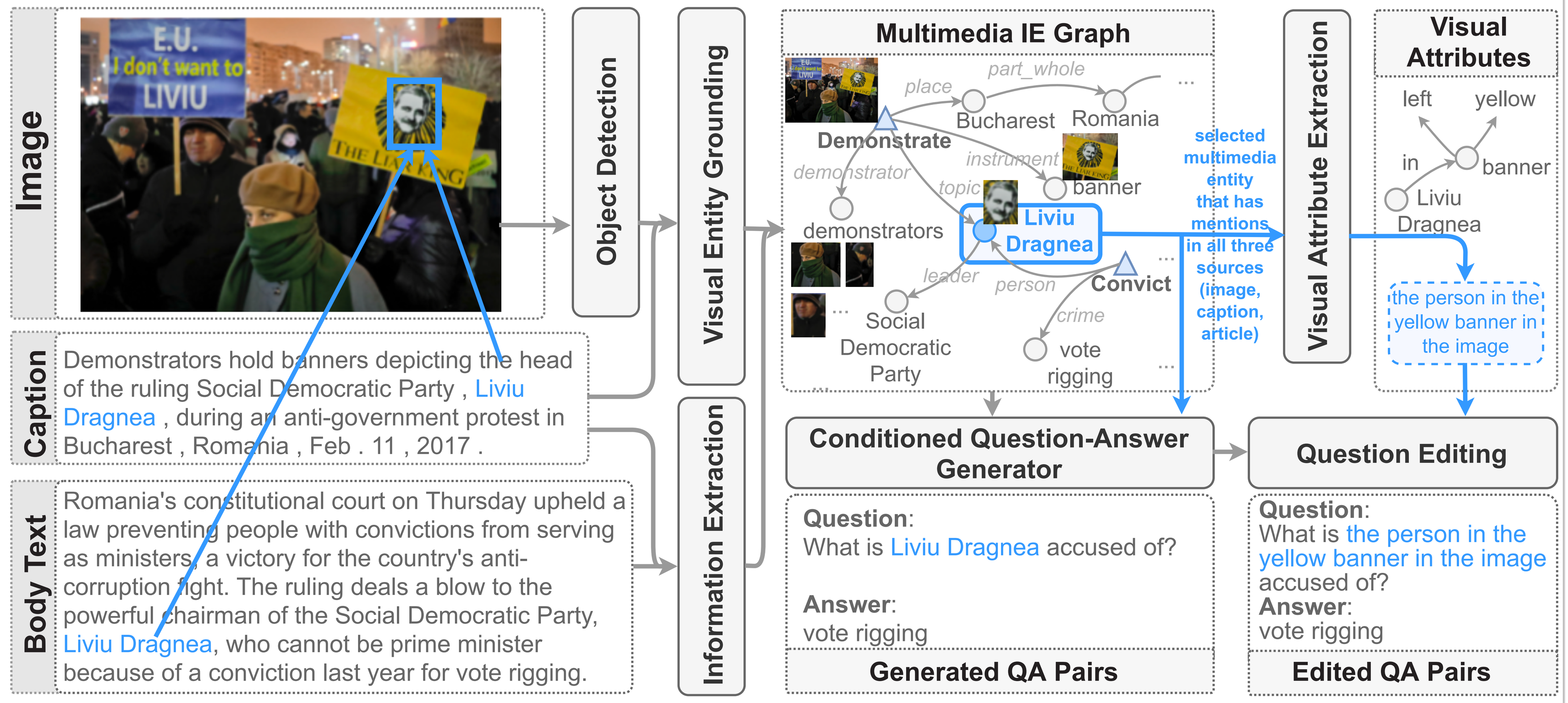
Benchmark Construction

- Our benchmark consists of an evaluation set that is human-annotated and a silver-standard training set that is automatically generated.
- News articles are shown in the interface along with their images and corresponding captions.
- The annotator first looks at the image-caption pair to identify which objects in the image are grounded in the caption and then creates a question about the grounded entity.
- For automatic quality control, the interface also provides access to a single-hop text-only QA model to ensure the questions cannot be directly answered using news body text.
- The evaluation set contains 1384 questions with 263 in dev and 1121 in test.

Silver Training Set Generation

The automatic training set generation process consists of the following steps:

- **Multimedia Entity Grounding:** Identify objects in images that are grounded in text.
- **Visual Attribute Extraction:** Generate visual descriptions for the objects in images.
- **Question Generation:** Generate questions about the cross-media grounded entities.
- **Question Editing:** Replace grounded entity mention with its visual description.
- **Question Filtering:** Discard questions answerable using a text-only QA model.



Baselines

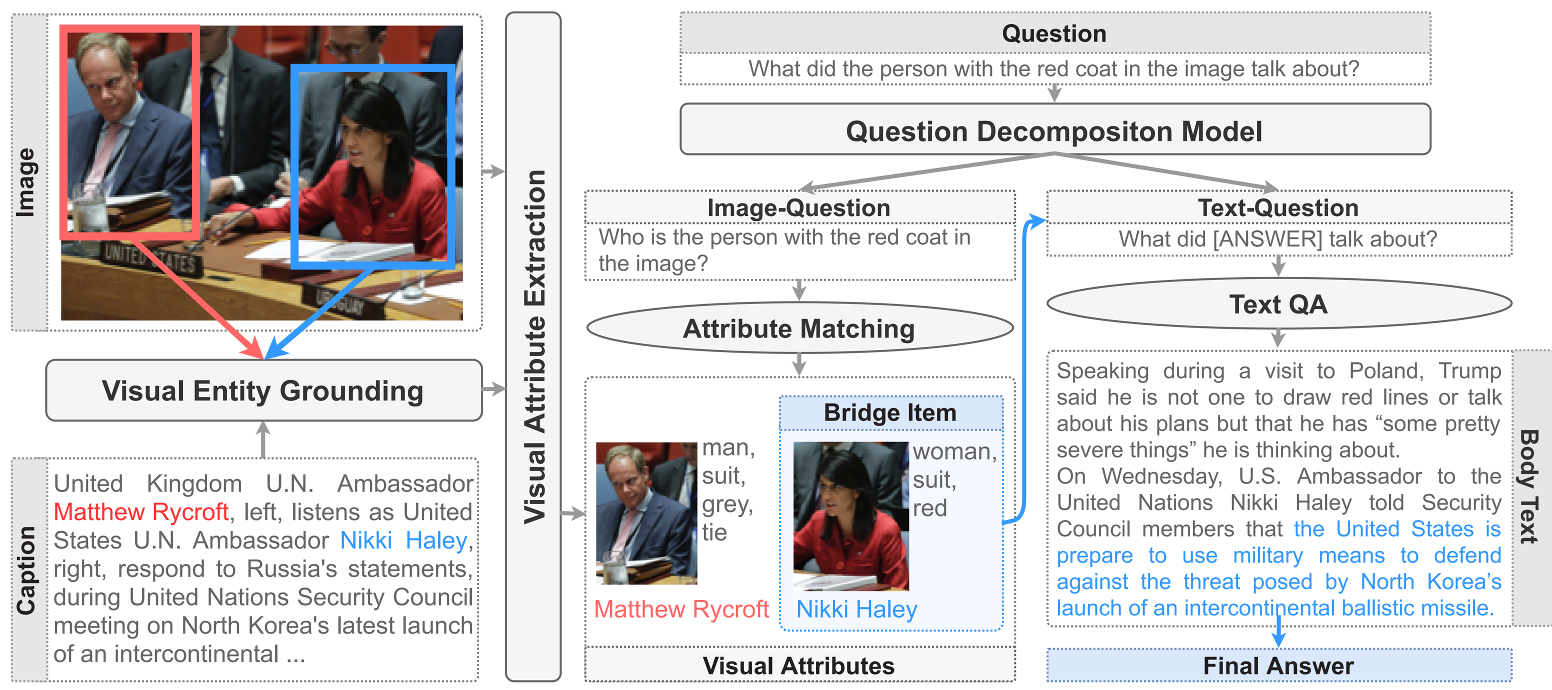
Multi-Hop Text-only QA

- We use an extractive text-only QA model that takes the question, caption and body text as input.
- The model is based on Bert-large and is trained on HotpotQA.

End-to-end Multimedia QA

- We finetune a pre-trained multimodal model for our task.
- We add an extractive answer predictor to OSCAR and finetune using 20k synthetic training examples.

Pipeline-based Multimedia QA



Experiments

- Pipeline-based QA system has bridge F1 of 29.8% compared to human performance of 78.8%.
- Underperformance of end-to-end multimedia QA system could be due to OSCAR being pre-trained with image-caption pairs, which makes it potentially not suited for reasoning over larger text input.
- Grounding system captures the bridge item in 45% of the cases.

Model	Dev	Test
Multi-hop Text-only QA	18.5	16.5
End-to-end Multimedia QA	12.1	11.5
Pipeline-based Multimedia QA	**33.9**	**30.8**
Human Baseline	-	66.5

Table: F1 performance of different baselines.

Conclusion

- We introduce a new challenging multi-hop QA task, **MuMuQA**, that requires cross-media grounding over images, captions and news body text.
- We demonstrate the benefit of using multimedia knowledge extraction, both for generating silver-standard training data and for a pipeline-based multimedia QA system.