

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

Revanth Gangi Reddy¹, Xilin Rui², Manling Li¹, Xudong Lin³, Haoyang Wen¹, Jaemin Cho⁴, Lifu Huang⁵, Mohit Bansal⁴, Avirup Sil⁶, Shih-Fu Chang³, Alexander Schwing¹, Heng Ji¹

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



Multimodal information in news

- Images in news have objects that are co-referential to the text, with complementary information provided in both modalities.
- Most QA research focuses on using information from a single modality (TextQA, VQA, VideoQA)
- It is necessary to incorporate information from multiple modalities to help understand the entire context necessary for answering questions in news.



Every New York Times front page since 1852.


Video Link: [Link](#)

Video credit: [Josh Begley](#)

MuMuQA: Multimedia Multi-hop QA

- Given a news article with an image-caption pair and a question, a system needs to answer the question 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 the news body text by using the bridge item to extract the final answer.
- Our benchmark reflects questions that news readers might have after looking at the visual information in the news article, without having read the relatively longer body text.

Image - Caption	Body Text
 <p>Israeli Prime Minister Benjamin Netanyahu (R) speaks with Finance Minister Moshe Kahlon during the weekly cabinet meeting in Jerusalem</p>	<p>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.</p> <p>...</p> <p>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 center-right partner in Netanyahu's ...</p>
<p>Question: What party does the person with the blue tie in the image belong to? Answer: Likud</p>	

Image - Caption	Body Text
 <p>Opposition supporters clash with security forces during a rally against Venezuela's President Nicolas Maduro in Caracas, Venezuela, April 26, 2017.</p>	<p>Venezuelan security forces fired scores of tear gas volleys and turned water cannons on rock-throwing protesters on a bridge in Caracas on Wednesday as the death toll from this month's anti-government unrest hit at least 29.</p> <p>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."</p>
<p>Question: What are the people in the image accused of behaving like? Answer: terrorists</p>	

Contributions

- We release a new QA evaluation benchmark, **MuMuQA**, based on multi-hop reasoning and cross-media grounding of information present in news articles.
- Our work is the first to attempt using information grounded in the image in an extractive QA setting.
- To automatically generate silver-standard training data for this task, we introduce a novel pipeline that incorporates cross-media knowledge extraction, visual understanding and synthetic question generation.
- We provide competitive baselines that leverage different modalities and demonstrate the benefit of using multimodal information.

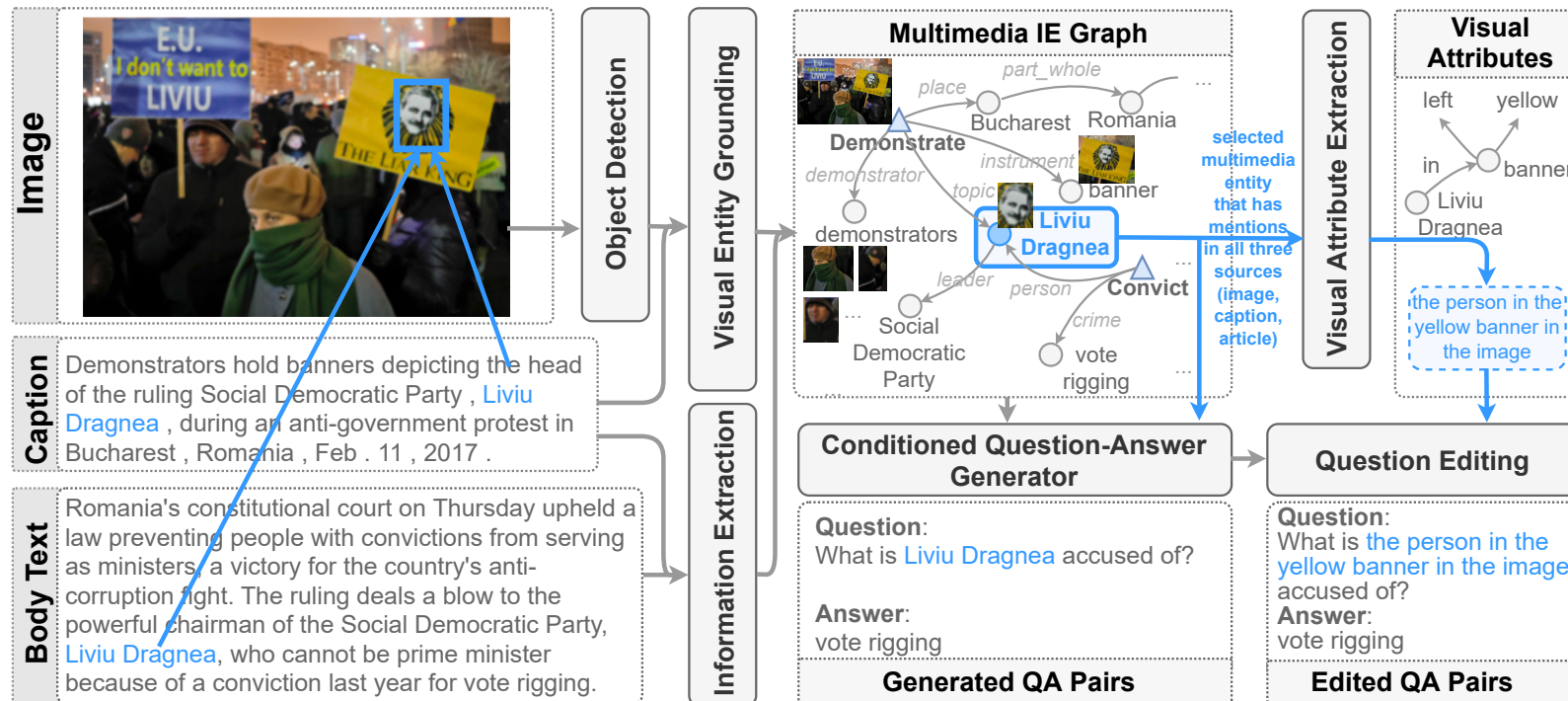
Benchmark Construction

- Our dataset 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 and annotators are asked to create questions.
- The process requires the annotator to first look at the image-caption pair to identify which objects in the image are grounded in the caption and then create 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:** To identify objects in images that are grounded in captions and body text.
- **Visual Attribute Extraction:** To generate visual descriptions of the objects in images.
- **Conditioned Question Generation:** To generate questions about the cross-media grounded entities.
- **Question Editing:** Replacing grounded entity mention with its visual description
- **Question Filtering:** Discarding synthetic questions answerable using a single-hop 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 HotpotQA¹.

End-to-end Multimedia QA

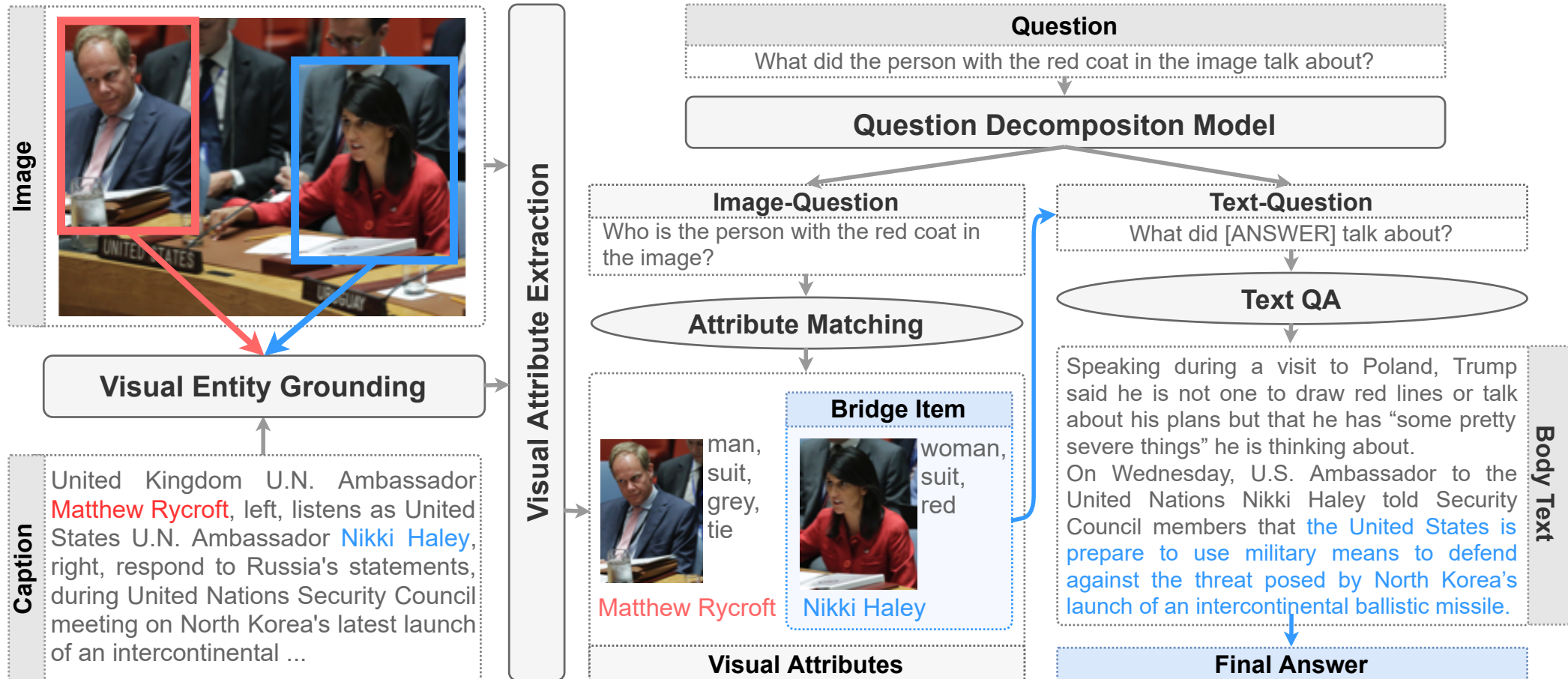
- We finetune a pre-trained multimodal model for our task.
- We add an extractive answer predictor to OSCAR² and finetune using 20k synthetic training examples.
- The corresponding image features and object-labels are obtained using Faster-RCNN³

¹ HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering; Yang et. al., EMNLP 2018

² Oscar: Object-semantics aligned pre-training for vision-language tasks; Li et. al., ECCV 2020

³ Bottom-up and top-down attention for image captioning and visual question answering; Anderson et. al., CVPR 2018

Pipeline-based Multimedia QA




Results and Analysis

- The pipeline-based QA system considerably outperforms other systems, but still lacks behind human performance.
- Interesting to see end-to-end multimedia QA system underperform multi-hop text-only QA model.
- This could be due to OSCAR being pre-trained with image-caption pairs, which makes it potentially not suited for reasoning over larger text input.
- Pipeline-based QA system has bridge F1 of 29.8% compared to human performance of 78.8%
- We observe that the bridge item is covered by grounding system in 45% of the cases.
- Whenever bridge item is covered in grounding, attribute matching system picks the correct bridge item in 60% 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
<i>Human Baseline</i>	-	66.5

F1 Performance (%) of different baselines on the MuMuQA evaluation benchmark.



Caption: A **woman** places flowers on an altar set up in honor of **Berta Caceres** during a demonstration outside Honduras' embassy in Mexico City, June 15, 2016.

Question: Where was the person in the photo in the image from?

Bridge Item: **Berta Caceres**

An example where the grounding system failed to capture the gold bridge item (in green). The grounded entity is in blue in the caption and its corresponding bounding box is shown in blue in the image.

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.
- Future work will explore incorporating other forms of media in news, such as video and audio, to facilitate information seeking across diverse data sources.