Question Generation Evaluation Guidelines

Nedjma Ousidhoum and Andreas Vlachos

November 22, 2022

We are evaluating a question generation system for fact-checking. First, we assess each question independently, then we evaluate the whole set of generated questions.

1 Criteria for a good question

We assess each generated question based on the following criteria.

1.1 Intelligibility

The question should be fluent but does not have to be perfectly grammatical as long as it is understandable. The intelligibility of a question should be judged without looking at the claim.

Examples of intelligible vs. unintelligible questions

- **Unintelligible question** How many less do Florida's teachers pay? (*Incomprehensible.*)
- **Intelligible question** What made Rep. Paul Gosar to ask for the arrest of the illegal immigrants? (*Not perfectly grammatical but still intelligible.*)
- **Intelligible question** What is the average pay for Florida's teachers? (*Grammatical and intelligible.*)

1.2 Clarity

Questions should be precise enough to be answered confidently using a search engine regardless of the context. A clear question should not be too broad and should include all the necessary details, such as dates, and names of people/speaker, etc. If the details can be induced by looking at the claim, the question remains clear.

Examples of clear vs. unclear questions

- **Unclear question** What is the name of the state that New Jersey elects a Republican to the Senate? (*Unintelligible and unclear.*)
- **Unclear question** What policies violate federal law? (*Too broad.*)
- **Unclear question** What did the author of the bill say about the bill? (*Intelligible but unclear.*)
- **Clear question** What is the definition of a sanctuary city?
- **Clear question** What is the United Nations?
- **Clear question** What was the name of the law that separated children from adults entering America?
- **Claim** Apprehension rates at the southern border have plummeted since the 1980s and apprehensions of Mexicans specifically have reached their lowest point in nearly half a century.
  - **Clear question when looking at the claim (otherwise, unclear since the name of the country is not specified)** What was the apprehensions rate at the southern border in the 1980s?

1.3 Relevance

The generated questions should mention entities that are related to the claim. The entities can either be mentioned in the claim or in the metadata since we may use the latter to train a question generation system.

Examples of relevant vs. irrelevant questions

- **Claim** Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine.
  - **Irrelevant** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Irrelevant** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Relevant** Who was the Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine?
  - **Relevant** Who is Miss Universe Guyana 2017?
  - **Relevant** What is the name of the person arrested at London Heathrow airport?

1.4 Informativeness

An informative question should return answers that provide information about the claim in order to help us reach a verdict for its veracity. The informativeness of a fact-checking question will depend on the type of the claim. For instance, if the claim is a quote, a question which focuses on the person or entity who made the statement can be an informative one. On the other hand, if the claim focuses on the narration of a certain event, then an informative fact-checking question may focus on the event itself. A yes/no question which is useful to reach a verdict is considered to be informative. Questions, however, should not directly ask or imply that the claim is true or false. Finally, an informative question should not (indirectly) imply that the claim is true or false.

We use a 4-point Likert scale to assess the informativeness of a question. A question can be:

1. **uninformative** i.e. useless (0),
2. **weakly informative** i.e. unlikely to be helpful but we do not mind to have it generated by the system*(score = 1)* ,
3. **potentially informative** or somewhat useful, i.e. a question that could be helpful depending on the context of the claim. For instance, a question whose answer is in the claim can be informative if it is worth verifying *(score = 2)*,
4. **informative** i.e. crucial *(score = 3)*.

Examples questions scored according to their informativeness

- **Claim** Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine.
  - **Irrelevant and uninformative** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Relevant and weakly informative** What is the name of the person arrested at London Heathrow airport?
  - **Relevant and potentially informative** Who is Miss Universe Guyana 2017?
- **Claim** You will learn more about Donald Trump by going down to the Federal Election Commission to see the financial disclosure form than by looking at tax returns.
  - **Relevant and uninformative** Where is Donald Trump's tax return?
  - **Relevant and weakly informative** How can you learn more about Donald Trump by looking at tax returns?

  - **Relevant and weakly informative** How can you learn more about Donald Trump by going down to the Federal Election Commission?
  - **Relevant and potentially informative** How much does Donald Trump donate to charity?
  - **Relevant and informative** What is Donald Trump's tax rate?
  - **Relevant and informative** What type of taxes does Donald Trump pay?

- **Claim** If Congress fails to act the Obama administration intends to give away control of the internet to an international body akin to the United Nations.
  - **Relevant and uninformative** What constitutes an international body?
  - **Relevant and uninformative** What would happen if Congress did not act?
  - **Relevant and weakly informative** Who has oversight over the internet in America?
  - **Relevant and weakly informative** What countries have officers involved in the Internet Corporation for Assigned Names and Numbers?
  - **Relevant and weakly informative** What is the United Nations?
  - **Relevant and potentially informative** What did the Obama administration intend to do with control of the internet?
  - **Relevant and informative question** What organization does the Obama administration want to give control of the internet to?

1.5 Prerequisites for the different criteria

1. Intelligibility should be judged without looking at the claim.
2. A clear question should be intelligible. When assessing the clarity, the claim can be checked for more details.
3. Relevance and informativeness should be annotated by looking at the claim.
4. A relevant question needs to be intelligible.
5. An informative question is intelligible, clear, and relevant.