Annotation Guidelines for AVeriTeC

v0.8 - July 5, 2021

1 Discussion

Here we list several issues for further discussions.

2 Introduction

We aim to construct a dataset which expands the factual verification task in several ways. We decompose the evidence retrieval process into steps, annotating each individual step as a question-answer pair (see Figure 1). Successful models must decompose claims, produce questions, retrieve answers, and combine answers into verdicts.

This allows us to progress beyond simple factoid claims by enabling partial evidence collection for complex statements. For example, in Figure 1, verifying the claim requires knowing the salary of the health commissioner, the governor, the vice president and Dr. Fauci, so that they can be compared. Therefore, four separate questions about salary need to be asked in order to reach a verdict (i.e. that the claim is *supported*).

By decomposing the evidence retrieval process, we also produce a natural way for systems to justify their verdicts. We furthermore annotate claims with a final justification, providing a textual explanation of how to combine the retrieved answers to reach a verdict. This allows users can follow each step of the retrieval and verification processes, and so understand the reasoning employed by the system.

Using question-answer pairs for evidence retrieval has the further benefit of allowing us to use state-of-the-art question answering systems for evidence retrieval, simplifying the process of building models.

3 Annotation Procedure

The claims for the dataset are sourced from existing fact checking datasets, specifically DATACOMMONS. We may also want to use the Google Fact Checking API (see https://github.com/dcorney/fact_check_explorer). The annotation consists of three phases as shown below:

1. Metadata Collection and Claim Normalization.
2. Question Generation and Answering.
3. Verdict Validation and Quality Control.

Claim: **The (Erie County, N.Y.) health commissioner** makes more than the **governor**, the **vice president of the United States**, but less than **Dr. Anthony Fauci.**

Q1: What is the salary of the health commissioner of the Erie County?
A1: Erie County Health Commissioner Dr. Gale Burstein received $103,374 in overtime since the start of the Coronavirus pandemic. That's an increase of $19,378 since the Office of Erie County Comptroller's last report that looked at overtime for the pay periods beginning in March and ending July 3, 2020, and is in addition to her salary of $207,292.

Q2: What is the salary of the governor of New York?
A2: The salary of the governor of New York in $225,000.

Q3: What is the salary of the vice president of the United States?
A3: The Vice President's salary is currently set at $235,100. This annual income has been frozen by Congress since at least 2014.

Q4: What is the salary of Dr. Anthony Fauci?
A4: Dr. Anthony Fauci made $417,608 in 2019, the latest year for which federal salaries are available.

Verdict: **True**

Figure 1: Example claim and question answer pairs.

Each phase requires different annotators, and as such each claim will be annotated by three different annotators. The first annotator conducts the claim assessment and normalization. Some claims that are not verifiable would be filtered out in this stage, while some incomplete claims should be rewritten. Furthermore, this annotator will collect metadata about the original claim. This also includes annotating the claims with a ground truth, in order to normalise the labels from the various fact checking sites.These labels should mostly map from those of the fact checking sites, as phase 1 annotators do not independently attempt to verify claims. They are mainly for validation.

The second annotator for the claim will conduct question generation and answer generation based on the articles provided by fact checkers. We furthermore ask annotators in phase 2 to collect statistics on what types of information was used to verify the claim, e.g. reading pdfs or consulting tables. Furthermore, annotators in the second phase are asked to provide a label on the basis of their retrieved evidence. The labels collected in phase 2 will help us validate the true labels generated in phase 3.

The last annotator performs verdict validation and quality control, which is a means to ensure that the data collected in phase 2 is of high quality. They are not able to see the verdict given by the fact checker from the fact checking websites, nor the fact checking article itself. Instead, they are required to predict one of the four labels based on the generated question answer pairs. If the predicted verdict in this phase is not consistent with the verdict from phase 2, we will ask for another judgment. We will then either use this new label, or we will discard the claim. Annotators in phase 3 are furthermore asked to identify obviously wrong or otherwise problematic answers to questions, so that such QA pairs can be filtered out. Annotators in phase 3 furthermore provide a textual justification for their decisions.

3.1 Labels

For the dataset we are creating, we will be using four labels:

1. The claim is **supported**. The claim is fully supported by the arguments and evidence presented.
2. The claim is **refuted**. The claim is fully contradicted by the arguments and evidence presented.
3. There is **not enough information** to support or refute the claim. The evidence either directly argues that appropriate evidence cannot be found, or leaves some aspect of the claim neither supported nor refuted. In cases where the evidence used by the original fact checking article is not available (or no longer available) online, annotators in Phase 1 should not use NEI, while annotators in Phase 2 and Phase 3 should use NEI.
4. The claim is misleading due to **missing context**, but not explicitly refuted. This includes cherry picking, true-but-misleading claims (e.g. the claim *"Alice has never lost a game of chess"* with evidence showing Alice has never played chess), as well as cases where conflicting or internally contradictory evidence can be found. Missing context may also be relevant if a situation has changed over time, and the claim fails to mention this (e.g. *"Alice is a strong supporter of subsidies for chess players"* with evidence showing that Alice currently supports the position, but in the past opposed the position).

4 Phase 1: Metadata Collection and Claim Normalization

In the first phase, annotators collect metadata about claims, as well as producing a normalized version of each claim. We furthermore filter out claims featuring certain challenges that we do not intend for our dataset to include.

4.1 Claim Splitting

In our pilot, some claims consisted of multiple, easily separable, independent parts (e.g. *"The productivity rate in Scotland rose in 2017, and similarly productivity rose in Wales that year."*). The first step is to split these claims into their parts. Metadata collection and normalization will then be done independently for each part, and in subsequent phases parts will be treated as separate claims.

When splitting a claim, it is important to ensure that each part is independent from the others. This can be done either through metadata, or through rewriting. For example, for the claim *"Amazon is doing great damage to tax paying retailers. Towns, cities and states throughout the U.S. are being hurt - many jobs being lost!"*, it should be clear what is causing job loss in the second part (see https://www.politifact.com/article/2017/aug/16/trump-takes-aim-amazon-barbed-tweet/). A possible split would be *"Amazon is doing great damage to tax paying retailers"* and *"Towns, cities and states throughout the U.S. are being hurt by Amazon - many jobs being lost"*.

4.2 Metadata Collection

We ask annotators to collect metadata through the following three steps.

4.2.1 General Information

- A hyperlink to the original claim, if that is provided by the fact checking site. Examples of this include Facebook posts (see https://www.politifact.com/factchecks/2019/jan/03/facebook-posts/did-michelle-obama-once-say-white-folks-are-whats-/), the original article or blog post being fact checked (see https://www.politifact.com/factchecks/2017/aug/23/blog-posting/fake-news-rosa-parks-had-no-daughter-could-praise-/), and embedded video links (see https://www.politifact.com/factchecks/2017/jul/27/gavin-newsom/inland-empire-second-only-san-francisco-california/). If the original claim has a hyperlink on the fact checking site, but that hyperlink is dead, annotators should leave the field empty. We will use archive.is to ensure the persistence of links that are still up.
- The date of the original claim, regardless of whether it is necessary for verifying the claim. This date is often mentioned by the fact checker, but not in a standardized place where we could automatically retrieve it. Note that the date of origin for the *original claim* and the *fact checking article* may be different and both stated in text. We specifically need the original claim date, as we intend to filter out results published after that date during search. Furthermore, that date may be necessary for checking the claim.
- The speaker (or source) of the original claim. This will also help resolve ambiguities when producing questions.
- If the original source is an image that contains text (for example, the Facebook meme about Michelle Obama listed above), we ask the annotators to transcribe whatever text occurs in the image as metadata. This is an easy way to add additional training data for anyone wishing to build models without an image processing component, and should not take much extra time for the annotators to gather.
- If previous rounds of fact checking are mentioned in the article (e.g. this claim was previously investigated by *xyz*), annotators should mark the claim and (if provided) add a hyperlink. This may be possible to identify automatically simply by detecting hyperlinks to other fact checking sites.
- If the fact checking article is paywalled, annotators should report this. We will discard all such claims.

4.2.2 Claim Type

The type of the claim itself, independent of the approach taken by the fact checker to verify or refute it, should be chosen from the following list. Reading the fact checking article may still be necessary to resolve ambiguity. This is not a mutually exclusive choice – a claim can be speculation about a numerical fact, for example. As such, annotators should choose one *or several* from the list.

- **Speculative claims**: such as "the price of crude oil will rise next year." The primary task is to assess whether the prediction is plausible or realistic. This is beyond the scope of this project, and we will discard all such claims.
- **Opinion claims**: such as *"cannabis should be legalized"*. This contrasts with factual claims on the same topic, such as *"legalization of cannabis has helped reduce opioid deaths."* This type of claim belongs to an opinion that is not factual. We will discard all claims in this category.

- **Numerical claims**. The primary task is to verify whether a numerical fact is true, to verify whether a comparison between several numerical facts hold, or to determine whether a numerical trend or correlation is supported by the evidence.
- **Quote verification**. The primary task is to identify whether a quote was actually said by the supposed speaker.
- **Position statements**. The primary task is to identify whether a public figure has taken a certain position, e.g. supporting a particular policy or idea. For example, *"Edward Heath opposed privatisation"*. This also includes statements that opinions have changed, e.g. *"Edward Heath opposed privatisation before the election, but changed his mind after coming into office"*. Factual claims about the actions of people (e.g. *"Edward Heath nationalised Rolls-Royce"*) are not position statements; claims about the attitudes of people (e.g. *"Edward Heath supported the nationalisation of Rolls-Royce"*) are.
- **Event or property claims**. The primary task is to determine the veracity of a narrative about a particular event or series of events, or to identify whether a certain non-numerical property is true, e.g. a person attending a particular university.
- **Doctored media identification**. The primary task is to determine whether an image, video, or soundbite has been doctored. This also includes HTML-doctoring of social media posts. We will discard all claims in this category.
- **Complex media claims**. The primary task is to perform complex reasoning about pieces of media, distinct from doctoring. This could for example be geolocating an image, or analysing audio. An example is https://www.factcheck.org/2016/06/expert-voice-analyst-its-trump/, where the fact checkers used voice analysis to identify Trump as the speaker on a phone recording. The claim itself *directly requires* media analysis to check; e.g. "the speaker of these two clips is the same". We will discard all claims in this category. Claims where the original source is video, but which can be understood and verified without viewing the original source, do not fall under this category. An example of this is https://www.politifact.com/factchecks/2017/jul/27/gavin-newsom/inland-empire-second-only-san-francisco-california/, where Politifact has extracted the claim (*"The Inland Empire is 'the second fastest growing region' for jobs in California."*) from the video. This can cause confusion for annotators if we do not specify, as the original video/audio can feature on the PolitiFact page as metadata.

This list is, in part, inspired by the analysis presented at https://www.herox.com/factcheck/5-practise-claims.

4.2.3 Fact Checking Strategy

After identifying the claim type, we ask annotators to classify the approach taken by the fact checker. Again, one *or several* options should be chosen from the following list:

- **Written Evidence**. The fact checking process involved finding contradicting written evidence, e.g. a news article directly refuting the claim.
- **Numerical Comparison**. The fact checking process involved numerical comparisons, such as verifying that one number is greater than another.
- **Consultation**. The fact checkers directly reached out to relevant experts or people involved with the story, reporting new information from such sources as part of the fact checking

article.

- **Satirical source identification**. The fact checking process involved identifying the source of the claim as satire, e.g. The Onion. We will discard all claims that were refuted only through satirical source identification.
- **Image analysis**. The fact checking process involved image analysis, such as comparing two images.
- **Other media analysis**. The fact checking process involved analysing other media, such as video. For example, for the claim *"Trump family hosted a riot watch party to view the storming of the U.S. Capitol from a private tent"*, the fact checker found the time of the video shooting based on the background music. The time of video shooting was before the Capitol riot, and therefore the Trump family did not watch the riot as it happened after the party. We will discard all claims that are refuted only through other media analysis.

As an example of a case with overlapping strategies, the claim at https://www.politifact.com/factchecks/2017/aug/23/blog-posting/fake-news-rosa-parks-had-no-daughter-could-praise-/ was fact checked both by finding contradicting evidence *and* by identifying the source as satirical. Since the claim can also be verified without relying on the nature of the original source, we do not want to discard it.

4.3 Claim Normalization

Some claims in the DATACOMMONS dataset are not complete, which means they lack enough context to be verified. For example, in the claim *"Says Rep. Brian Fitzpatrick, R-Pa., 'stood with Trump and the lies... He stood with QAnon, not you."'*, there are unresolved pronouns without which the claim cannot be verified (e.g. *you* refers to the American voters). Another example is *"Israel already had 50% of its population vaccinated."* We need to know when this claim was made to verify its veracity, as the time is crucial for this verification.

We will rely primarily on metadata to resolve the ambiguity in these cases. However, this may not always be enough. In the claim normalization stage, annotators are asked to contextualize claims to the original post – beyond what metadata provides – by gathering the necessary information. We furthermore ask annotators to do some small housekeeping actions, such as rephrasing questions as claims. We include both the normalized and the original claim in the final dataset (as well as in subsequent annotation phases). We ask annotators to follow this checklist:

1. Is the claim referring to entities which can only be identified by reading the associated fact checking article, even if all metadata is taken into consideration? If so, add the name of the entity (e.g. *"Quotes former first lady as saying, 'White folks are what's wrong with America'."* becomes *"Quotes former first lady Michelle Obama as saying, 'White folks are what's wrong with America'."*).
2. Does the claim have unnecessary quotation marks, or references to a speaker (such as the word *says* in the example here)? If so, remove them (e.g. *"Says 'Monica Lewinsky Found Dead' in a burglary."* becomes *"Monica Lewinsky found dead in a burglary."*). Do NOT remove the reference to the speaker if the central problem is to determine if that person actually said the quote, e.g. in the case of quote verification.
3. Is the claim a question? If so, rephrase it as a statement (e.g. *"Did a Teamsters strike hinder aid efforts in Puerto Rico after Hurricane Maria?"* becomes *"A Teamsters strike hindered aid efforts in Puerto Rico after Hurricane Maria in 2017."*).

4. Does the claim contain pronominal references to entities only mentioned in the fact checking article? If so, replace the pronoun with the name of that entity.

5. For some examples in the dataset, the claim field contains the headling of the fact checking article, not the actual claim. In such cases, find the original claim in the article, and use that for producing the normalized version.

It is important that all rewrites should be done *after* reading the article and understanding what the central problem is. Furthermore, information should *only* be added to claims if it is absolutely necessary; if metadata is enough to resolve all ambiguity, there is no need to rewrite. Placing metadata collection first ensures this.

Important! Notice that annotators are not allowed to modify the terms used in the original claim being fact checked. For example, we have the claim "Our economy wouldn't reach pre-pandemic levels until 2025." Annotators should not replace "economy" with any other term, such as "real GDP", "potential GDP" – even if that is necessary to reach an unambiguous interpretation of the claim. This will be handled in the question generation and answering phase.

4.4 Ground Truth Production

We furthermore ask annotators to produce a ground truth from the claim, relying *only* on the information on the fact checking site (and assuming that everything reported there is accurate). This should reflect the decision of the fact checker, not the interpretation of the annotator; if the annotators disagree with the original judgment, they should err on the side of the fact checker.

We are using four categories – see Section 3.1. We have also discussed using a heuristic to map labels from the fact checking sites to these four categories as well, in order to do statistics on how close the judgments of phase 1 annotators and fact checkers are. We expect these to be close. As such, we can use the labels collected in phase 1 as a representation of how fact checking sites have validated the claims, and we can perform comparisons in subsequent phases. These comparisons are not the primary basis for discarding claims, as fact checkers use different evidence from the phase 2 & 3 annotators (and ultimately trained systems), but it is relevant data.

5 Phase 2: Question and Answer Generation

The next round of annotation aims to produce pairs of questions and answers providing evidence to verify the claim. The primary sources of evidence are the URLs linked in the fact checking article. We also provide access to a custom search bar, for the cases where those URLs do not yield appropriate information, as well as a reverse image search function. We intend to use the Google search API (or similar) to retrieve evidence – this grants access to a reverse image search API, which we can also use.

The annotator is first asked to read the claim and the associated fact checking article (including the verdict). Following this, the annotator is asked to generate the questions the answers to which are the evidence that allowed the fact checker to reach a verdict. We then ask the annotator to go through the following iterative process to verify or refute the claim:

1. Ask a question to collect evidence about the claim.

2. Attempt to answer the question. The answer can come from any of the sources listed on the fact checking article (e.g. any hyperlinks to other sites), from the internet using our search bar, or our reverse image search tool.

3. If the question was not answerable from any of the links on the fact checking article, or from searching, mark it as unanswerable. It may be useful to ask a rephrased version of the question, or a version with more context, as the next question.
4. If the claim is not verifiable from the retrieved answers, ask another question or rephrase the current question, proceeding from (1). Otherwise, or if more than n minutes have passed, indicate that you are ready to give a verdict for the claim.
5. Assign a verdict from the list in Section 3.1, then proceed to the next hit.

Many questions would be answerable from the information on the fact checking page. However, it may be the case that links are broken, or that information is missing. We should encourage annotators to try the hyperlinks on the fact checking page first, and then move on to searching if that fails.

Before proceeding to the next hit, the annotator will be shown a warning with the QA-pairs they have generated. They will also be shown their assigned label, along with the label from the fact checking site. They will be asked to check an ”are you sure?”-box, indicating that the collected evidence is sufficient to assign the label they have chosen to the claim.

Annotators should not choose a label if the retrieved evidence does not support it; for example, if the label **missing context** is chosen, there should be evidence documenting the missing context.

In the following two sections, we detail further how questions and answers should be constructed.

5.1 Question Generation

To ensure the quality of the generated questions, we ask the annotators to create their questions as follows:

- Questions should be well-formed, rather than search engine queries (e.g. “where is Cambridge?” rather than “Cambridge location”).
- Questions should be standalone and understandable without the context of the claim, or of any previous questions. All relevant details should be specified, such as dates, names of people and places, timeframes, etc.
- Questions should refer to at least one entity appearing in the claim, in metadata, or in other answers.
- The annotators should avoid any question that directly asks whether or not the claim holds, e.g. *“is it true that [claim]”*.
- The annotators should ask all questions necessary to gather the evidence needed for the verdict, including world knowledge that might seem obvious, but could depend for example on where one is from, e.g. Europeans will have better knowledge of European geography/history than Americans, and vice-versa.
- As a guiding principle, at least 2 questions should be asked. This is not a hard limit, however, and the annotators can proceed with only one question asked if they do not feel more are needed.

The following are examples used to illustrate how questions should be asked. These are based on the real claim *“the US in 2017 has the largest percentage of immigrants, almost tied now with the historical high as a percentage of immigrants living in this country”*:

- Good: What was the population of the US in 2017?
- Good: How many immigrants live in the US in 2017?
- Bad: What was the population of the US? [No time specified to find a statistic]
- Bad: What was the population there in 2017? [What does *there* refer to?]
- Bad: Is it true that the US in 2017 has the largest percentage of immigrants, almost tied now with the historical high as a percentage of immigrants living in this country? [Directly paraphrases the claim]

5.2 Value Judgments

As a part of the question generation process, annotators may have to make judgments about how to interpret the claim. For example, for the claim *"Shakira is Canadian"*, it may be necessary to choose what it means to be Canadian. This is expressed in how questions are formulated, e.g. *"does Shakira have Canadian citizenship?"* or *"where does Shakira live?"*. This may also involve politically charged judgments. For example, some First Nations people are classed as Canadian by the Canadian government, but do not use that label for themselves.

In such cases, we ask annotators to follow – as closely as possible – the judgments made by the fact checking websites. If the annotators feel that these are incomplete or misleading, they can add additional questions.

For example, for the claim *"Edward Heath opposed privatisation"*, a fact checker might provide his party manifesto as evidence. A corresponding question could then be *"what did the 1970 Conservative Party manifesto say about privatisation?"* An annotator could encounter evidence for (or have prior knowledge of) the nationalisation of Rolls Royce during Heath's government. In that case, the annotator might want to add an additional question, such as *"did Heath's government nationalise any companies?"*. The annotators should ask both questions.

5.3 Answer Generation

To find answers to questions, the annotators can rely on any sources linked from the factchecking site. Where these fail to produce appropriate information – either because they are not relevant to an asked question or because they refer to sources which have been taken down – we provide search functionalities as an alternative. Note that the annotators are not allowed to use the fact checking article itself as a source, only the *links* in the fact checking article (and only when they are not from fact checking websites).

In order to control what information annotators can find, we rely on custom search bars/search fields. As a backend, we use a commercial search engine such as the Google or Bing API. We give access to two information gathering actions: Text search, and reverse image search. The Google and the Bing API allow access to both of these. We filter the retrieved results as follows:

1. We discard any result dated after the publication of the original claim, as identified by the annotators in the normalization step (see Section 4).
2. We discard any result from a fact checking website. Some claims have gone through multiple cycles of publication and fact checking, and step 1 is not sufficient to filter these.
3. We furthermore discard any result from an article with a hyperlink to a fact checking site, to avoid press mentions of previous fact checks of claims.

Once an answer has been found, the annotators can choose between the following four options to enter it:

- **Extractive:** The answer can be copied directly from the source. We ask the annotators to use the copy-paste mechanism to enter it. At most one sentence, and at most 50 words, can be pasted.
- **Abstractive:** A freeform answer can be constructed based on the source, but it cannot be directly copy-pasted. The annotator should write the answer as a single sentence of at most 50 words.
- **Boolean:** This is a special case of abstractive answers, where a yes/no is sufficient to answer the question. A second box must be used to give an explanation for the verdict grounded in the source (e.g. "yes, because.."). The explanation must be a single sentence with at most 50 words.
- **Unanswerable:** No source can be found to answer the question. Unlike QABrief, we do not require annotators to explain why this is the case.

For extractive, abstractive, and boolean answers, the annotators are also asked to copy-paste a link to the source URL they used to answer the question. Extractive answers are preferred to abstractive and boolean answers.

We curate a list of known sources of misinformation. In the event that an annotator enters an answer with a source from a website on that list, that will be marked in the system. We do not prevent them from using the source, but we provide a message asking for additional evidence. This could be e.g. *"[Website] is on our list of known biased websites. Can you add similar evidence from an additional source, in order to further substantiate the evidence?"* Similarly, we curate a list of satirical websites, and mark those. Annotators can use this to either avoid such sources, or to show that the claim originates from a satirical site.

We furthermore prioritise easily read sources, e.g. websites, over pdfs. We use a system similar to biased sources – we do not reject information sourced from a pdf, but we ask the annotator if they can add additional information from a non-pdf source.

We may also wish to collect some statistics regarding what evidence was needed, especially when search was used. The search query will be an important piece of data to gather, in that case. That is, did the annotators need to do anything on the page where the answer originates from. This may be relevant especially for abstractive answers, if they are poorly phrased. We can do so with a checklist. Did the annotator read a pdf? Consult a table? Watch a video? Look at a graph? This can simply be checkboxes on the annotation page.

While answering a question, we furthermore ask annotators to adhere to the following:

Important!

- DO NOT use any other search bar to find an answer. You MUST use the provided search bar only.
- Do NOT answer the claim or predict a verdict at this stage, and do NOT update the answer to any previous question. Only answer the CURRENT question.
- DO NOT submit answers from fact checking websites, such as politifact.com or factcheck.org. These answers will not be accepted. If you use our provided search bar, you will not have this problem.

5.4 Persistence

Once evidence has been retrieved, accepted and used to provide an extractive/abstractive answer as discussed below by the annotator, we use archive.is to ensure persistence of the evidence page. We can then disseminate links to the archived versions of the evidence with the dataset. We do not place any restrictions on what data formats the annotator can choose as evidence – that includes pdfs, word documents, and multimodal data.

5.5 Reasoning Chain of Claims

Annotators can build up reasoning chain across multiple questions, meaning that answers of one question can be used in the next question. For example, the first question is "What is the fastest Japanese train?". The answer is "The fastest Japanese train is Shinkansen ALFA-X". Based on the answer, we can further ask the second question to get more details, "What is the maximum operating speed of the Shinkansen ALFA-X".

While entering data, annotators are not required to use placeholders for entities introduced by previous answer (e.g. Shinkansen ALFA-X in the above example). We may want to do so in the final dataset, e.g. to facilitate evaluation or training, but we do so posthoc.

6 Phase 3: Verdict Validation and Quality Control

Once we have collected evidence in the form of retrieved questions and answers, we want to provide a measure of quality. Given a claim with associated evidence, we ask a third round of annotators to give a verdict for the claim. Crucially, we do not give annotators at this round access to the original fact checking article, or to the claim label. We furthermore use this round to perform quality control on the question-answer pairs, similar to the QABriefs paper. We give instructions similar to the following:

1. Read the claim, the metadata, and the question-answer pairs. This is the only information which should be used during this phase. Do NOT use web search to find additional information, or rely on background knowledge which an average English speaker might not have.
2. Flag any answers that seem incorrect, e.g. if the answer to *"how many people live in California?"* is *"three billion"*. The QA pairs will be presented one by one.
3. After reviewing the kept QA pairs, assign a label to the claim (see the four labels introduced for phase 1 annotators). Any pairs marked incorrect will not be shown.
4. If you feel the presented sources did not allow you to reach an unbiased verdict for the claim, flag it. We will review any claims flagged as such.

We use this information to decide which claims to keep, and which to discard. Malformed QA pairs should be discarded, but the claims can be kept if there is sufficient evidence otherwise. We can potentially replace or repair these QA-pairs, and reannotate the claims.

If the phase 3 annotators reach a verdict different from that obtained during phase 2, another annotator can review the claim. That annotator should have access to the fact checking article. Such claims can then either be discarded, re-annotated, or the phase 3 label accepted.

In addition to the verdict, we ask annotators in Phase 3 to write a short statement justifying their verdict. This should be at most 2 sentences, and at most 100 words. There should not be any new information presented in this statement, it should just describe how the annotators used the

information present in the claim, the metadata, and the QA-pairs to reach their verdict. These justifications can be used by systems to learn to provide explanations along with verdicts.

7 Postprocessing

After all annotation steps, we will carry out several small steps to increase the quality of the data. This section documents any such changes.

7.1 Quality Check for Sources

There is a small chance that some annotators will retrieve answers from sources that contain biased, unreliable, or otherwise harmful information – as an example, an annotator might source an answer to *Infowars*. We are marking these during the annotation process. If manageable, we may also wish to go over the list of used sources manually to ensure we do not miss anything. Similarly, we can manually go over claims which phase 3 annotators have flagged as relying on biased or otherwise problematic sources.

7.2 Previously Checked Claims

During phase 1, we collect claims which have been checked multiple times. In those cases, it is possible that both fact checking articles are in our dataset. To identify this, we will compare the retrieved URLs from previously checked claims to all URLs in the dataset, searching for similarities. We will also compare all original claim URLs, in order to find cases where the same claim occurs as a fact checking article from different institutions. When we eventually construct splits for our dataset, we will ensure that all such cases are placed in the training split in order to prevent data leakage. Potentially, we could also merge these claims, so that we get a single claim with more evidence question-answer pairs.

7.3 Label Bias in Question Generation

A potential concern is that annotators may generate different questions depending on whether they believe the claim to be true or false. Since systems ultimately must learn to generate their own questions (rather than relying on annotated questions), we do not believe this to be a large issue – however, it is something to be aware of. It is a way for bias about the state of the world to enter into training (e.g. certain questions are asked about certain topics on the basis of bias).

The use of validation in phase 3 will hopefully filter out some of these cases. To measure how serious a problem this poses afterwards, we will along with other baselines produce a ”gold questions only”-system. That is, a system which uses the claim and the gold questions, but not the answers, to reach a verdict. We will include this along with the eventual dataset publication.

8 Previous discussion points

To keep a bit more structure, let us move items from Section 1 that we already discussed here (rather than commenting them out) so we can refer to them if needed.

8.1 A concern regarding leakage

Something I realised might be a concern is that the label leaks through the form of the questions annotators ask, similarly to how the surface form of claims in FEVER-style annotators can give

away the game (see e.g. Derczynski et al. (2020)). For example, annotators might ask certain types of questions more often when substantiating true claims than when refuting false claims.

I am not sure how large a concern this is. It is rather hard to explicitly design the annotation process to avoid it, I think – phase 2 annotators necessarily know the validity of the claim because they read the fact checking article. Thoughts?

With that said, once we reach the stage of building systems, we can detect such issues by using a ”gold questions only”-baseline. That is, a system which uses the claim and the questions, but not the answers. I think we should definitely remember to run such a system and include it when the time comes to publish.

AV: True. But I think this is less of an issue here. The surface form of the claim (and the metadata in our case) could give away the answer and they shouldn't, so we should definitely check for these. However, the questions are something the systems will have to generate and won't be given to them, so I would worry less about it. Having said that, it would be interesting to see the bias in the human question generation process; I can definitely see myself asking different questions when I believe a claim is false, and this bias is not great, same questions should be asked for the same kind of claim. In a way, phase 3 tries to avoid the word part of this by asking someone to validate the question-answer-based verdict, but still this would only avoid some obvious cases.

8.2 ”No answer explanation” in QABriefs

I (Michael) went through the first 1000 examples from the QABriefs, classifying explanations for cases where no answer where given. My findings:

- 42 questions where appropriate information was simply not found in time through the search engine.
- 13 questions that were too vague or malformed for the annotator to even start searching.
- 5 questions for which the search engine did not return any hits with the keywords the annotator used.
- 3 questions that misunderstand the subject to such a degree that the question is nonsense; e.g. asking what effect the ACA had on Obamacare when the ACA *is* Obamacare.
- 2 questions where the annotator wrote in personal experience in the ”no answer explanation” box rather than search for an answer.
- 2 questions with unresolvable pronominal references to entities in the claim.
- 1 question where the annotator could not find evidence from *reliable* sources (although the annotator mentions that they did find an answer from some random person's blog).
- 1 question seemingly asking for the annotator's opinion on a subject.
- 1 question where the annotator could not figure out how to enter the absence of information as evidence; e.g. showing that a company did not donate to a certain politician by showing that the company do not appear on the politician's list of donors.
- 1 question where the information is contradictory. The question asks for a politician's stance on a subject. The politician took one stance while campaigning, then switched to the other stance after being elected.

For our purposes I think we should figure out how to deal with ”absence of information” as evidence. This is also related to the infamous Shakira example – if we cannot find information

supporting her being Canadian, but also cannot find evidence directly opposing it, how do phase 2 annotators show that using QA pairs? Contradictory evidence we actually already deal with by having the fourth label. With that sorted I don't think we need the annotators to explain why they chose no answer, the other cases are not very informative. And given the two who used the "no answer explanation"-box to enter personal experience for answering a question, maybe we are better off without it?

Phase 2 is where annotators interpret statements... use exmaple

8.3 Explicit questions for complex calculations

In the pilot, some numerical claims required complex calculations on the basis of retrieved evidence to verify (for example, calculating a percentage increase). We briefly discussed making this explicit. We agreed that it should not be, as models should learn to perform this reasoning on their own. However, I thought of another option that we might want to consider – we could include them explicitly as questions (e.g. what is the percentage increase from X to Y), and mark them as a special answer type? Not *extractive* or *abstractive*, but *calculation*. That way people using the dataset (or we, once we see how often this occurs but before we release it) can choose whether to include these cases or not; in some cases they might actually be helpful for teaching models the reasoning process.

AV: It is definitely something to discuss. My intuition has been to leave any (complex) calculation for the verdict stage. I was thinking that any answers obtained should be trivial for a human to assess by looking at a source, and hopefully possible to assess automatically with some word overlap measure. I.e. even if it is an abstractive answer, we should be able to check it automatically. My thinking is that I anticipate evidence finding (formulated as question answering) to be the hardest part of the annotation (using search engine over the web for this is something I feel less certain about), so it would be great to be able to check it automatically.

Actually this leads me to revise what I said earlier about Boolean questions: While we should allow for them, the answers should be more than "yes"/"no" so that we can assess them. One idea would be to have a Yes/No followed by an extractive/abstractive answer which should contain parts of the source webpage.

Justifications in phase 3 (?)

8.4 Add boolean answer type

A question that came up – answers are currently either abstractive or extractive. Do we want to add an extra category for yes/no, or just use abstractive for such answers?

AV: I say yes. In fact we should be more precise about extractive vs abstractive. Do we mean at the sentence level? I would suggest so. Which means that one can copy paste a sentence but not a whole paragraph. If bits of the whole paragraph are needed, then best to summarize it in a sentence. This implies that questions should be answerable with a sentence roughly speaking, at least those where some sort of free text is expected.

Enforce 1 sentence limit in addition to 50 words

8.5 Prioritize easily accessible sources?

We did not have time to discuss this last Monday – should we ask annotators to prioritise easily accessible sources, if they can find multiple? E.g. use a web page rather than a pdf, if one is

available? If so, how do we concretely do this without making search take longer?

AV: Yes, we should. I think google search kind of does this to an extent (and in a way going far down the search results would be undesirable for both humans and machines). I would definitely prioritise the sources linked from the fact checking article. Assuming that these are not pdfs (perhaps we could do a quick check on the link the annotator gives us), then we ask them to search google. I think we can tell the API to omit some kinds of results. If we find that we can get enough claims annotated this way, then we might decide not to have anything that is not a webpage.

Same/similar approach to biased sources (e.g. give me another source please)

8.6 Split NEI into two labels in phase 3?

I realised there are two distinct cases where annotators in phase 3 may use NEI. (1) if there is evidence that the claim cannot be verified or that the appropriate information cannot be retrieved, and (2) if the phase 2 reviewer has not managed to verify the fact through their asked questions. The latter could happen e.g. because they ran out of time. Do we want to use separate annotations for these? We could mark (1) as true NEI in the dataset, and use a second pass (either by us or by another annotator) to verify (2). Then, (2) could either be discarded (because the questions were not appropriate) or kept with another label.

Keep as one

8.7 Time limit for questions

We should discuss the time limit annotators have in phase 2 to answer questions. Should it be there, and if so how long?

8.8 Rephrasing for comparisons

We have discussed rephrasing claims involving (some) comparisons to instead be about percentage errors (see Vlachos and Riedel (2015)). How exactly should annotators do this? When should they rephrase, what percentages should we suggest?

This is only for models producing evidence, not for the dataset.

8.9 Answer Search

In QABrief and MultiFC, the search engine will block contents from fact checking websites when they collecting evidence. For Snopes, they directly use the sentences in the summary section of the fact checking article as evidence. I think there are three ways to search for the answer.

- **Use a custom search engine**: similar to QABrief, annotators will have to search information through the internet without using any contents from the fact checking website.
- **Get the answers from article**: similar to Snopes but relax a little bit, annotators can just extract or rewrite the sentences from the full fact checking article rather than only from the summary section.
- **Get the answers from references**: based on my hand annotations, many evidence is hard to retrieve from the Internet without prior knowledge. For example, most claims on Politifact involves politics and economics. One will need related knowledge to find evidence from specific websites, such as reports, laws, bills and rules from the government or organizations. On the other hand, some evidence especially related to social media is no longer

exist. Under this circumstances, getting answers from the references of the fact check articles might be a solution. First, fact checkers have provided evidence that hard to retrieve from the internet, second evidence from the article is arxiv, which means they kept the copy for a certain time.

How do we present this in a way that encourages using web search over the article if possible, but allows the option in case web search fails?

We discussed two options:

1. During the initial normalisation round, annotators also flag cases where the fact checking site does NOT contain the appropriate information to support their conclusion (or perhaps better, give a guess for the verdict without seeing it). During the final verdict validation phase, the annotators then do not see the fact checking site at all; their conclusion is based only on retrieved evidence. Cases where evidence from the fact checking article is necessary to validate the claim are then exactly those which could be validated during initial normalisation, but not during verdict validation.

2. During verdict validation, we first do not grant access to the text from the fact checking site. We place a button on the site which the annotators can press if they feel the retrieved evidence is not sufficient to make a judgment. Potentially, they have to give their best guess before pressing. When pressed, they are then presented with the text from the fact checking article, which can then be used as further evidence.

AV: I think the main source of answers should be the references of the fact checking article. If this is not enough, (could be validated in a second round as suggested) then a search engine with a block list (I think this is possible in some APIs, but I guess we can hack it too to ensure that the results shown to the annotators exclude known FC websites). Actually the idea of asking them to provide verdict given just the evidence annotated/retrieved can also be used to filter claims, as I mentioned above, so it might help us with many problems.

8.10 Question and Answer Generation Process

In the document as previously written, QG and QA are two separate steps. However, they will be performed by the same annotator. From my (Michael) experience annotating claims, there are situations where completely separating the process (e.g. first generating all the questions, then all the answers) is not ideal. Instead, we could use an iterative process where the annotator generates and answers one question, then another, and so on. The situations I encountered were:

1. Some claims require multihop reasoning, as mentioned also by Rami. With separate phases we would need placeholders for entities retrieved by initial questions.

2. With some claims, strictly following the reasoning of the fact checking website fails, but the claim can still be verified with information from the internet. This can be because evidence is not available online, or has been removed since the initial fact checking. I found trial-and-error an easy way to find alternative sources to verify claims. This may be outside the scope of what we can ask annotators to do – worth discussing, though. With separate phases, annotators cannot use trial-and-error to find the best questions.

3. For one claim, I needed to add extra context to my initial question in order to get useful results on Google. *"What did the pope say in 2017 about migrants?"* was not specific enough; *"what did the pope say in 2017 about migrants and satefy concerns?"* retrieved the necessary evidence.

Are we committed to a two-step process, or is an iterative approach (e.g. ask question, answer question, choose whether to keep QA pair, repeat) more suitable? What is the reasoning behind the current two-step structure?

AV: I think we should go with the iterative, single step approach, probably involiving multiple annotators per claim, for the reason you mentioned. The two step process was used in the fact checking briefs, but that was because we wanted to see what lay people thought was worth asking. Here what we really want is to reverse-engineer the pro fact checker.

Zhijiang and I discussed this, and we think an iterative approach is best. That way the annotators can refine their claims, and possibly get around the cases where the original evidence is not available/has been taken off the internet. We should also:

1. Encourage annotators to formulate their questions based on the fact checking site, to give them an idea of how to easiest find evidence. **AV**: Agreed.
2. Limit the number of questions that can be asked (including revisions), maybe by limiting the number of times the search bar can be used. This means annotators do not get stuck on claims where google does not retrieve appropriate evidence. **AV**: Not sure about this though. I think we should limit the amount of time spent on a claim instead, which will limit the questions.

This does break down the distinction between ”questions” and ”search queries”, which might actually be a problem. Having the two phases separate encourages well-formed questions (e.g. ’what is the average yearly temperature in Cambridge’ rather than ’average temp Cambridge’).

AV: Good point. In fact translating the questions into queries is that something we should (learn to) tackle.

Also, do we give the annotators for this stage access to the truth ratings assigned by the fact checking websites?

AV: I thought of using this as a validation process; if the verdict of our annotator disagrees with that in the website, this might be indicative of poor work by the annotator, but also highly ambiguous/controversial claim, which we might want to ignore. However, I see the point in the comments, and also the way the article is written often gives away the verdict. So I would say that the validation step involves someone looking at the questions and answers only, returns the verdict (if they think there is conclusive enough evidence) and then we compare it to the rating from the website.

I also thought of a hypothetical that we should probably consider how to deal with before it occurs – what do we do if some annotators retrieve and use unreliable, biased, or otherwise harmful evidence documents? E.g. what do we do if an annotator answers a question with an extractive fragment from a site like infowars?

We discussed that a simple check for which URLs annotators used might be enough – if we see 1-2 datapoints with evidence from unreliable sites, we can discard those.

AV: Yep, for these we could also use known lists of websites with poor/extremely biased content, e.g. https://en.wikipedia.org/wiki/List_of_fake_news_websites

8.11 Claim Normalization

For example, we have the claim “A CBO report found that USA economy wouldn’t reach pre-pandemic levels until 2025.” The team “economy” inside the claim is not well-defined. For such

claims, we might have two solutions. The first one is that we can design questions that can give the claim reasonable concrete interpretations so that it can be checked against evidence. The second one is to rewrite some of the concepts in the claim during the normalization stage. For example, we can change the "economy" as "real GDP".

Another example is "The (Erie County, N.Y.) health commissioner makes more than the governor, the vice president of the United States, and comparable to Dr. Anthony Fauci." The word "comparable" may be vague, but we can substitute it into "less than". Then the claim is much more clear.

8.12 Rewrite Multi-Modal Claims

After discussing with Rami, we have some thoughts about the multi-modal claims. Due to the difficulty of including images or videos in the dataset, we can rewrite the original claim based on the content of the images or videos.

For example, we have "Video of Sikh farmers attacking a bus in Punjab.", which is also incomplete as it missed the time stamp. We can rewrite the claim to be "In order to protest against the new farm laws in Delhi, Sikh farmers attacked a PRTC bus in Punjab in February 2021." Therefore, we can use the claim solely based on text instead of using claim with a video.

However, the fact checking procedure provided by the fact checkers should be adjusted accordingly, I am not sure how to deal with it at this moment.

For such claims, fact checker usually used reversed image search for the screenshot to find the original video from reliable news websites. Then check the texts along with the video in the news article. By imitating the process, we can use several questions to build the reasoning chain.

The first question can be "What is the original source of this video?". The annotators can provide an abstractive answer by using reversed image search or references in the fact check websites. The answer can be "The video originally from an article titled The horse touched a bus published by Asianet News Hindi website on 22 September 2019." Based on the first answer, we can further ask "What is event described in the article titled The horse touched a bus published by Asianet News Hindi website on 22 September 2019." The answer can be abstractive or extractive, "In the article, it is reported that a group of Nihang Sikh men attacked the PRTC bus traveling from Nakodara to Kapurthala. This incident reportedly took place near the Sundan Bridge after the PRTC bus driver accidentally hit one of the horses ridden by the Nihang Sikh men." Finally, the last question can be "Is ther Nihang Sikh men attacked the PRTC bus related to the ongoing protest against the new farm laws?"

8.13 Difference Between QABrief

The main difference is that the evidence retrieved through the questions will allow for a verdict to be deduced without further information. QABriefs doesn't do this, the questions and answers help, but they are not sufficient for a verdict to be deduced.

8.14 Evidence formats

After looking through a few claims, I (Michael) have found evidence to come in several different forms. Often, it is impossible to verify the claim within the first 10 pages of google without the ability to parse the proper source. I have seen:

- Raw text. This is the most frequent form, and also the easiest for us to parse.

- Structured data. Tables and lists are also relatively common, especially when discussing official statistics.
- One claim required *either* reading a table *or* extracting a value from a picture of a trend line. We probably do not want/need to tackle this "graph reading" problem at this time?
- Several of the claims only had appropriate evidence retrieved by google in the form of PDFs or Word documents. Word documents we cannot do much about. We may want to keep our eyes open for a PDF to text parser.

8.15 Label Standardization

Different fact checking websites have different labels, I think we should standardize the labels in the claim assessment and normalization stage. Do we use three classes classification similar to FEVER (refuted, supported, not enough information) or adding more classes, such as mixture?

AV: yes. I think the three way classification is the way to go, but with evidence for the NEI option; even if one can't reach a verdict, one must have evidence to say this, as we do in the new dataset.

8.16 Multi-Modal Claims

8.16.1 Fact Check Procedure

For natural claims, we have non-textual ones. such as "Video of Sikh farmers attacking a bus in Punjab." This type of claim includes information from more than texts, such as videos, images, etc,. It is not verifiable solely based on the claim itself.

Actually, we have a considerable of multi-modal claims. Especially, for natural claims in Indian and Filipino fact checking websites. Here are the statistics of claims in DATACOMMONS dataset: 26% of claims belong to multi-modal claims. 16% of claims are related to images and 10% of claims are related to videos. For claims from Indian fact checking websites, 64% of claims are multi-modal. For Filipino fact checking websites, 60% of claims are multi-modal.

For claims related to images, fact checkers mainly relied on reverse image search to collect evidence.

- **Image Search**: This method leverages reverse image search for the image along with the claim. After finding the photo from reliable news websites, the fact checker can give the verdict based on the comparison between description in the news and the sentence in the claim. For example, we have the claim "Photo shows Democratic protesters who "stormed the U.S. Capitol in 2018." Fact checkers used the photo along with claim to search its source. Then we can find that the photo is originated from the news article related to protest in Milwaukee. Therefore, we can give the verdict False here. The image along with the claim is not doctored.

Statistics of Image Search : For reverse image search, I think Google's Vision API.7 is a potential choice, which is also used by previous work to fact check image related claims from Snopes (Zlatkova et al., 2019).

To find out if it is possible to use the image reverse search, I tested on 7 instances from the DATACOMMONS dataset (the number of samples is not enough). All of images related to the claim are from Facebook posts. Existing claims with images can be categorized into 2 types. The

first one is True image with False claim, where the image is real from other reports. The second type is False image with False claim, where the image is doctored intentionally. At previous discussion, Andreas mentioned that the second type is not in the scope of our project.

For images without doctoring (5 out of 7), I am able to find all of their original sources based on the reverse image search. It is much easier than checking the claims with videos (see stats in the next subsection).

As for doctored images (2 out of 7), I can also find lots of related results based the reverse search. However, basically all of them are from the fact checking websites.

For claims related to videos: one relied on the description of the video while the other required to understand the content of the video.

- **Screenshot Search**: This method leverages reversed image search for the screenshot of the video. After finding the video from reliable news websites. Fact checker can give the verdict based on the comparison between description in the news and the sentence in the claim. For example, we have the claim "Video of Sikh farmers attacking a PRTC bus in Punjab" Fact checkers used the search shot in the video to search its source. Then we can find that the photo is originated from the news article related to a protest back in 2018 rather the current one. Therefore, we can give the verdict False here.
- **Keywords Search**: This method uses keywords along with the photos to search. For example, we have the claim "Video of a Japanese train traveling 515 km in 10 minutes at a maximum speed of 4800 kmph" as shown in Figure 2. Fact checkers use two methods to verify the claim. The first method finds the original video at its description, which denotes that the video is a fiction. The other method uses texts with the video and extract keywords to search. For example, the post with the video also claim that it only took 10 minute to travel from Osaka to Tokyo by using this train. Based on the Google map, it took 2 hours to travel from Osaka to Tokyo, so the verdict is False. **For the second method, they did not rely on the video completely, they just fact checked "From Osaka to Tokyo 515km takes 10 minutes." I guess both methods (finding the source of the video and fact check the texts along with the video) are sufficient to verify that the claim is not true.**

Statistics of Screenshot Search : To find out if it is possible to use the screenshot reverse search, I tested on 5 instances from the DATACOMMONS dataset (the number of samples is not enough). 4 out of 5 videos are from Facebook and 1 video is from Youtube. I can only find the original source of 1 video. For the other four videos, two of them are deleted and the Indian fact checking website does not provide an archive version of the post as Politifact did. The original post of the other one is not provided in the fact checking article. For the other two that we can access the original post, I can find one based on the screenshot reverse search. It is non-trivial to do so, as it requires you to watch the video and fine the most distinguishable part. For example, we have the claim "Video of Uttar Pradesh government attacking protesting women farmers with a water tanker." The video lasts around 3 minutes. I can only find the original video from the news website by using the screenshot of the water tank. I also tried other screenshots, including women were crushed by the truck, etc,. but I could not find the original video based on such screenshots. For the other video, I adopted similar strategy and used several screenshots to search, however, I could not find any results except those fact checking websites.

It seems to me that it is extremely hard to use screenshots of the video to find the evidence based on the reverse image search. The first issue is we do not know how to find the correct screenshot to search, as most videos are from Facebook, where automatically chosen screenshots are not

Figure 2: Example claim related to video.

provided. The one from Youtube is deleted as it violated the policy. Another issue is language, many claims related to videos are from Indian fact checking websites. Many reverse image search results are not in English, but in Hindi (though we can use google translate).

I believe we have as such decided to use reverse image search for images, but not for video. Do we then want to filter out entirely video-form claims, or do we just not give any special tool to find evidence?

AV: As you both mention in the comments, we can keep videos and doctored images as long as the evidence doesn't require tools beyond reverse image search. I mostly the dataset we are creating as a dev/test set. Thus evidence to evaluate the models beyond just finding the verdict is necessary for any claim to be there, however this evidence doesn't need to be useful for training.

8.17 Claim types

Part of the metadata collection in phase 1 involves identifying the type of the claim. What are the options to include? These are potential types we have discussed:

- "Vanilla" fact verification.
- Quote verification. Identify whether a quote was actually said by the supposed speaker.
- Doctored image identification. Identify whether an image has been doctored. This is not within the scope of our dataset, and such claims will be filtered out.

Also add "excluded" options

Challenges from FEVEROUS annotation, resolving ambiguity https://www.herox.com/factcheck/5-practise-claims

References

Derczynski, L., J. Binau, and H. Schulte (2020, July). Maintaining quality in FEVER annotation. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Online, pp. 42–46. Association for Computational Linguistics.

Vlachos, A. and S. Riedel (2015, September). Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2596–2601. Association for Computational Linguistics.

Zlatkova, D., P. Nakov, and I. Koychev (2019). Fact-checking meets fauxtography: Verifying claims about images. *ArXiv abs/1908.11722*.