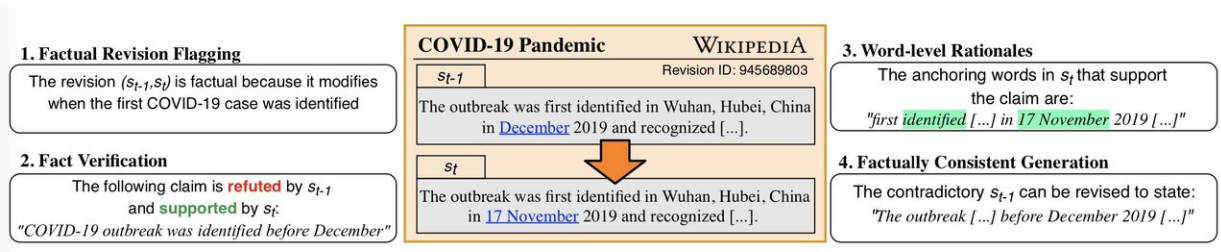
Auto-updating websites when facts change

March 30 2021



Credit: MIT Computer Science & Artificial Intelligence Lab

Many companies put millions of dollars towards content moderation and curbing fake news. But what about the old news and misinformation that is still out there?

One fundamental truth about the internet is that it has lots of outdated information. Just think about the many news articles written in the early weeks of the COVID-19 pandemic, before we knew more about how the virus was transmitted. That information is still out there, and the most we can do to minimize its impact is to bury it in search results or offer warnings that the content is old (as Facebook now does when users are about to share a story that's over three months old.)

The story becomes even more complicated when dealing with deep learning models. These models are often trained on billions of webpages, books, and news articles. This can help the AI models to catch up with

what's second nature to us humans, like grammatical rules and some world knowledge. However, this process can also result in undesirable outcomes, like amplifying social biases from the data that the models were trained on. Similarly, these models can also stick to some old facts that they memorized at the time they were created but were later on changed or proved to be false—for example, the effectiveness of certain treatments against COVID-19.

In a new paper to be presented at the NAACL Conference on Computational Linguistics in June, researchers from MIT describe tools to tackle these problems. They aim to reduce the amount of wrong or out-of-date information online and also create deep learning models that dynamically adjust to recent changes.

"We hope both humans and machines will benefit from the models we created," says lead author Tal Schuster, a Ph.D. student in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). "We can monitor updates to articles, identify significant changes, and suggest edits to other related articles. Importantly, when articles are updated, our automatic fact verification models are sensitive to such edits and update their predictions accordingly."

The last part—ensuring that the latest information is followed—is specific to machines in this project. Encouraging also humans to have a flexible mindset and update their beliefs in the presence of new evidence was beyond the scope here. Though, boosting the editing process of old articles can already at least reduce the amount of old information online.

Schuster wrote the paper with Ph.D. student Adam Fisch and their academic advisor Regina Barzilay, the Delta Electronics Professor of Electrical Engineering and Computer Science and a professor in CSAIL.

Studying factual changes from Wikipedia revisions

In order to examine how new information is being incorporated in articles, the team has decided to examine edits to popular English Wikipedia pages. Even with its open design, allowing anyone to make edits, its massive and active community helped Wikipedia become a safe place with reliable content—especially for newly developed situations like a pandemic.

Most of the edits in Wikipedia, however, do not add or update new information but only make stylistic modifications, for example, reordering sentences, paraphrasing, or correcting typos. Identifying the edits that express a factual change is important because it can help the community flag these revisions and examine them more carefully.

"Automating this task isn't easy," says Schuster. "But manually checking each revision is impractical as there are more than six thousand edits every hour."

The team has collected an initial set of about two hundred million revisions to popular pages like COVID-19 or famous figures. Using deep learning models, they ranked all cases by how likely they are to express a factual change. The top three hundred thousand revisions were then given to annotators that confirmed about a third of them as including a factual difference. The obtained annotations can be used to fully automate a similar process in the future.

To complete this manual annotation process, the team reached out to TransPerfect DataForce. In addition to filtering the significant revisions, annotators were also asked to write a short plausible claim that was correct before the revision but is not true anymore.

"Achieving consistent high-quality results on this volume required a well-orchestrated effort," says Alex Poulis, DataForce's creator and senior director. "We established a group of 70 annotators and industry-grade

training and quality assurance processes, and we used our advanced annotation tools to maximize efficiency."

This process resulted in a large collection of revisions, paired with claims that their truthfulness changes over time. The team named this dataset Vitamin C as they find its unique contrastive nature to improve the robustness of AI systems. Next, they turned to develop a number of AI models that can simulate similar edits and be sensitive to them.

They also publicly shared Vitamin C to allow other researchers to extend their studies.

Automating content moderation

A single event can be relevant to many different articles. For example, take the FDA's emergency approval for the first mRNA vaccine. This event led to edits not only in the mRNA page on Wikipedia but to hundreds of articles on COVID-19 and the pandemic, including ones about other vaccines. In this case copy-pasting is not sufficient. At each article, the information should be added at the relevant location, maintaining the coherence of the text, and possibly removing old contradicting details (for example, removing statements like "no vaccine is available yet").

Similar trends could be seen in news websites. Many news providers create dynamic webpages that update from time to time, especially about evolving events like elections or disasters. Automating parts of this process could be highly useful and prevent delays.

The MIT team decided to focus on solving two related tasks. First, they create a model to imitate the filtering task of the human annotators and can detect almost 85 percent of revisions that represent a factual change. Then, they also develop a model to automatically revise texts, potentially

suggesting edits to other articles that should also be updated. Their text revising model is based on sequence-to-sequence Transformer technology and trained to follow the examples collected for the Vitamin C dataset. In their experiments, they find human readers to rate the model's outputs the same as the edits written by humans.

Automatically creating a concise and accurate edit is difficult to do. In addition to their own model, the researchers also tried using the GPT-3 language model that was trained on billions of texts but without the contrastive structure of Vitamin C. While it generates coherent sentences, one known issue is that it can hallucinate and add unsupported facts. For example, when asked to process an edit reporting the number of confirmed COVID-19 cases in Germany, GPT-3 added to the sentences that there were 20 reported deaths, even though the source, in this case, doesn't mention any deaths.

Luckily, this inconsistency in GPT-3's output was correctly identified by the researchers' other creation: a robust fact verification model.

Making fact verification systems follow recent updates

Recent improvements in deep learning, have allowed the development of automatic models for fact verification. Such models, like the ones created for the FEVER challenge, should process a given claim against external evidence and determine its truth.

The MIT researchers found that current systems are not always sensitive to changes in the world. For around 60 percent of the claims, systems were not modifying their verdict even when presented with the opposite evidence. For example, the system might remember that the city of Beaverton Oregon had eighty thousand residents and say that the claim

"More than 90K people live in Beaverton" is false, even when the population of the city eventually grows above this number.

Once again, the Vitamin C dataset comes in handy here. Following its many examples of facts that change with time, the MIT team trained the fact verification systems to follow the currently observed evidence.

"Simulating a dynamic environment enforces the model to avoid any static beliefs," says Schuster. "Instead of teaching the model that the population of a certain city is this and this, we teach it to read the current sentence from Wikipedia and find the answer that it needs."

Next, the team is planning to expand their models to new domains and to support languages other than English. They hope that the Vitamin C dataset and their models will also encourage other researchers and developers to build robust AI systems that adhere to the facts.

More information: Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. arXiv:2103.08541v1 [cs.CL] 15 Mar 2021, arxiv.org/abs/2103.08541

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by MIT Computer Science & Artificial Intelligence Lab