Automated system can rewrite outdated sentences in Wikipedia articles

February 12 2020, by Rob Matheson

MIT researchers have created an automated text-generating system that pinpoints and replaces specific information in relevant Wikipedia sentences, while keeping the language similar to how humans write and edit. Credit: Christine Daniloff, MIT

A system created by MIT researchers could be used to automatically

update factual inconsistencies in Wikipedia articles, reducing time and effort spent by human editors who now do the task manually.

Wikipedia comprises millions of articles that are in constant need of edits to reflect new information. That can involve article expansions, major rewrites, or more routine modifications such as updating numbers, dates, names, and locations. Currently, humans across the globe volunteer their time to make these edits.

In a paper being presented at the AAAI Conference on Artificial Intelligence, the researchers describe a text-generating system that pinpoints and replaces specific information in relevant Wikipedia sentences, while keeping the language similar to how humans write and edit.

The idea is that humans would type into an interface an unstructured sentence with updated information, without needing to worry about style or grammar. The system would then search Wikipedia, locate the appropriate page and outdated sentence, and rewrite it in a humanlike fashion. In the future, the researchers say, there's potential to build a fully automated system that identifies and uses the latest information from around the web to produce rewritten sentences in corresponding Wikipedia articles that reflect updated information.

"There are so many updates constantly needed to Wikipedia articles. It would be beneficial to automatically modify exact portions of the articles, with little to no human intervention," says Darsh Shah, a Ph.D. student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and one of the lead authors. "Instead of hundreds of people working on modifying each Wikipedia article, then you'll only need a few, because the model is helping or doing it automatically. That offers dramatic improvements in efficiency."

Many other bots exist that make automatic Wikipedia edits. Typically, those work on mitigating vandalism or dropping some narrowly defined information into predefined templates, Shah says. The researchers' model, he says, solves a harder artificial intelligence problem: Given a new piece of unstructured information, the model automatically modifies the sentence in a humanlike fashion. "The other [bot] tasks are more rule-based, while this is a task requiring reasoning over contradictory parts in two sentences and generating a coherent piece of text," he says.

The system can be used for other text-generating applications as well, says co-lead author and CSAIL graduate student Tal Schuster. In their paper, the researchers also used it to automatically synthesize sentences in a popular fact-checking dataset that helped reduce bias, without manually collecting additional data. "This way, the performance improves for automatic fact-verification models that train on the dataset for, say, fake news detection," Schuster says.

Shah and Schuster worked on the paper with their academic advisor Regina Barzilay, the Delta Electronics Professor of Electrical Engineering and Computer Science and a professor in CSAIL.

Neutrality masking and fusing

Behind the system is a fair bit of text-generating ingenuity in identifying contradictory information between, and then fusing together, two separate sentences. It takes as input an "outdated" sentence from a Wikipedia article, plus a separate "claim" sentence that contains the updated and conflicting information. The system must automatically delete and keep specific words in the outdated sentence, based on information in the claim, to update facts but maintain style and grammar. That's an easy task for humans, but a novel one in machine learning.

For example, say there's a required update to this sentence (in bold): "Fund A considers 28 of their 42 minority stakeholdings in operationally active companies to be of particular significance to the group." The claim sentence with updated information may read: "Fund A considers 23 of 43 minority stakeholdings significant." The system would locate the relevant Wikipedia text for "Fund A," based on the claim. It then automatically strips out the outdated numbers (28 and 42) and replaces them with the new numbers (23 and 43), while keeping the sentence exactly the same and grammatically correct. (In their work, the researchers ran the system on a dataset of specific Wikipedia sentences, not on all Wikipedia pages.)

The system was trained on a popular dataset that contains pairs of sentences, in which one sentence is a claim and the other is a relevant Wikipedia sentence. Each pair is labeled in one of three ways: "agree," meaning the sentences contain matching factual information; "disagree," meaning they contain contradictory information; or "neutral," where there's not enough information for either label. The system must make all disagreeing pairs agree, by modifying the outdated sentence to match the claim. That requires using two separate models to produce the desired output.

The first model is a fact-checking classifier—pretrained to label each sentence pair as "agree," "disagree," or "neutral"—that focuses on disagreeing pairs. Running in conjunction with the classifier is a custom "neutrality masker" module that identifies which words in the outdated sentence contradict the claim. The module removes the minimal number of words required to "maximize neutrality"—meaning the pair can be labeled as neutral. That's the starting point: While the sentences don't agree, they no longer contain obviously contradictory information. The module creates a binary "mask" over the outdated sentence, where a 0 gets placed over words that most likely require deleting, while a 1 goes on top of keepers.

After masking, a novel two-encoder-decoder framework is used to generate the final output sentence. This model learns compressed representations of the claim and the outdated sentence. Working in conjunction, the two encoder-decoders fuse the dissimilar words from the claim, by sliding them into the spots left vacant by the deleted words (the ones covered with 0s) in the outdated sentence.

In one test, the model scored higher than all traditional methods, using a technique called "SARI" that measures how well machines delete, add, and keep words compared to the way humans modify sentences. They used a dataset with manually edited Wikipedia sentences, which the model hadn't seen before. Compared to several traditional text-generating methods, the new model was more accurate in making factual updates and its output more closely resembled human writing. In another test, crowdsourced humans scored the model (on a scale of 1 to 5) based on how well its output sentences contained factual updates and matched human grammar. The model achieved average scores of 4 in factual updates and 3.85 in matching grammar.

Removing bias

The study also showed that the system can be used to augment datasets to eliminate bias when training detectors of "fake news," a form of propaganda containing disinformation created to mislead readers in order to generate website views or steer public opinion. Some of these detectors train on datasets of agree-disagree sentence pairs to "learn" to verify a claim by matching it to given evidence.

In these pairs, the claim will either match certain information with a supporting "evidence" sentence from Wikipedia (agree) or it will be modified by humans to include information contradictory to the evidence sentence (disagree). The models are trained to flag claims with refuting evidence as "false," which can be used to help identify fake

news.

Unfortunately, such datasets currently come with unintended biases, Shah says: "During training, models use some language of the human written claims as "give-away" phrases to mark them as false, without relying much on the corresponding evidence sentence. This reduces the model's accuracy when evaluating real-world examples, as it does not perform fact-checking."

The researchers used the same deletion and fusion techniques from their Wikipedia project to balance the disagree-agree pairs in the dataset and help mitigate the bias. For some "disagree" pairs, they used the modified sentence's false information to regenerate a fake "evidence" supporting sentence. Some of the give-away phrases then exist in both the "agree" and "disagree" sentences, which forces models to analyze more features. Using their augmented dataset, the researchers reduced the error rate of a popular fake-news detector by 13 percent.

"If you have a bias in your dataset, and you're fooling your model into just looking at one sentence in a disagree pair to make predictions, your model will not survive the real world," Shah says. "We make models look at both sentences in all agree-disagree pairs."

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 Massachusetts Institute of Technology