How AI can help journalists find diverse and original sources

December 19 2023, by Leila Okahata

Credit: Unsplash/CC0 Public Domain

What would news stories be without proper sources? To tell a compelling story, reporters need to find newsworthy narratives and trustworthy information. Such information typically comes from a wide

pool of publications, official records and experts, all with their own biases, expertise, opinions and backgrounds. The pool of interview candidates is plentiful yet overwhelming to navigate.

Artificial intelligence, however, may serve as a guide.

Researchers from the USC Information Sciences Institute are creating a source-recommendation engine designed to suggest references for journalists. "In practice, the software application would analyze a given text or topic and suggest relevant sources by cross-referencing against a database of potential interviewees, experts or informational resources," said Emilio Ferrara, a professor of computer science and communication at the USC Viterbi School of Engineering. "The tool could provide contact details, areas of expertise and previous work of the sources," he added.

The tool's development is being led by Alexander Spangher, a computer science Ph.D. student at USC Viterbi who previously worked as a data scientist at the New York Times. While immersed in the journalism industry, Spangher witnessed the pressure of traditional newsrooms. "I haven't spoken to a single local journalist that was not totally overstretched," he remarked. "There have been news deserts and papers shutting down. It's areas like this that we really want to assist and build tools for."

Motivated to provide helpful resources for reporters, Spangher is creating various AI gadgets, including a source-recommendation system prefaced in his paper, "Identifying Informational Sources in News Articles," that was accepted to the 2023 Conference on Empirical Methods in Natural Language Processing and is now posted to the *arXiv* preprint server.

To create an AI model that can suggest sources, the researchers first laid

the groundwork: how are human journalists currently using sources in news writing? To study this, they gathered a dataset of sentences from over a thousand news articles and annotated the source of the information, as well as the sourcing category (e.g., "direct quotes," "indirect quotes," "published works" and "court proceedings").

A thousand annotated news articles, however, were not enough data for the researchers to draw firm conclusions about all the myriad ways journalists use sources across reporting genres. But, it was enough to train a language model (LM) to continue the annotation process. "Language models are AI frameworks that process and understand [human language](#) by analyzing large volumes of text for patterns and context," explained Ferrara, senior author of the paper.

The LMs the researchers trained could detect source attributions with 83% accuracy, revealed the authors. Now equipped with these LMs, they annotated roughly 10,000 news articles and delved further into understanding the compositionality of news writing: when and how do journalists currently use sources?

The AI models found that, on average, roughly half the information in news articles came from sources and, in each article, there are usually one to two major sources (i.e., those contribute 20% or more of the information in the article) and two to eight minor ones (those that contribute less). "The AI also discovered that the first and last sentences were the most likely to be sourced," Spangher explained, adding that reporters often lead with cited information and close with a quote to send the reader off.

The researchers challenged their new algorithm with one more test: could they detect if a source was missing? If AI can recognize when information is lacking, then it can be configured to know when to recommend a particular expert to complete the full picture.

Analyzing 40,000 articles with some sources randomly removed, the AI models easily noticed when a major source was absent but had difficulties with minor ones. Although they may be the least crucial to a story, less obvious sources may also be the most valuable recommendations that an AI could one day make, Spangher said.

"You're going to draw a lot of information from the main participants, but supplementary voices are going to provide extra color and details to the article," he noted. "It's going to be a challenge to get the engine to recognize and recommend minor sources, but they may be the most helpful."

The researchers also think the tool will be significant if it can diversely recommend sources. "It can introduce journalists to new, diverse voices beyond their usual network, thus reducing the reliance on familiar sources and potentially bringing in fresh perspectives," Ferrara said.

However, every AI system is prone to bias if not appropriately designed, he added. "To ensure diversity in source databases, standards should include representation from a wide range of demographics, disciplines and perspectives," he noted.

Jonathan May, a research associate professor of computer science at USC Viterbi and ISI lead researcher, imagines a future where the sourcing engine jumpstarts the reporting process, allowing journalists to be more efficient.

"Technology that can help us do creative work and be our creative best is a good thing," said May, a co-author of the paper. "That's why I'm hopeful for it."

The team plans to collaborate with journalists to gather feedback for further improvements.

"With projects like this, I really thrive off talking to journalists and understanding their needs, viewpoints and what they think will or won't work," Spangher said. "Any solution to local journalism will require a bunch of different people with a bunch of different backgrounds coming together."

Provided by University of Southern California