Designing natural language processing tools for teachers

October 26 2023, by Allison Whitten

Credit: Unsplash/CC0 Public Domain

According to a 2022 Gallup poll, U.S. teachers in grades K-12 are the most burnt-out workers in any profession. Could natural language processing tools (NLP) support them?

While today's chatbots are so far helping teachers to increase their efficiency on tasks like quickly generating lesson plans or writing emails to parents, two Stanford scholars are studying options that could go beyond shaving a few minutes off the workday.

Stanford Graduate School of Education professor Dora Demszky and computer science Ph.D. student Rose Wang are creating models that provide teachers with the kind of rare feedback and suggestions that will help them make the most impact on a student's learning and well-being in the classroom.

Demszky and Wang emphasize that every tool they design keeps teachers in the loop—never replacing them with an AI model. That's because even with the rapid improvements in NLP systems, they believe the importance of the human relationship within education will never change.

"Such a fundamental aspect of learning is the human connection between the student and the teacher, and the motivation to learn is triggered by that relationship. It's not really possible to build that with a robot," said Demszky. "We should never circumvent the teacher but think about the ways in which we could augment the teacher's work."

Experts in the loop

Building classroom technology requires extensive background knowledge of pedagogy and student learning techniques that only experienced teachers have gained.

As a result, Demszky and Wang begin each of their NLP education projects with the same approach. They always start with the teachers themselves, bringing them into a rich back and forth collaboration. They interview educators about what tools would be most helpful to them in

the first place and then follow up with them continuously to ask for feedback as they design and test their ideas. "We couldn't do our research without consulting the teachers and their expertise," said Demszky.

Sometimes, teacher feedback took the research in a new direction. For one preprint project posted to *arXiv*, developing a large language model (LLM) to tutor math students step by step in an online chat-based format, the scholars originally assumed that math teachers first pick a strategy to help the student understand.

But after interviewing math teachers, they learned that a teacher's first step is to try to pinpoint exactly where the student's misconception is coming from. "We would have never been able to actually get to that detail if we hadn't been able to talk to teachers that can share their own math teaching experiences," said Wang. "The devil is really in the details."

When they asked students to rate the feedback generated by LLMs and teachers, the math teachers were always rated higher. However, when they re-prompted the LLM with help from the teachers—who labeled the type of student mistake and offered a specific strategy to use—the LLM responses were rated much higher. While still not considered as valuable as a teacher, the LLMs rated more highly than a layperson tutor.

"It indicates that there's a lot of promise in using these models in combination with some expert input, and only minimal input is needed to create scalable and high-quality instruction," said Demszky.

In an additional preprint paper posted on June 23, they studied math at the college level using online courses from the MIT OpenCourseWare YouTube channel. In these projects, they examined whether LLMs could provide feedback to online instructors on when they lose students during

a lecture, based on analyzing online student comments during the discussion.

Here, they created SIGHT, a large dataset of lecture transcripts with linked student comments, and trained an LLM to categorize the comments into categories like confusion, clarification, and gratitude. Additionally, they are working on developing and publishing a framework called Backtracing, which is a task that prompts LLMs to retrieve the specific text that caused the most confusion in a student's comment.

"We're trying to incentivize these online lecture creators to be able to revise their content in a much more targeted manner," said Wang.

Meet your new AI coach

Another promising direction that Demszky and Wang have been working on is an NLP system that could act as a teacher's aide to observe an in-person lesson and offer suggestions to improve. Demszky sees this option almost like a "nonjudgmental coach" that could tailor its suggestions while teachers are still new to the profession and then continue providing in-depth advice even as those teachers become more seasoned.

In a paper published this June at ACL's Workshop on Innovative Use of NLP for Building Educational Applications, the team tested ChatGPT as one possible coaching tool. They found 82% of the model's suggestions were ideas teachers were already doing, but the tool improved with more tailored prompts.

In a new paper posted to *arXiv*, which will be presented at the Conference on Empirical Methods in Natural Language Processing in December, they trained a model on "growth mindset" language. Growth

mindset is the idea that a student's skills can grow over time and are not fixed, a concept that research shows can improve student outcomes.

When they prompted GPT-4 to reframe a teacher's comments into growth mindset language, 174 students and 1,006 students rated the model's reframings as being 24% to 85% better (depending on the task) than teachers in its use of growth mindset language.

For example, one teacher responded to a student's question on fractions by saying, "Erase what you have please. Everything. And I want you to look at it again, okay? Everything. Everything. Everything." The LLM recommended this language instead: "Thank you for sharing your thoughts; it's great that you're actively participating! Let's erase the fraction and work together to understand how to locate fractions on the number line. We'll figure this out as a team and continue improving."

Demszky and Wang are currently working with David Yeager at the University of Texas at Austin, who offers annual trainings for teachers on growth mindset strategies. They're aiming to develop an LLM teacher coaching tool that Yeager and others could soon deploy as part of these workshops.

A vision for the future

Thus far, Demszky and Wang have focused on building and evaluating NLP systems to help with one teaching aspect at a time. But the two envision a future where many NLP tools are used together in an integrated platform, avoiding "tech fatigue" with too many tools bombarding teachers at once.

But Demszky and Wang worry this future may exacerbate inequity in schools. "What I'm seeing at the moment, at least, is more just that the rich get richer," said Demszky. She worries that children in more

privileged settings might get access to both high-quality teaching and AI teaching support, while children in underserved settings may eventually get access to AI without high-quality teaching.

Wang adds that it will be just as important for AI researchers to make sure that their focus is always prioritizing the tools that have the best chance at supporting teachers and students.

"It's not about thinking, how do we raise the ceiling of educational tech, but how do we raise the floor of education with these tools?" Wang said.

More information: Rose E. Wang et al, Step-by-Step Remediation of Students' Mathematical Mistakes, *arXiv* (2023). DOI: 10.48550/arxiv.2310.10648

Rose E. Wang et al, SIGHT: A Large Annotated Dataset on Student Insights Gathered from Higher Education Transcripts, *arXiv* (2023). DOI: 10.48550/arxiv.2306.09343

Rose E. Wang et al, Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction, *arXiv* (2023). DOI: 10.48550/arxiv.2306.03090

Kunal Handa et al, "Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms, *arXiv* (2023). DOI: 10.48550/arxiv.2310.10637

Provided by Stanford University