DORTY WORDS

ENGONEERONG A LOTERARY CLEANUP

In an interdisciplinary effort, Northwestern computer scientists use machine learning to correct glitches in digitized historic texts.

For years, Martin Mueller has been plagued by ugly black dots. A professor emeritus of English and classics at Northwestern's Weinberg College of Arts and Sciences, Mueller studies and enjoys early English texts. Over the past decade, unsightly black dots have crept into the words of the digitized versions of some of his favorite books, giving them what Mueller calls a serious "yuck factor."

"I don't like dirty things," he says. "These texts are cultural heritage objects, and they should be clean. We don't have much of the past left. We want to keep it right."

WHENCE COMETH THE BLACK DOT?

In 1999, several universities and libraries established the Text Creation Partnership (TCP) to digitize English books published before 1700. By transcribing texts and putting them online, the TCP developed a searchable, navigable, free database for students, scholars, and readers everywhere. Transcribed by non-English speakers, these digital versions were not copied from the original books but from digital scans of microfilms—some of which were more than 60 years old.

"You can imagine that with a process like that, a lot of things can go wrong," Mueller says. "And a lot of things did."

The resulting 50,000 transcribed texts have roughly five million incomplete words. Many of the aged books were browned and splotchy from the start, and their legibility was further compromised by poor quality scans. If transcribers could not read or understand a portion of the text—for example, if it was cut off at the margins, obscured by previous owners' handwritten notes, or included arcane abbreviations and spellings—they replaced the unknown character with a black dot. Thus the word "love," for example, became "lo●e."

To solve the problem, modern readers could arguably comb through the texts and fix all the errors, but Mueller estimated it could take several minutes for a human to fix just one error. To tackle all of the errors, it would take one person years of nonstop work—an impractical, if not humanly impossible, task.

HENCE LIT MEETS COMPSCI

Northwestern Engineering's Doug Downey embraced Mueller's dilemma as a challenge. Four years ago, then associate dean Stephen Carr connected Downey, an associate professor of computer science, with Mueller. An expert in statistical language modeling, Downey proposed using machine learning to help correct the texts.

Language modeling is most popularly known for its role in autocorrect and voice recognition programs. It assigns probabilities to sequences of words to estimate the likelihood of what word is missing or coming next. For example, the word "lo●e" might be "love," but it also might be "lone," "lore," or "lose." A language model evaluates the context of the word in question—comparing it to known contexts where the candidate words have been used—to choose the correct option. If the context is "she was in lo●e with him," then the program assumes the missing word is, indeed, "love."

Downey's group first trained the language model on 363 relatively clean texts from the same era. Once the computer program understood common semantics, it was ready to work through a sample of 359 flawed texts, which included plays, textbooks, court transcripts, treatises, biblical commentary, romance novels, and more.

The program finds "blackdot words" and spelling errors, and evaluates 35 characters to the left and right of each character in question. It then submits zero to three replacement possibilities, assigning a probability to each option based on the context. This past summer, Weinberg undergraduates combed through the options to select the correct one.

"The project is a lot more involved than it originally appears," says Katie Poland, a junior studying English and philosophy. "We're just working with single words, but there can be so much hidden in there."

In many cases, the undergraduate students did not just deal with blackdot words but also non-standardized spellings—"France" and "color" were spelled "Fraunce" and "coulloure." Other times, the transcribers might have mistaken old letters for modern letters. The Old English letter "s," for example, looks very similar to a modern-day "f."

SPREADING THE RESULTS

While the Weinberg students work to solve the language riddles, Northwestern Engineering students are working on an interactive website for the project. They have built a site where humanities scholars can search for words in different texts and fix errors right on the spot. Super users then either accept or reject the corrections. Accepted fixes are automatically updated into the system.

"Machine learners can also learn from that feedback," Downey says. "A little bit of crowdsourcing like that could go a long way. Eventually we could have super high-quality transcriptions."

Yue "Hayley" Hu, a sophomore studying computer science, has been working on the site's database. Although she previously had little exposure to early English texts, she has enjoyed discussing them with the English students. "It's neat that we can offer a computer science-based solution."

The collaboration's initial results indicate that approximately three-quarters of the incompletely or incorrectly transcribed works can be definitively corrected with a combination of machine learning and machine-assisted editing—without the need to consult the original printed text. This could drastically reduce the human-time cost from minutes to a mere seconds per word.

Corrections are automatically reintroduced and reintegrated into the texts, which form a larger and more accurate set of training data used for fixing blackdot and misspelled words in other texts. Eventually, Downey posits, the machines will be confident enough to require hand verification for just small text samples.

"It's energizing that we can apply research in a way that will impact a real community of users," Downey says. "It's rare to have a customer of your basic research where you can deliver something today that makes a difference."

AMANDA MORRIS