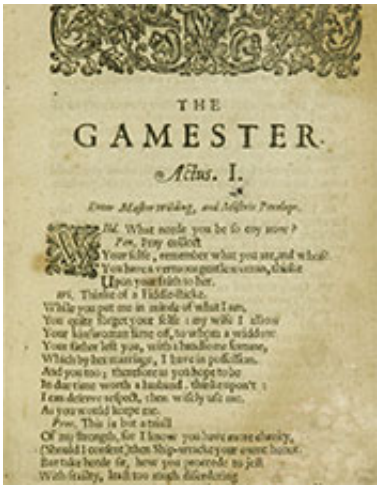
Engineering students fix common glitch in digitization of books published before 1700

December 28 2015


THE
GAMESTER.
Actus. I.

Computer scientists use machine learning to correct glitches.

Digitizing books published before 1700 has created an aesthetic as well as quite pragmatic "black-dot problem" in translated texts, with the word "love," for example, showing up as "lo•e."

Taking the digital savvy of today's age one step farther, Northwestern University engineering students in the McCormick School of Engineering and Applied Sciences have come to the rescue of the marred and sometimes indecipherable words that populate the translated versions of the early English texts.

Working in conjunction with undergraduates from the Weinberg College

of Arts and Sciences, the engineering students designed a computer program that uses language modeling, akin to autocorrect and voice-recognition programs, to help fill in the blanks of the incomplete words.

The dots creep into the process because of the difficulties of translating aged texts that often are browned, splotchy and cut off at the margins. When translators could not read or understand a portion of a text, they replaced an unknown character with a black dot.

Since 1999, about 50,000 texts have been transcribed by the non-profit Text Creation Partnership, but the works have roughly 5 million incomplete words. The translations of the tattered books also were further compromised by poor-quality scans.

Language modeling finds misspellings and "blackdot words" created when the computer encounters an unknown character. Once an error is found, nearby characters are evaluated and replacement suggestions are made, with a probability assigned to each option based on the context.

The word "lo•e" might be "love," but it also might be "lone," "lore," or "lose." A language model uses context 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."

Last summer, Weinberg students worked on the language riddles by combing through the options and selecting the correct one. Engineering students, meanwhile, have built a site where humanities scholars can search for words in different texts and fix errors on the spot. Super users then either accept or reject the corrections.

"Machine learners can also learn from that feedback," said project leader Doug Downey, associate professor of electrical engineering and computer science at the McCormick School of Engineering. "A little bit

of crowdsourcing like that could go a long way. Eventually, we could have super high-quality transcriptions."

Modern readers could arguably comb through the texts and fix all the errors, but it could take several minutes for a human to fix just one error, said Martin Mueller, professor emeritus of English and classics at Northwestern. To tackle all of the errors, it would take one person years of non-stop work—an impractical, if not humanly impossible, task.

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 seconds per word.

Provided by Northwestern University