Crowdsourced Transcription of Handwritten Mental Health Records

UNMIL P. KARADKAR

Abstract:
Problem statement: While optical character recognition (OCR) technology is routinely used to automatically retrieve text from print document images, this technology is insufficient for recognizing handwritten, especially cursive text, where words and lines often overlap. In order to transcribe such documents (for example, Korean war letters) National Archives among other institutions, have developed transcription interfaces. However, this approach exposes the document content to an unknown human transcriber and thus, is unsuitable for transcribing privacy-sensitive records, such as (historical) medical records. As a part of the Central State Hospital Digital Archives & Library Project, we are developing technologies and workflows that will enable digital access to privacy-sensitive records.

Methodology: Using our privacy requirements as guidelines, we have created a crowdsourcing workflow for public transcription of handwritten content. We have developed a Captcha-like mechanism that segments documents at the word-level, thus removing the textual context around words. The workflow uses open source infrastructure such as MongoDB document store, Zooniverse transcription engine, and Python and C++ scripting.

Results & Findings: In a preliminary evaluation, the segmentation algorithm successfully retrieved between 84% and 92% of the words in a document. However, the algorithm also generated segments that includes artifacts in the documents (due to spots and lines on paper, errant pen markings, and digitization artifacts like text from the other side of the page). The work to improve the algorithms is ongoing. We are also evaluating the workflow in terms of its usefulness and effectiveness.

About the Author:

Unmil P. Karadkar studies how data shapes and is shaped by human practices. His research contributes to areas such as Digital Humanities, Human Data Interaction, and Visualization. As an Assistant Professor in the School of Information at The University of Texas at Austin, Unmil teaches courses in Digital Libraries, automatic techniques for metadata generation. His research has been supported by the National Science Foundation, Andrew W. Mellon Foundation, and Texas General Land Office. He has published in leading digital libraries and conferences and journals and serves on the editorial board of the International Journal of Digital Libraries. Unmil holds a Ph.D. in Computer Science from Texas A&M University.