A STUDY OF TEXT SIMPLIFICATION ON BREAST CANCER INFORMATION TARGETING A LOW HEALTH LITERACY POPULATION

Ginger Dragon, Josefina Guzman, Xiwei Wang* and Francisco Iacobelli*

Adequate health literacy is important for maintaining good health, managing disease and self-advocation. Internet searching is one of the most efficient tools for patients, yet the written grade level of medical information is higher than many patients' level of literacy. Consequently, patients with low health literacy are less likely to understand online health information to treat their health issues and make better informed decisions. Although there are some automatic text simplification tools, they work off of assumptions of what simplified text should be and readability scores, which may not be reliable indicators of simplification. The purpose of this study is to understand what makes a text simple in the domain of breast cancer. We used machine learning techniques to train an automatic classifier that can identify if a text is simplified or not. This study was conducted using two coders that freely simplified texts to identify the strategies used. They then created a code-book of strategies to simplify text and trained to obtain a satisfactory inter rater reliability score. Then, they simplified 100 texts about breast cancer. Next, the 100 simplified and 100 original (complex) texts were converted into a set of numeric attributes obtained from Coh-Metrix –a tool that automatically analyzes text and returns 107 surface features of text, such as length, number of syllables, overlap between paragraphs, etc. A logistic-regression based classifier obtained an accuracy of 88% with a solid ROC curve indicating robustness of the classifier. The coefficients of the regression highlighted the importance of five attributes: word polysemy, standard deviation of sentence length, number of syllables, and word frequency in CELEX. As expected, simpler text uses a vocabulary with less syllables, less precise words (increased polysemy), words that are more frequent in regular speech, and sentences of less uniform length. Moreover, word polysemy alone can accurately separate 82.6% of texts into simple and complex. In the future, we would like to test our findings on low literacy populations to validate our work, as well as design algorithms to manipulate these surface features thus, automating the simplification of texts.