

Student Research Symposium 2003

Kevyn Collins Thompson

Predicting Reading Difficulty of WebPages With Statistical Language Models

A potentially useful feature of information retrieval systems for students is the ability to identify documents that are not only relevant to the query, but also a good match for the student's reading level. Manually obtaining an estimate of reading difficulty for each document is not feasible for very large collections, so we require an automated technique. Traditional readability measures such as Flesch-Kincaid perform very poorly on Web pages and other non-traditional documents, because of unreliable sentence length estimates and other factors. Recasting the well-studied problem of readability in terms of text categorization, we describe a new method based on simple statistical language modeling techniques. We show that by using a mixture model to interpolate evidence of a word's frequency across grades, it is possible to build a classifier achieving good performance across a range of individual grade levels. In addition, the classifier is not specific to any subject area and can be built using relatively little training data.