SO SÁNH CÁC VĂN BẢN TIẾNG VIỆT THEO ĐỘ KHÓ

***Lương An Vinh*[1], *Nguyễn Thị Như Điệp*[2], *Đinh Điền*[3]**
[1,3]Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
[2]Bộ môn Ngôn ngữ học, Trường Đại học Khoa học Xã hội và Nhân văn, ĐHQG-HCM
anvinhluong@gmail.com, nhudiep2004@gmail.com, ddien@fit.hcmus.edu.vn

Tóm tắt

Độ khó của văn bản là chỉ số xác định văn bản dễ hay khó đọc ở mức nào. Độ khó của văn bản đóng vai trò vô cùng quan trọng trong việc soạn thảo, phát hành và lựa chọn sách, đặc biệt là trong lĩnh vực giáo dục. Nghiên cứu về độ khó của văn bản đã được quan tâm từ lâu nhưng chủ yếu là cho tiếng Anh và một số ngôn ngữ phổ biến khác. Trong bài báo này, chúng tôi trình bày một phương pháp so sánh độ khó của các văn bản tiếng Việt với nhau sử dụng bộ phân lớp SVM. Bộ ngữ liệu được sử dụng là các tác phẩm văn học Việt Nam được đánh giá độ khó tương quan với nhau thông qua một số người đọc. Phương pháp này không đòi hỏi quá nhiều chi phí để xây dựng bộ ngữ liệu huấn luyện nhưng cũng đạt được độ chính xác xấp xỉ 80%. Đây cũng là tiền đề cho việc so sánh và lựa chọn các văn bản sao cho phù hợp với trình độ đọc của người đọc.

Từ khóa: Độ khó của văn bản, So sánh văn bản, Tiếng Việt

COMPARING VIETNAMESE TEXTS BY READABILITY

***Luong An Vinh*[1], *Nguyen Thi Nhu Diep*[2], *Dien Dinh*[3]**
[1,3]Faculty of Information Technology, University of Science, VNU-HCM
[2]Department of Linguistics, University of Social Sciences & Humanities, VNU-HCM
anvinhluong@gmail.com, nhudiep2004@gmail.com, ddien@fit.hcmus.edu.vn

Abstract

Readability is a concept that describes the degree to which a text is easy or difficult to read. It has an important role in text drafting, publishing and document selecting, especially in education. Research on text readability has long been concerned but mainly for English and some other popular languages. In this paper, we present a method of comparing the readability of Vietnamese texts using SVM classifier. The corpus we used for experiment is Vietnamese literary texts which are evaluated for their relative readability by some readers. This method does not require too much effort to build training corpus but also achieves approximately 80% accuracy. This is also a prerequisite for the comparison and selection of text to fit the reading level of the reader.

Key words: Text Readability, Text comparing, Vietnamese