Tạp chí Khoa học Đại học Huế: Khoa học Xã hội và Nhân văn
ISSN 2588-1213
Tập 130, Số 6A, 2021, *Tr. 123–130*; **DOI:** 10.26459/hueunijssh.v130i6A.5744

MÔ HÌNH SO SÁNH CÁC VĂN BẢN TIẾNG VIỆT THEO ĐỘ KHÓ

Nguyễn Sơn[1], Hồ Viết Hoàng[1], Đinh Điền[2], Lương An Vinh[2], Nguyễn Thị Như Điệp[3]
[1] Trường Đại học Ngoại ngữ, Đại học Huế, 57 Nguyễn Khoa Chiêm, Huế, Việt Nam
[2]Trường Đại học Khoa học Tự Nhiên, ĐHQG - HCM
[3]Trường Đại học Khoa học Xã hội và Nhân văn, ĐHQG - HCM

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 bằng máy tính, 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

1. Giới thiệu

Theo định nghĩa của Bailin & Grafstein, *độ khó của văn bản* là độ đo xác định xem một văn bản là dễ hay khó đọc như thế nào. Độ khó của văn bản phụ thuộc nhiều vào các yếu tố ngôn ngữ như cách dùng từ, ngữ, câu, phong cách của văn bản… Độ khó của một văn bản có ảnh hưởng rất lớn trong quá trình đọc và hiểu văn bản đó. Dựa trên độ khó của văn bản, độc giả có thể xác định xem một văn bản có phù hợp với khả năng đọc của họ hay không. Soạn giả cũng có thể sử dụng độ khó của văn bản đang soạn thảo để định hình đối tượng độc giả hoặc có những điều chỉnh cần thiết để văn bản của họ phù hợp hơn với đối tượng người đọc đang hướng đến [1].

Xây dựng một mô hình đánh giá độ khó của văn bản có ý nghĩa rất lớn trong khoa học và thực tiễn: hỗ trợ các nhà khoa học viết các báo cáo dễ đọc hiểu hơn; hỗ trợ các nhà làm giáo dục

**Liên hệ:* viethoang.vnh@gmail.com
Nhận bài: 27-03-2020; Hoàn thành phản biện: 23-11-2020; Ngày nhận đăng: 10-12-2020

soạn thảo các giáo trình và tài liệu phù hợp với từng lứa tuổi học sinh; hỗ trợ các nhà xuất bản trong việc định hình đối tượng độc giả; hỗ trợ chính phủ khi soạn thảo các văn bản quy phạm pháp luật cho phù hợp hơn với trình độ đọc hiểu của công dân; hoặc hỗ trợ các nhà sản xuất trong việc chuẩn bị các tài liệu hướng dẫn sử dụng các sản phẩm của họ… Ngoài ra, mô hình xác định độ khó của văn bản còn có thể hỗ trợ rất hiệu quả cho quá trình lựa chọn học liệu giảng dạy tiếng Việt cho người nước ngoài.

Các nghiên cứu về độ khó của văn bản đã được thực hiện từ những năm đầu thế kỉ XX, hầu hết các nghiên cứu đó được thực hiện trên tiếng Anh và một số ngôn ngữ phổ biến khác như các công thức của Dale-Chall [2], Flesch [4], Flesch-Kincaid [6], SMOG [8], các công trình của Si và Callan [12], Schwarm và Ostendorf [11], Heilman và cộng sự [5], Tanaka-Ishii và cộng sự [13], Vajjala và Meurers [14]…

Trong tiếng Việt, không có nhiều nghiên cứu về độ khó của văn bản. Một số nghiên cứu đã được thực hiện từ những năm 1985 của nhóm Nguyễn và Henkin [9], [10] cho người Việt ở hải ngoại. Năm 2017, khi khảo sát các đặc trưng của văn bản trong sách giáo khoa ngữ văn, nhóm tác giả Lương và các cộng sự đã chỉ ra yếu tố độ dài văn bản có ảnh hưởng lớn đến việc phân loại các văn bản ngữ văn này theo độ khó [7]. Tuy nhiên, nguồn ngữ liệu mà nhóm Lương và các cộng sự khảo sát giới hạn trong các văn bản sách giáo khoa phổ thông với độ khó đã có sẵn theo từng cấp lớp học. Trong thực tế, nguồn ngữ liệu văn bản bên ngoài là rất phong phú và cũng không có những ràng buộc về thời gian đọc như trong giáo dục phổ thông, nên vẫn cần có những giải pháp khác để đánh giá độ khó của các văn bản này.

Trong bài báo này, chúng tôi trình bày một phương pháp so sánh độ khó của các văn bản tiếng Việt với nhau. Những văn bản ban đầu không cần có độ khó thuộc lớp nào mà chỉ cần độ khó tương quan với một số văn bản khác. Theo đó, phần còn lại của bài báo được trình bày như sau: phần 2 trình bày các đặc trưng của văn bản mà chúng tôi sử dụng; phần 3 trình bày thực nghiệm so sánh độ khó của các văn bản dựa trên những đặc trưng chúng tôi đã nêu ra; phần 4 là các kết luận của nghiên cứu này.

2. Đặc trưng sử dụng

Chúng tôi mô tả những đặc trưng được sử dụng để so sánh độ khó của các văn bản. Các đặc trưng được lựa chọn hầu hết là các thống kê về hình thái, tần suất và ngữ pháp ở mức từ. Những đặc trưng này có thể được rút trích một cách tự động thông qua các công cụ máy tính để có thể tự động hóa quá trình so sánh văn bản. Với các đặc trưng mức cao hơn như ngữ pháp mức câu, ngữ nghĩa… mặc dù có thể có ảnh hưởng lớn tới độ khó của văn bản, nhưng do các công cụ xử lý tự động cho tiếng Việt vẫn còn nhiều hạn chế khi rút trích các đặc trưng này nên chúng tôi không sử dụng trong bài báo này.

Độ dài trung bình của câu: Độ dài trung bình câu của một văn bản là một trong những đặc trưng dễ thống kê và phổ biến nhất khi đánh giá độ khó của văn bản. Trong bài báo này, chúng tôi sử dụng độ dài trung bình câu tính theo từ và theo tiếng.

Độ dài trung bình từ: Trong tiếng Việt, một từ có thể là từ đơn tiết (chỉ có 1 âm tiết) hoặc từ đa tiết (có từ 2 âm tiết trở lên). Trong bài báo này, độ dài trung bình từ mà chúng tôi sử dụng được tính theo số âm tiết.

Tỉ lệ từ khó trong văn bản: trong rất nhiều nghiên cứu, tỉ lệ các từ khó trong văn bản là đặc trưng rất quan trọng khi đánh giá độ khó của văn bản. Tuy nhiên, việc xây dựng một danh sách từ khó là rất tốn kém, đòi hỏi sự khảo sát trên một lượng lớn người đọc với một lượng lớn từ, do đó hầu hết các tác giả đều sử dụng danh sách tần số từ để thay thế cho danh sách từ khó với ý tưởng: nếu một từ mà có tần suất sử dụng cao thì nhiều khả năng từ đó là từ dễ và ngược lại. Trong bài báo này, chúng tôi sử dụng danh sách 3.000 từ phổ biến nhất trong tiếng Việt được trích xuất từ các thống kê của nhóm tác giả Dien Dinh và cộng sự [3] đã công bố vào năm 2018. Theo nhóm tác giả Dien Dinh và cộng sự, 3.000 từ phổ biến này đã chiếm tới gần 90% số lượng từ thường được sử dụng trong văn bản tiếng Việt. Trong nghiên cứu này, những từ xuất hiện trong danh sách 3.000 từ này được xem như là những từ dễ, ngược lại, những từ không xuất hiện trong danh sách này được xem như là những từ khó.

Tương tự, không chỉ có từ khó, chúng tôi cũng lấy **Tỉ lệ các âm tiết khó** làm đặc trưng trong nghiên cứu này. Chúng tôi sử dụng danh sách 3.000 âm tiết phổ biến nhất trong tiếng Việt cũng trích xuất từ nghiên cứu của nhóm tác giả Dien Dinh và cộng sự [3].

Tỉ lệ số từ Hán - Việt trong văn bản: Theo nhiều nghiên cứu, có hơn 60% số từ vựng tiếng Việt có nguồn gốc từ tiếng Hán - những từ mà chúng ta gọi là từ Hán - Việt. Các từ Hán - Việt thường được sử dụng trong các văn bản khoa học, văn bản kỹ thuật, và các văn bản có phong cách trang trọng, chính thức, vì vậy các từ Hán - Việt thường được xem là khó hơn so với các từ khác trong tiếng Việt. Chính vì thế, tỉ lệ số từ Hán - Việt cũng được chúng tôi sử dụng trong nghiên cứu này. Tương tự, chúng tôi cũng sử dụng thêm một số danh sách từ mượn khác như danh sách từ Pháp - Việt, Anh - Việt cùng lúc với danh sách từ Hán - Việt để sử dụng làm đặc trưng tỉ lệ số từ mượn trong văn bản.

Tỉ lệ từ địa phương: Lãnh thổ Việt Nam trải dài trên hơn 3.000km với nhiều vùng miền khác nhau, mỗi vùng đều có những nét văn hóa và cách sử dụng ngôn ngữ khác nhau. Nhiều vùng có những từ mà thường chỉ sử dụng ở vùng đó mà không hoặc ít có ở những nơi khác. Chính vì thế, trong các văn bản thông thường, các văn bản toàn dân, sự xuất hiện của các từ ngữ địa phương có thể ảnh hưởng tới độ khó của văn bản đó.

Tỉ lệ danh từ riêng: Trong văn bản, khi số lượng các danh từ riêng (tên người, tên địa danh…) càng nhiều, người đọc cần phải tốn nhiều công sức hơn để ghi nhớ và nhận diện được đầy đủ các thực thể mà các danh từ đó đề cập đến. Chính vì vậy, tỉ lệ số danh từ riêng trong văn bản có thể là một đặc trưng tốt để đánh giá độ khó của văn bản. Trong bài báo này, chúng tôi cũng sử dụng thêm một đặc trưng tương tự là tỉ lệ danh từ riêng phân biệt trên bộ từ vựng của văn bản (số từ phân biệt trong văn bản).

Với mỗi văn bản, chúng tôi trích xuất tất cả các đặc trưng đã nêu ở trên làm vector đặc trưng của văn bản để sử dụng trong bước thực nghiệm sẽ được mô tả trong phần 3.

3. Thực nghiệm

Trong các bài toán đánh giá độ khó của văn bản, việc xây dựng bộ ngữ liệu dùng để huấn luyện cần rất nhiều công sức bỏ ra: cần phải xác định xem văn bản thuộc mức độ khó nào. Nếu chưa có thang đo làm chuẩn mực để xác định mức độ khó của văn bản, chúng ta phải khảo sát trên nhiều người đọc với các độ tuổi, trình độ đọc khác nhau để xác định xem những nhóm người đọc nào có thể đọc hiểu được một văn bản nào đó và dùng những nhóm đó làm căn cứ để quyết định mức độ khó của văn bản. Đó là chưa kể đến số lượng các văn bản cần phải đánh giá phải đủ lớn để huấn luyện trong các phương pháp máy học.

Chúng tôi sử dụng hơn 200 văn bản thuộc lĩnh vực văn học được thu thập thủ công từ nhiều trang web làm nguồn ngữ liệu ban đầu. Nguyên nhân chủ yếu của việc lựa chọn lĩnh vực văn học là vì các văn bản thuộc lĩnh vực này dễ dàng thu thập và người đọc cũng không cần phải có nhiều kiến thức chuyên ngành để có thể đọc và hiểu được nội dung văn bản như trong các lĩnh vực chuyên môn khác. Các văn bản sau khi được thu thập sẽ trải qua nhiều công đoạn tiền xử lý như kiểm và sửa lỗi chính tả, chuẩn hóa các dấu thanh, chuẩn hóa bảng mã tiếng Việt, tách câu, tách từ và tách dấu câu. Bước kiểm và sửa lỗi chính tả được chúng tôi thực hiện thủ công. Các bước còn lại, chúng tôi sử dụng công cụ "CLC_VN_Toolkit" do Trung tâm Ngôn ngữ học Tính toán[1] phát triển. Đây là bộ công cụ hỗ trợ tiền xử lý văn bản, tách từ, gán nhãn từ loại, gán nhãn các thực thể có tên trong văn bản.

Các văn bản sau khi được tiền xử lý sẽ được lấy ra theo phương pháp tổ hợp để được các cặp văn bản. Với hơn 200 văn bản được thu thập và xử lý ban đầu, chúng tôi có thể chọn ra được hơn 20.000 cặp văn bản theo phương pháp tổ hợp. Tuy nhiên, vì một số hạn chế về chi phí và phần cứng máy tính, chúng tôi chỉ chọn ra 10.000 cặp văn bản để thực nghiệm. Các cặp văn bản này được đưa cho một số người đọc để xác định văn bản nào khó hơn. Những người đọc này là các học viên cao học chuyên ngành Ngôn ngữ học tại một số trường đại học ở Thành phố Hồ Chí Minh. Các cặp nào không xác định được văn bản nào khó hơn sẽ bị loại ra. Còn lại, mỗi

[1] CLC - Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh (http://www.clc.hcmus.edu.vn)

cặp văn bản sẽ được trích xuất thành 2 vector đặc trưng khác nhau tương ứng với 2 phân lớp: Văn bản A khó hơn Văn bản B; và Văn bản B dễ hơn Văn bản A. Hai vector đặc trưng này được tạo ra bằng cách lấy hiệu của vector đặc trưng của văn bản A với văn bản B và ngược lại.

Trong bài báo này, chúng tôi sử dụng thư viện sklearn trên python để xây dựng một mô hình máy học phân lớp các cặp văn bản theo độ khó, sử dụng thuật toán tính SVM (Support Vector Machine - Máy vector hỗ trợ). SVM là một thuật toán máy học có giám sát được sử dụng rất phổ biến ngày nay trong các bài toán phân lớp (classification), đặc biệt là các bài toán phân lớp văn bản. Ý tưởng của SVM là tìm một siêu phẳng (hyper lane) để phân tách các điểm dữ liệu được tạo ra bởi các vector đặc trưng của dữ liệu - ở trong nghiên cứu này là vector đặc trưng của các văn bản thu thập được. Siêu phẳng này sẽ chia không gian vector thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu. Trong nghiên cứu này, các miền dữ liệu chính là các nhãn tương ứng với phân lớp: Văn bản A khó hơn Văn bản B; và Văn bản B dễ hơn Văn bản A.

Nhằm tránh tình trạng quá khớp xảy ra, chúng tôi sử dụng phương pháp đánh giá chéo theo k-fold với k=10: chúng tôi ngẫu nhiên chia 10.000 cặp văn bản ra thành 10 phần và thực hiện 10 lần xây dựng và đánh giá mô hình phân lớp. Cứ mỗi lần thực hiện, 9 phần (tương đương với 9.000 cặp văn bản) sẽ dùng để huấn luyện mô hình và phần còn lại (tương đương 1.000 cặp văn bản) sẽ được dùng để đánh giá mô hình. Ở giai đoạn huấn luyện, toàn bộ 18.000 vector tương ứng được trích xuất từ 9.000 cặp văn bản (mỗi cặp văn bản trích được 2 vector đặc trưng) sẽ được đưa vào thuật toán SVM để xác định siêu phẳng phân tách dữ liệu. Sau đó, siêu phẳng này sẽ được kiểm tra trên 2.000 vector của 1.000 cặp văn bản còn lại để đánh giá độ chính xác của mô hình phân lớp. Kết quả, chúng tôi có một mô hình so sánh độ khó tương quan giữa 2 văn bản với đầu vào là vector đặc trưng được rút ra từ cặp văn bản đó và đầu ra là một chỉ số xác định xem trong 2 văn bản đó thì văn bản nào khó hơn. Độ chính xác đạt được của mô hình là 79,95%.

Một số ví dụ về kết quả so sánh của mô hình :

● Văn bản có độ khó cao nhất (kết quả so sánh khó hơn nhiều văn bản nhất): bài đọc "Rô-bin-xơn ngoài đảo hoang" (Sách giáo khoa môn Ngữ văn lớp 9, tập 2, Nhà xuất bản Giáo dục, tái bản lần thứ 6 năm 2011).

● Văn bản có độ khó thấp nhất (kết quả so sánh dễ hơn nhiều văn bản nhất): bài đọc "Điện thoại" (Sách giáo khoa môn Tiếng Việt lớp 2, tập 1, Nhà xuất bản Giáo dục, tái bản lần thứ 11 năm 2014).

● Cặp văn bản có kết quả so sánh không chính xác: bài đọc "Vẽ về cuộc sống an toàn" (Sách giáo khoa môn Tiếng Việt lớp 4, tập 2, Nhà xuất bản Giáo dục, tái bản lần thứ 9 năm 2014)

được mô hình máy tính đánh giá là khó hơn bài đọc "Những ngôi sao xa xôi" (Sách giáo khoa môn Ngữ văn lớp 9, tập 2, Nhà xuất bản Giáo dục, tái bản lần thứ 6 năm 2011). Nguyên nhân có thể là do bài đọc "Vẽ về cuộc sống an toàn" mặc dù có ít câu nhưng mỗi câu lại khá dài với các từ chủ yếu tên các địa danh hoặc tên riêng, dẫn tới độ dài trung bình của câu và tỉ lệ số danh từ riêng trong bài đọc này tăng cao hơn so với các bài đọc khác, làm ảnh hưởng tới đánh giá của mô hình máy tính.

4. Kết luận

Nghiên cứu về độ khó của văn bản trong tiếng Việt vẫn còn ít được quan tâm mặc dù đây là một tiền đề rất quan trọng trong việc soạn thảo và lựa chọn văn bản. Hơn nữa, nguồn ngữ liệu có thể dùng để khảo sát độ khó của văn bản tiếng Việt còn nhiều hạn chế do các vấn đề về bản quyền. Trong nghiên cứu này, chúng tôi đã trình bày một phương pháp so sánh độ khó của các văn bản tiếng Việt với nhau thông qua SVM với các đặc trưng ở mức từ được rút trích từ các văn bản. Phương pháp này không đòi hỏi quá nhiều ngữ liệu ban đầu mà chỉ cần một lượng nhỏ văn bản có tương quan độ khó chênh lệch nhau. Kết quả thực nghiệm trên 10.000 cặp văn bản cho thấy phương pháp này có thể đánh giá tương quan độ khó của các cặp văn bản với độ chính xác xấp xỉ 80%. Trong các nghiên cứu kế tiếp, chúng tôi sẽ tìm cách sắp xếp và phân nhóm độ khó của văn bản dựa trên các kết quả so sánh độ khó tương quan. Các ngữ liệu cụ thể hơn thuộc nhiều lĩnh vực hơn sẽ được thu thập để xây dựng các mô hình so sánh, đánh giá độ khó của văn bản tiếng Việt cho các lĩnh vực đó.

TÀI LIỆU THAM KHẢO

1. Bailin, A., & Grafstein, A. (2016). *Readability: Text and Context*: Palgrave Macmillan UK.
2. Dale, E., & Chall, J. S. (1949). The Concept of Readability. *Elementary English, 26*(1), 19 - 26.
3. Dinh, D., Nguyen, T. N., & Ho, H. T. (2018). Building a corpus-based frequency dictionary of Vietnamese. In, pp. 72 - 98.
4. Flesch, R. (1949). *The Art of Readable Writing*. New York: Harper and Brothers Publishers.
5. Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts.* Paper presented at the Human Language Technologies 2007: The

Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York.

6. Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training, Research B*(February), pp. 49.

7. Luong, A.-V., Nguyen, D., & Dinh, D. (2017). *Examining the text-length factor in evaluating the readability of literary texts in Vietnamese textbooks.* Paper presented at the 2017 9th International Conference on Knowledge and Systems Engineering (KSE).

8. Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*(8), pp. 639 - 646.

9. Nguyen, L. T., & Henkin, A. B. (1982). A Readability Formula for Vietnamese. *Journal of Reading, 26*(3), pp. 243 - 251.

10. Nguyen, L. T., & Henkin, A. B. (1985). A Second Generation Readability Formula for Vietnamese. *Journal of Reading, 29*(3), pp. 219 - 225.

11. Schwarm, S. E., & Ostendorf, M. (2005). *Reading Level Assessment Using Support Vector Machines and Statistical Language Models.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

12. Si, L., & Callan, J. (2001). *A Statistical Model for Scientific Readability.* Paper presented at the Proceedings of the Tenth International Conference on Information and Knowledge Management, New York, NY, USA.

13. Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting Texts by Readability. *Comput. Linguist., 36*(2), pp. 203 - 227.

14. Vajjala, S., & Meurers, D. (2012). *On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition.* Paper presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Montr\'{e}al, Canada.

COMPARING VIETNAMESE TEXTS BY READABILITY

Nguyễn Sơn[1], Hồ Viết Hoàng[1], Đinh Điền[2], Lương An Vinh[2], Nguyễn Thị Như Điệp[3]

[1] University of Foreign Languages - Hue University, 57 Nguyen Khoa Chiem, Hue

[2] University of Sciences, HCM National University

[3] University of Social Scinces and Humanities, HCM National University

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 an SVM classifier. The corpus we used for the experiment is Vietnamese literary texts evaluated for their relative readability by some readers. This method does not require too much effort to build a training corpus but also achieves approximately 80% accuracy. This is also a prerequisite for the comparison and selection of text to fit the reader's reading level.

Keywords: Text Readability, Comparing text, Vietnamese literary texts