TRANG THÔNG TIN LUẬN ÁN

Tên đề tài luận án: Xây dựng mô hình đánh giá độ khó của văn bản tiếng Việt
Ngành: Khoa học Máy tính
Mã số ngành: 62480101
Họ tên nghiên cứu sinh: Lương An Vinh
Khóa đào tạo: 2016
Người hướng dẫn khoa học: PGS. TS. Đinh Điền
Cơ sở đào tạo: Trường Đại học Khoa học Tự nhiên, ĐHQG.HCM

1. TÓM TẮT NỘI DUNG LUẬN ÁN:

Độ khó của văn bản là hệ thống các yếu tố ngôn ngữ của nội tại văn bản ảnh hưởng đến tính dễ đọc hay khó đọc của một văn bản. Các nghiên cứu về độ khó đã được bắt đầu từ lâu, nhưng hầu hết các nghiên cứu đó đều được thực hiện trên tiếng Anh và một số ngôn ngữ phổ biến trên thế giới. Trong tiếng Việt, trước đây chỉ có hai công trình nghiên cứu về Độ khó của văn bản, thực hiện trên hai bộ ngữ liệu khá nhỏ. Chính vì thế, rất cần có thêm các công trình nghiên cứu khác về độ khó của văn bản tiếng Việt thực hiện trên các bộ ngữ liệu cập nhật hơn, lớn hơn và trên các đặc trưng sâu hơn của văn bản như vai trò của từ, ngữ, cấu trúc ngữ pháp, ngữ nghĩa của câu…

Luận án đã xây dựng 2 bộ ngữ liệu dùng để khảo sát và thực nghiệm đánh giá độ khó văn bản tiếng Việt, gồm: (1) Bộ ngữ liệu 370 văn bản thu thập từ sách giáo khoa tiếng Việt và Ngữ văn; và (2) Bộ ngữ liệu 1.825 văn bản thuộc lĩnh vực văn học và ngôn ngữ học. Đây là 2 bộ ngữ liệu lớn và công khai đầu tiên về độ khó văn bản trong tiếng Việt. Luận án đã khảo sát 262 đặc trưng được trích xuất từ các văn bản này để xây dựng các công thức, các mô hình đánh giá độ khó văn bản. Các đặc trưng này được thuộc nhiều cấp độ của văn bản như các đặc trưng ở mức bề mặt (độ dài câu, độ dài từ, …), các đặc trưng về tần suất từ và tần suất chữ, các đặc trưng ở cấp độ ngữ pháp mức từ, mức câu, các đặc trưng thuộc về mô hình ngôn ngữ, các đặc trưng đơn giản ở cấp độ ngữ nghĩa và các đặc trưng của riêng tiếng Việt (như tỉ lệ từ mượn, tỉ lệ phương ngữ).

Luận án cũng đã giới thiệu 3 mô hình đánh giá độ khó văn bản tiếng Việt theo từng hướng tiếp cận: Với hướng tiếp cận thống kê, luận án đã thực hiện phân tích tương quan để chọn ra những đặc trưng có tương quan cao nhất với độ khó của văn bản, sau đó thực hiện phân tích hồi quy với một số cải tiến khi thực nghiệm để xây dựng công thức tính độ khó văn bản tiếng Việt. Kết quả cho thấy công thức mới được xây dựng có độ tương quan với độ khó văn bản vượt trội so với tất cả các nghiên cứu khác.

Với hướng tiếp cận máy học, luận án đã đề xuất sử dụng thuật toán RFECV để tự động chọn ra các đặc trưng có đóng góp tốt vào các mô hình máy học đánh giá độ khó văn bản dùng các thuật toán phân lớp truyền thống. Nhờ đó, mô hình mà luận án xây dựng đã đạt độ chính xác cao so với các nghiên cứu trước đây.

Ngoài ra, luận án cũng đề xuất một mô hình học sâu để phân lớp văn bản theo độ khó dựa trên mô hình ngôn ngữ tiền huấn luyện BERT và mạng LSTM. Độ chính xác của mô hình có giảm nhẹ so với các mô hình máy học truyền thống nhưng chúng ta tiết kiệm được chi phí để gán nhãn và trích xuất đặc trưng từ văn bản. Khi tích hợp thêm một số đặc trưng ngôn ngữ trích xuất từ văn bản vào mô hình học sâu, độ chính xác của mô hình đã được cải thiện và cao hơn so với các mô hình phân lớp truyền thống.

2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:

Việc khảo sát nhiều đặc trưng đã giúp luận án chọn ra được những đặc trưng có tương quan cao nhất với độ khó của văn bản. Công thức hồi quy được xây dựng từ những đặc trưng này có hệ số tương quan cao với độ khó của văn bản so với các nghiên cứu trước đây: thực nghiệm của luận án đạt hệ số tương quan 0,89, cao hơn 0,01 điểm so với nghiên cứu của Luong và các cộng sự [148] (sau khi điều chỉnh trọng số) và cao hơn 0,04 điểm so với Nguyen và Henkin (1982) [2]. Ngoài ra, luận án còn đề xuất cải tiến thực nghiệm phân tích hồi quy: sử dụng tất cả các đặc trưng

có tương quan cao với độ khó của văn bản để thực hiện phân tích. Kết quả thực nghiệm cho thấy công thức mới được xây dựng có kết quả về độ tương quan với độ khó văn bản là vượt trội so với tất cả các nghiên cứu khác (đạt hệ số tương quan 0,94).

Việc ứng dụng thuật toán xếp hạng đặc trưng RFECV để lựa chọn những đặc trưng tốt nhất ứng với từng thuật toán phân lớp đã giúp mô hình đánh giá độ khó văn bản mà luận án xây dựng được có độ chính xác cao so với các nghiên cứu trước đây: độ chính xác của mô hình phân lớp đạt ~85,7% với bộ ngữ liệu sách giáo khoa và 95,72% với bộ ngữ liệu văn học - ngôn ngữ học.

Sử dụng kỹ thuật học sâu dựa trên BERT và mạng LSTM mà luận án đề xuất, độ chính xác của mô hình có giảm nhẹ so với các mô hình máy học truyền thống (95,2% so với 95,72% trên bộ ngữ liệu văn học - ngôn ngữ học) nhưng chúng ta tiết kiệm được chi phí để gán nhãn và trích xuất đặc trưng từ văn bản. Khi tích hợp thêm một số đặc trưng ngôn ngữ trích xuất từ văn bản, độ chính xác của mô hình đã được cải thiện hơn so với các thuật toán phân lớp truyền thống (đạt 96,57%).

3. CÁC ỨNG DỤNG/ KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU

Kết quả của luận án có thể áp dụng vào nhiều lĩnh vực quan trọng như hỗ trợ biên soạn sách giáo khoa, giáo trình, viết báo, viết hướng dẫn sử dụng, viết định nghĩa trong từ điển giải thích bằng tiếng Việt, hỗ trợ dạy tiếng Việt cho người nước ngoài…

Do các hạn chế hiện hữu của các công cụ xử lý ngôn ngữ tiếng Việt, các đặc trưng mà luận án sử dụng chỉ dừng lại tới các đặc trưng đơn giản ở cấp độ ngữ nghĩa như tỉ lệ các từ đơn nghĩa, tỉ lệ các từ đa nghĩa , … Các đặc trưng ở sâu hơn ở mức ngữ nghĩa và ngữ dụng của văn bản hay các yếu tố liên kết trong văn bản không được xét đến trong luận án. Ngoài ra, do khuôn khổ của một luận án tiến sĩ, nên chúng tôi chỉ mới thu thập và khảo sát hai bộ ngữ liệu (ngữ liệu sách giáo khoa và ngữ liệu văn học - ngôn ngữ học). Việc áp dụng các mô hình đã đề xuất trong luận án sang các miền ngữ liệu khác có thể sẽ không tốt do các miền ngữ liệu đó có cách sử dụng từ, ngữ, câu và các yếu tố ngôn ngữ khác so với miền văn học - ngôn ngữ học. Trong tương lai, việc thu thập, xây dựng thêm các bộ ngữ liệu lớn hơn, khảo sát trên các đặc trưng sâu hơn ở cấp độ ngữ nghĩa cần được thực hiện để cải thiện độ tương quan và độ chính xác của các mô hình đánh giá độ khó văn bản tiếng Việt trong các lĩnh vực khác nhau.

TẬP THỂ CÁN BỘ HƯỚNG DẪN	**NGHIÊN CỨU SINH**
(Ký tên, họ tên)	(Ký tên, họ tên)
Đinh Điền	Lương An Vinh

XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO
HIỆU TRƯỞNG

THESIS INFORMATION

Thesis title: Building a model to assess the readability of Vietnamese texts
Speciality: Computer Science
Code: 62480101
Name of PhD Student: Lương An Vinh
Academic year: 2016
Supervisor: Assoc. Dr. Dinh Dien
At: VNUHCM - University of Science

1. SUMMARY:

Text readability is the system of linguistic factors of the text's internals that affect the easiness or difficulty of a text. Readability studies have been done for a long time, but most of them are in English and some popular languages. In Vietnamese, there were previously only two studies on text readability, performed on two relatively small corpora. Therefore, it is necessary to have other studies on the readability of Vietnamese texts on more updated, larger corpora and deeper features of the text such as the role of words, phrases, grammatical structure, semantics of sentences...

This thesis has built two corpora for examining and experimenting, including: (1) The corpus of 370 documents collected from textbooks of Vietnamese language and Literature; and (2) The corpus 1,825 texts in the field of literature and linguistics. These are the first two large and public corpora for text readability in Vietnamese. The thesis has examined 262 features extracted from these corpora to build formulas and models for assessing text readability. These features are included at many levels of the text such as surface-level features (sentence length, word length, etc.), word and word-frequency features, word and sentence-level grammatical features, language model features, simple semantic features and Vietnamese-specific features (such as the ratio of borrowed words, the ratio of dialects).

This thesis has also introduced three models for assessing the readability of Vietnamese texts according to each approach: With the statistical approach, the thesis has performed correlation analysis to select the features that have the most significant correlation with the readability of the text, then performed regression analysis with some improvements when experimenting to build a formula to calculate the readability of Vietnamese text. The results show that the newly formulated formula has a superior correlation with text readability compared to all other studies.

With the machine learning approach, the thesis has proposed to implement the RFECV algorithm to automatically select the features that make good contributions to machine learning models to evaluate text readability using traditional classification algorithms. As a result, the model developed by the thesis has achieved high accuracy compared to previous studies.

In addition, the thesis also proposes a deep learning model to classify documents according to difficulty based on BERT pre-training language model and LSTM neural network. The accuracy of the model is slightly reduced compared to traditional machine learning models, but we save the cost to label and extract features from the text. When integrating some more linguistic features extracted from text into the deep learning model, the accuracy of the model has been improved and is higher than that of traditional classification models.

2. NOVELTY OF THESIS:

Examining many features helped this thesis to select the features that are most correlated with the readability of the text. The regression formula built from these features has a high correlation coefficient with the readability of the text compared to previous studies: the thesis's experiment achieved a correlation coefficient of 0.89, higher than 0.01 point compared to the study of Luong et al. [148] (after weight adjustment) and 0.04 points higher than Nguyen and Henkin (1982) [2]. In addition, the thesis also proposes to improve the regression analysis experiment: use all the features highly correlated with the readability of the text to perform the analysis.

Experimental results show that the newly formulated formula has a superior correlation with text readability compared to all other studies (the correlation coefficient achieved is 0.94).

The application of the RFECV feature ranking algorithm to select the best features for each classification algorithm has helped the assessment model have high accuracy compared to previous studies: the accuracy of the classification model reached ~85.7% with the textbook corpus and 95.72% with the literary - linguistic corpus.

Using the deep learning technique based on BERT and LSTM network proposed by the thesis, the accuracy of the model is slightly reduced compared to traditional machine learning models (95.2% compared to 95.72% on the corpus literature-linguistic data) but we save the cost to label and extract features from the text. When integrating some more linguistic features extracted from the text, the accuracy of the model has been improved compared to the traditional classification algorithms (reaching 96.57%).

3. APPLICATIONS/ APPLICABILITY/ PERSPECTIVE

The results of the thesis can be applied to many important fields such as supporting the compilation of textbooks, writing articles, writing user manuals, writing definitions in explanatory dictionaries in Vietnamese, support Vietnamese language teaching for foreigners…

Due to the existing limitations of Vietnamese language processing tools, the features used by the thesis limited to simple features at the semantic level such as the ratio of monosyllabic words, the ratio of polysemantic words,... The deeper features at the semantic and pragmatic level of the text, or the linking elements in the text are unconsidered in the thesis. In addition, due to the scope of a doctoral thesis, we have only collected and examined two corpora (textbook corpus and literary-linguistic corpus). Applying the models proposed in the thesis to other domain may not be good because those domain have different usage of words, phrases, sentences and other linguistic elements compared to the literary-linguistic domain. In the future, the collection and construction of larger corpora, the examination on deeper features at the semantic level should be done to improve the correlation and accuracy of Vietnamese text readability assessment models in different domains.

SUPERVISOR	**PhD STUDENT**
Đinh Điền	Lương An Vinh

CERTIFICATION
UNIVERSITY OF SCIENCE
PRESIDENT