VỀ MỘT PHƯƠNG PHÁP XÁC ĐỊNH MỤC TIÊU VĂN BẢN TRONG TIẾNG VIỆT

Nguyễn Cảnh Hùng*, Đặng Hoàng Minh

Tóm tắt: *Trong bài báo này, chúng tôi giới thiệu mô hình xác định mục tiêu của văn bản tiếng Việt dựa trên cơ sở áp dụng hai giải thuật: giải thuật phân tách từ tiếng Việt sử dụng trường điều kiện ngẫu nhiên (CRFs) [1] và giải thuật phân loại văn bản StarSpace [2]. Kết quả thử nghiệm cho thấy, mô hình đề xuất đã tiến hành phân loại văn bản theo mục tiêu với độ chính xác tốt (hơn 90%) trên tập dữ liệu kiểm tra.*

Từ khóa: Phân loại văn bản; Tách từ; Các trường điều kiện ngẫu nhiên.

1. ĐẶT VẤN ĐỀ

Bài toán phân loại văn bản là một trong các lĩnh vực thu hút được sự chú ý rất lớn của cộng đồng nghiên cứu khóa học. Thực tế này xuất phát từ ý nghĩa thực tiễn của nó. Có thể định nghĩa, một bài toán phân loại văn bản (Text Classification) là một phép ánh xạ một văn bản (hoặc câu văn) sang một tập hữu hạn các chủ đề dựa trên nội dung của văn bản đó. Chính vì vậy, khi giải thuật phân loại văn bản được xây dựng, nó có thể được ứng dụng theo nhiều cách như: phân loại văn bản theo cảm xúc của người viết (tích cực hay tiêu cực); phân loại văn bản theo chủ đề (như: thể thao, chính trị, kinh tế,...). Bài toán xác định mục tiêu của văn bản cũng là một dạng không tách rời của bài toán phân loại văn bản. Trên thế giới hiện nay, rất nhiều giải thuật phân loại văn bản dựa trên Deep Learning[3] đã chứng minh được tính ưu việt của nó so với các công nghệ trước đó [4].

Tuy nhiên, việc áp dụng trực tiếp các giải thuật này vào ngôn ngữ tiếng Việt thường đem lại kết quả không cao. Lý do là vì, tiếng Việt là loại hình ngôn ngữ đơn lập. Nghĩa là mỗi tiếng được phát âm tách rời nhau và được thể hiện bằng một chữ viết. Mỗi từ có thể được cấu thành bởi một hoặc nhiều tiếng. Tiếng, về hình thức, nó trùng với một đoạn phát âm tự nhiên gọi là âm tiết. Về nội dung, nó là đơn vị nhỏ nhất có nội dung được thể hiện. Về ý nghĩa, có những tiếng tự thân nó đã mang một ý nghĩa, phản ánh một đối tượng hoặc khái niệm, ví dụ: cây, trời, cỏ, lá, ăn, nói, cười,… Có những tiếng không phản ánh hay thể hiện một ngữ nghĩa hay đối tượng nào cả. Nhưng bản thân sự có mặt của nó trong từ có thể tạo nên một sự khác biệt lớn. Nghĩa là, nó kết hợp với một hay nhiều tiếng khác có nghĩa để tạo nên từ (ví dụ: tiếng "sá" trong từ "đường sá", "e" trong từ "e lệ", "khúc" trong từ "khúc mắc"…).

Bên cạnh đó, cách viết tách từng tiếng của một từ ra cũng là sự khác biệt lớn giữa tiếng Việt và các ngôn ngữ khác, đặc biệt là tiếng Anh (ngôn ngữ nền tảng của các thử nghiệm giải thuật phân loại văn bản). Nói cách khác, trong tiếng Anh, mỗi từ mang ý nghĩa sẽ được phân tách với nhau bởi một khoảng trắng. Do đó, khi xử lý tiếng Anh, các giải thuật có xu thế phân tách các câu thành từng đơn vị ngữ nghĩa nhỏ dựa trên khoảng trắng. Trong khi đó, với tiếng Việt, phương pháp này sẽ phá với ý nghĩa của từ.

Trong bài báo này, nhóm nghiên cứu đề xuất áp dụng kết hợp 02 giải thuật là: giải thuật tách từ cho tiếng Việt và giải thuật phân loại văn bản StarSpace nhằm nâng cao hiệu quả của quá trình xác định mục tiêu văn bản.

2. CÁC GIẢI THUẬT SỬ DỤNG

2.1. Giải thuật tách từ tiếng Việt sử dụng các trường điều kiện ngẫu nhiên (Conditional Random Fields - CRFs)

Ta có thể quy bài toán tách từ trong tiếng Việt thành bài toán gán nhãn cho các âm tiết

tiếng Việt. Dựa vào các nhãn đó, ta có thể xác định được ranh giới của từng từ trong văn bản tiếng Việt. Các nhãn được sử dụng ở đây là:

• B_W: nhãn đánh dấu bắt đầu một từ;

• I_W: nhãn đánh dấu ở trong một từ.

Ví dụ, câu văn: "*Hôm nay là ngày Quốc Khánh nước Hà Lan*" sẽ được gán nhãn như sau:

Hôm	nay	là	ngày	Quốc	Khánh	nước	Hà	Lan
B_W	I_W	B_W	B_W	B_W	I_W	B_W	B_W	I_W

Dựa trên việc gán nhãn này, giải thuật sẽ đánh dấu các từ trong câu như sau:

"*Hôm_nay là ngày Quốc_Khánh nước Hà_Lan*"

Như vậy, bài toán phân đoạn từ tiếng Việt có thể phát biểu là:

"Hãy xây dựng một mô hình để gán nhãn {B_W, I_W} cho các âm tiết của văn bản tiếng Việt chưa được tách từ".

Bài toán này được giải khi mô hình tìm thấy nhãn phù hợp nhất cho từng âm tiết. Việc định nhãn này được biểu diễn bằng:

$$y^* = argmax\{p(y|x\} \quad (1)$$

Trong đó, y^* là nhãn cho âm tiết x. y^* là một trong các nhãn thuộc tập nhãn y.

Người ta có thể giải quyết bài toán này bằng nhiều mô hình như Markov ẩn [5]. Tuy nhiên, hiện nay CRFs thường được sử dụng hơn do kế thừa các ưu việt của mô hình trước đó, đồng thời, hoạt động tốt hơn trong trường hợp dữ liệu tồn tại nhiều ràng buộc phức tạp [6].

Giải phương trình trên bằng CRFs, ta có:

$$p(y|x) = \frac{1}{Z(x)} \exp\Big(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x) + \sum_i \sum_k \mu_k g_k(y_i, x)\Big) \quad (2)$$

Trong đó, x là chuỗi dữ liệu, y là chuỗi trạng thái tương ứng. f_k (y_i-1, y_i, x) là thuộc tính của chuỗi quan sát ứng và các trạng thái ứng với vị trí thứ i và i-1 trong chuỗi trạng thái. g_k (y_i, x) là thuộc tính của chuỗi quan sát và trạng thái ứng với trí thứ i trong chuỗi trạng thái. Các thuộc tính này được rút ra từ tập dữ liệu và có giá trị cố định. VD:

f_i = 1 nếu x_{i-1} = "Quyết", x_i = "định" và y_{i-1}=B_W, y_i=I_W

f_i = 0 nếu ngược lại

g_i = 1 nếu x_i = "Quyết" và y_i = B_W

g_i = 0 nếu ngược lại.

λ_i và μ_i là các tham số sẽ được ước lượng (học) trong quá trình huấn luyện. Quá trình ước lượng các tham số này được thực hiện bởi giải thuật tối ưu số bậc hai LBFGS (limited memory BFGS).

2.2. Giải thuật phân loại văn bản StarSpace

Trong thử nghiệm của mình, chúng tôi sử dụng mô hình giải thuật StarSpace cho bài toán xác định mục tiêu của văn bản. Giải thuật StarSpace do Facebook phát triển và công bố năm 2017. Kết quả thử nghiệm cho bài toán phân loại văn bản trên các tập dữ liệu tiếng Anh cho thấy: mô hình này đạt độ chính xác tốt hoặc tương đương so với các kiến trúc nổi tiếng như fastText.

Bên cạnh đó, việc lựa chọn giải thuật này cũng đến từ khả năng cho phép so sánh các thực thể không cùng loại của mô hình. Chính tính năng chỉ ra rằng, giải thuật có thể hoạt động tốt đối với nhiều ngôn ngữ mà không chỉ hoạt động tốt đối với tiếng Anh hoặc các ngôn ngữ có quy luật tương tự tiếng Anh.

Mô hình StarSpace bao gồm việc học các thực thể. Mỗi thực thể được mô tả bằng một tập hợp các tính năng riêng biệt. Mục tiêu là học ma trận có kích thước *Dxd*, trong đó *D* là số lượng các đặc trưng và *d* là chiều dài của vectơ embedding. Một thực thể *a* được biểu diễn dưới dạng $\sum_{i \in a} F_i$, trong đó, F_i là hàng thứ *i* (có kích thước *d*) trong ma trận embedding.

Hàm loss sau sẽ được cực tiểu hóa trong quá trình huấn luyện:

$$\sum_{\substack{(a,b) \in E^+ \\ b^- \in E^-}} L^{batch}(sim(a,b), sim(a,b_1^-), \dots, sim(a,b_k^-)) \quad (3)$$

Trong đó, việc tạo ra các cặp thực thể dương (a,b) thuộc E^+ và thực thể âm b^- thuộc E^- (phương pháp lấy mẫu k-âm (tương tự như trong word2vec) được sử dụng để lấy mẫu cho b_i^-) phụ thuộc vào từng ứng dụng cụ thể của mô hình (nội dung này sẽ được giải thích rõ hơn ở bên dưới).

Hàm $sim(.,.)$ là hàm tương tự, trong mô hình được đề xuất, nhóm tác giả triển khai cả hai phương pháp tính tương tự là cosine (cosine similarity) và tích trong (inner product), sau đó, để mô hình tự lựa chọn phương pháp phù hợp trong quá trình huấn luyện. Thông thường, các phương pháp này đều hoạt động tốt đối với số lượng nhãn nhỏ, tuy nhiên đối với tập nhãn kích thước lớn, hàm cosine cho kết quả tốt hơn.

Hàm loss L_{batch} sẽ so sánh cặp thực thể dương (a,b) với các cặp thực thể âm (a, b_i^-) với *i=1,...,k.* Quá trình huấn luyện được tối ưu hóa dựa vào giải thuật Stochastic gradient descent (SGD). Sau khi huấn luyện xong, hàm $sim(.,.)$ sẽ được sử dụng. Ví dụ trong các bài toán phân loại, nhãn *b* cho thực thể *a* sẽ được tính bằng $max_{\hat{b}}\, sim(a,\hat{b})$ đối với mọi nhãn $\hat{b}$. Hiểu một cách đơn giản là nhãn nào có tính tương đồng với thực thể *a* nhất sẽ được lựa chọn. Tùy vào ứng dụng cụ thể, mô hình này có thể được lựa chọn cấu hình khác nhau.

Đối với bài toán phân loại văn bản, cặp thực thể dương (a,b) được lấy trực tiếp từ tập huấn luyện, trong đó, *a* là nhóm từ đầu vào và *b* là nhãn tương ứng trong tập huấn luyện. Các thực thể âm b^- là các nhãn còn lại trong tập huấn luyện. Mô hình sẽ học cách cực đại hóa $sim(a,b)$ và cực tiểu hóa $sim(a,b_i^-)$.

Bằng việc kết hợp hai giải thuật trên vào một chuỗi xử lý thống nhất, nhóm đề tài tiến hành xây dựng mô hình phân loại văn bản tiếng Việt theo các mục tiêu cho trước.

3. CÁC THỬ NGHIỆM VÀ KẾT QUẢ

3.1. Bộ dữ liệu thử nghiệm

Bộ dữ liệu thử nghiệm của mô hình là các câu văn được lấy từ những văn bản trong mạng nội bộ của Viện CNTT. Các văn bản được lần lượt tách thành từng câu riêng biệt. Mỗi câu có nghĩa sẽ được phân về một trong các nhóm mục tiêu tương ứng:

- Công tác Đào tạo;
- Công tác Tài chính;
- Công tác Đảng công tác chính trị;
- Công tác hành chính hậu cần;
- Công đoàn và các tổ chức quần chúng khác;
- Công tác quản lý Khoa học công nghệ.

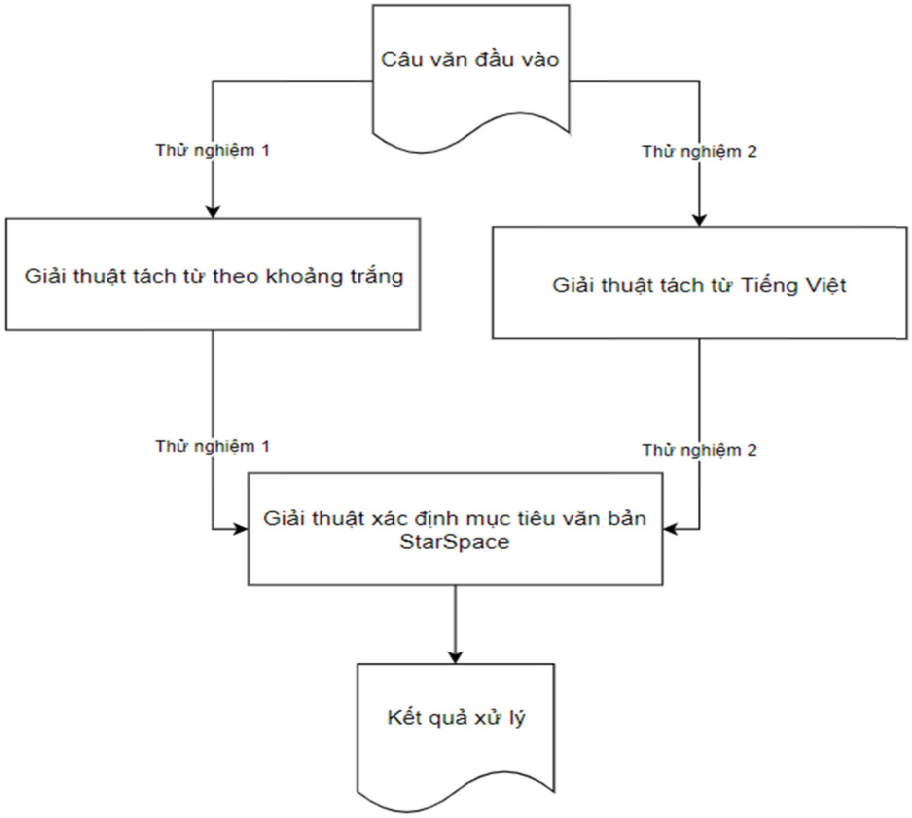
Kết quả, bộ dữ liệu xây dựng được gồm tổng 1200 câu với trung bình 200 câu cho một mục tiêu.

3.2. Phương pháp thử nghiệm và kết quả

Quá trình thử nghiệm được tiến hành trên cùng tập dữ liệu với hai sô đồ xử lý khác

nhau. Trong đó, thử nghiệm 1, dữ liệu được đưa qua giải thuật tách từ theo khoảng trắng (tức là coi mỗi tiếng là một từ). Trong thử nghiệm 2, dữ liệu được đưa qua giải thuật tách từ tiếng Việt trước khi đi vào giải thuật xác định mục tiêu văn bản.

Mô hình thử nghiệm tổng thể được xây dựng theo sơ đồ sau:



Hình 1. *Mô hình thử nghiệm.*

Bộ dữ liệu được tách thành 02 phần, trong đó: 900 mẫu được đưa vào huấn luyện và 300 mẫu (với 50 câu cho mỗi mục tiêu) được sử dụng để kiểm tra độ chính xác.

Bảng 1. *Kết quả xử lý đối với dữ liệu trong tập kiểm tra.*

Thử nghiệm	Độ chính xác
Thử nghiệm 1	88.1%
Thử nghiệm 2	93.7%

Kết quả thử nghiệm cho thấy, việc áp dụng thêm giải thuật tách từ tiếng Việt vào quá trình tiền xử lý trước khi đưa vào giải thuật phân loại mục tiêu văn bản tiếng Việt sẽ cho kết quả tốt hơn. Kết quả thử nghiệm này chỉ ra rằng, đối với các bài toán phân loại văn bản tiếng Việt nói chung và bài toán phân loại mục tiêu văn bản nói riêng, việc áp dụng giải thuật phân tách từ tiếng Việt là hết sức cần thiết.

4. KẾT LUẬN

Trong bài báo này, chúng tôi đã phân tích các giải thuật cần thiết để xây dựng một mô hình phân loại văn bản tiếng Việt. Trong đó, 02 giải thuật được sử dụng để tạo nên mô hình này là giải thuật phân tách từ tiếng Việt dựa trên CRFs và giải thuật phân loại văn bản StarSpace. Qua nội dung nghiên cứu này, chúng tôi hy vọng sẽ áp dụng kết quả vào các bài toán thực nghiệm như tìm kiếm, tra cứu thông minh.

Mặc dù kết quả thử nghiệm là khá khả quan, tuy nhiên, nó có thể đến từ tính độc lập tương đối của bộ dữ liệu. Trong các trường hợp khi bộ dữ liệu được phân tách thành các mục tiêu chứa nhiều nội dung, thuật ngữ trùng nhau (như mục tiêu "bóng đa", "bóng chuyền",...) chúng ta sẽ cần thêm nhiều cải thiện khác để nâng cao hiệu năng của giải thuật.

TÀI LIỆU THAM KHẢO

[1]. Lafferty, J., McCallum, A., Pereira "*Conditional random fields: Probabilistic models for segmenting and labeling sequence data*". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289, (2001).

[2]. Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, Jason Weston, "*StarSpace: Embed All The Things!*", Computation and Language (2017).

[3]. Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov. "*Enriching word vectors with subword information*". Transactions of the Association for Computational Linguistics 5:135–146 (2017)

[4]. Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, "*A neural probabilistic language model*". Journal of machine learning research 3(Feb):1137–1155

[5]. Baum, L. E.; Petrie, "*Statistical Inference for Probabilistic Functions of Finite State Markov Chains*". The Annals of Mathematical Statistics. 37 (6): 1554–1563. doi:10.1214/aoms/1177699147, (2011).

[6]. Sutton, Charles; McCallum, Andre, "*An Introduction to Conditional Random Fields*". arXiv:1011.4088v1 (2010).

ABSTRACT

A SUITABLE MODEL FOR CLASSIFYING VIETNAMESE DOCUMENTS

In this paper, we proposed a text classifying model for Vietnamese document. Our model is a combination of two separated components: A tokenization algorithm based on Conditional Random Fields (CRFs)[1] and StarSpace[2] – a general text classification model. Experiments results indicate that our model performed well on classifying task (with accuracy above 90% on the testing dataset).

Keywords: Text Classification; Tokenization; Conditonal Random Fields - CRFs.

Nhận bài ngày 02 tháng 01 năm 2020
Hoàn thiện ngày 15 tháng 02 năm 2020
Chấp nhận đăng ngày 10 tháng 4 năm 2020

Địa chỉ: Viện Công nghệ thông tin/Viện KH-CN quân sự.
*Email: hungbka48@gmail.com.