Tạp chí Tin học và Điều khiển học, T. 18, S. 3 (2002), 268–272

NHẬN DẠNG TỰ ĐỘNG NGÔN NGỮ TIẾNG ANH

TRẦN DUY HƯNG, NGUYỄN NGỌC CƯƠNG

Abstract. In practice, there are many problems which require the recognition of the language used in a text, especially a short message. This study flexibly applies some mathematical formula to determine whether the language used in a message of tens of characters is English or not.

Tóm tắt. Trong thực tế, có nhiều bài toán yêu cầu phải giải quyết vấn đề tự động nhận dạng ngôn ngữ sử dụng trong văn bản, đặc biệt là đối với những đoạn văn ngắn. Nghiên cứu này áp dụng mềm dẻo một số công thức toán học nhằm kiểm định ngôn ngữ sử dụng trong một đoạn văn có độ dài cỡ vài chục ký tự có là tiếng Anh hay không.

1. MỞ ĐẦU

Mỗi bản văn có ý nghĩa đều được viết bởi một hoặc một vài ngôn ngữ nhất định. Ví dụ: một câu tiếng Anh, một ghi chép cuộc hội thoại bằng tiếng Việt có xen lẫn một vài từ (câu) tiếng Anh, một đoạn chương trình máy tính viết bằng Pascal,... Hầu hết những đoạn văn ngắn trong thực tế thường sử dụng một ngôn ngữ duy nhất.

Đã có nhiều nghiên cứu kết luận ngôn ngữ có thể được biểu diễn bằng một mô hình xích Markov hữu hạn có bậc lớn hơn không. Quá trình thiết lập đoạn văn là quá trình xích Markov chuyển từ trạng thái này sang trạng thái kia với xác suất chuyển giữa chúng thể hiện đặc trưng của ngôn ngữ. Nói cách khác, ngôn ngữ viết được mô hình hóa bằng tập các trạng thái của xích Markov và ma trận xác suất chuyển giữa chúng. Người ta có thể dựa vào mô hình Markov này để giải quyết bài toán nhận dạng ngôn ngữ. Tuy nhiên, việc nhận dạng tự động bằng máy tính đòi hỏi độ dài của đoạn văn phải đủ lớn để các qui luật này thể hiện rõ ràng, giảm thiểu sai lầm.

Người ta cho rằng bài toán nhận dạng tự động ngôn ngữ của một văn bản có thể xếp vào một trong bốn bài toán tổng quát sau:

Bảng 1. Bốn bài toán nhận dạng ngôn ngữ tổng quát

Stt	**Mô tả bài toán**	**Giả thiết H_0**	**Đối thiết H_1**
1	Nhận dạng ngôn ngữ đã biết B	$P_M = P_B$	$P_M \neq P_B$
2	Phân biệt ngôn ngữ đã biết B với nhiễu ngẫu nhiên	$P_M = P_B$	$P_M = P_U$
3	Phân biệt nhiễu bậc 0 với ngôn ngữ bậc 1 chưa biết	$b(M) = 0$	$b(M) = 1$
4	Phát hiện ngôn ngữ chưa biết	$P_M = P_U$	$P_M \neq P_U$

Trong bảng trên, P_M là ma trận xác suất chuyển trạng thái của xích Markov M tạo ra đoạn văn bản, P_B là ma trận xác suất chuyển của xích Markov B mô tả ngôn ngữ đã biết, P_U là ma trận xác suất chuyển của nguồn nhiễu ngẫu nhiên. Hàm số $b(M)$ là bậc của mô hình xích Markov M.

Nghiên cứu của Ravi Ganesan và Alan T. Sherman [2] đã xác định các tiêu chuẩn kiểm định mạnh nhất cho bốn bài toán cơ bản kể trên. Vận dụng chúng một cách mềm dẻo, chúng tôi đã thu được những kết quả khá tốt trong nghiên cứu của mình.

Trong nghiên cứu này, chúng tôi giới hạn bài toán nhận dạng ngôn ngữ trong việc phân biệt đoạn văn tiếng Anh ngắn không có giãn cách từ và dấu câu với các xâu ký tự cùng độ dài được tạo ngẫu nhiên. Hai bài toán được áp dụng là:

Bài toán 4. Phát hiện một ngôn ngữ có qui luật nhưng chưa biết qui luật đó. Giả thiết H_0 được phát biểu: *Đoạn ký tự đang xem xét là một dãy các ký tự được sinh ngẫu nhiên ($P_M = P_U$).* Đối

thiết H_1 sẽ là $P_M \neq P_U$, có nghĩa là: *Đoạn ký tự đang xét không phải là một dãy ký tự ngẫu nhiên,* nó là đoạn văn được viết bằng một ngôn ngữ nào đó.

Bài toán 2. Phân biệt tiếng Anh và nhiễu ngẫu nhiên. Giả thiết H_0: *Đoạn ký tự đang xét được viết bằng tiếng Anh* ($P_M = P_B$). Đối thiết H_1: *Đoạn ký tự được xem xét là một dãy các ký tự được sinh ngẫu nhiên* ($P_M = P_U$).

2. THỬ NGHIỆM NHẬN DẠNG XÂU KÝ TỰ CỦA TIẾNG ANH

2.1. Các tiêu chuẩn kiểm định giả thiết

Với Bài toán 4, tiêu chuẩn kiểm định mạnh nhất mà [2] đưa ra là

$$\Lambda_4 = -N\ln m - \sum_{1\leq i,j\leq m} n_{ij}\ln\hat{p}_{ij} \quad \text{với} \quad \hat{p}_{ij} = \frac{n_{ij}}{n_{i^*}},$$

n_{ij} là tần số của bộ đôi ij trong đoạn văn, n_{i^*} là tần số xuất hiện các bộ đôi có chữ cái đầu là i; m là số trạng thái của mô hình ($m = 2$); N là tổng số các bộ đôi có trong bản văn. Nếu đếm số bộ đôi theo cách tuần tự (nghĩa là chia đoạn văn thành các cặp ký tự liên tiếp và rời nhau rồi đếm) của đoạn văn có độ dài X, thì $N = [X/2]$. Trường hợp đếm số bộ đôi theo cách móc xích (ký tự sau của cặp này lại là ký tự đầu của cặp tiếp theo) của đoạn văn trên, thì $N = X - 1$.

Tiêu chuẩn này yêu cầu mẫu có độ dài khá lớn. Chúng tôi đã giải quyết hạn chế này bằng cách chia 26 trạng thái (26 ký tự A-Z) thành hai nhóm trạng thái 0-1. Trong đó, 13 ký tự xuất hiện nhiều hơn trong tiếng Anh (E, T, A, O, I, N, S, R, H, L, C, U) được thể hiện bằng trạng thái 1; còn 13 ký tự ít xuất hiện hơn trong tiếng Anh được thể hiện bằng trạng thái 0.

Theo [2], đại lượng $-2\Lambda_4$ có phân bố tiệm cận đến phân bố χ^2 với bậc tự do $m(m-1) = 2$. Chọn mức ý nghĩa $\alpha = 0{,}005$ khi đó $\chi^2_2(0{,}005) = 10{,}597$. Như vậy, nếu $-2\Lambda_4 \geq 10{,}597$ ta bác bỏ giả thiết H_0 và công nhận đối thiết H_1, có nghĩa là công nhận ngôn ngữ sử dụng để viết đoạn văn là tiếng Anh. Việc lựa chọn hợp lý mức ý nghĩa α giúp chúng ta giảm thiểu được sai lầm trong kết luận.

Tương tự, với Bài toán 2, ta sử dụng tiêu chuẩn kiểm định mạnh nhất:

$$\Lambda_2 = \ln\frac{L(P_B)}{L(P_U)} = \sum_{1\leq i,j\leq m} n_{ij}\ln p_{ij} + \sum_{1\leq i,j\leq m} n_{ij}\ln\frac{1}{m} = \sum_{1\leq i,j\leq m} n_{ij}\ln p_{ij} + N\ln m.$$

Vì $N\ln(m)$ là thành phần tuyến tính, công thức trên tương đương với

$$S_2 = \sum_{1\leq i,j\leq m} n_{ij}\ln p_{ij}$$

trong đó n_{ij} là số lần xuất hiện của cặp bộ đôi ký tự ij, p_{ij} là tần suất xuất hiện cặp bộ đôi ij trong tiếng Anh (xác suất chuẩn). Tiêu chuẩn S_2 chính là tiêu chuẩn của Sinkov đã nêu trong [2].

Để có thể xác định ngưỡng quyết định, ta cần chuẩn hóa S_2 như sau:

$$\hat{S}_2 = \frac{(S_2/N) - \mu_S}{\sigma_S/\sqrt{N}},$$

các tham số được tính như sau:

$$\mu_S = \sum_{1\leq i,j\leq m} b_{ij}\ln p_{ij} \quad \text{và} \quad \sigma_S^2 = \sum_{1\leq i,j\leq m} b_{ij}(\ln p_{ij} - \mu_S)^2.$$

Nếu b_{ij} là tần suất bộ đôi tuần tự và p_{ij} là tần suất bộ đôi móc xích, các tham số sẽ là:

$$\mu_{S_1} = -2{,}6734; \quad \sigma^2_{S_1} = 4{,}2076$$

Nếu b_{ij} là tần suất bộ đôi móc xích và p_{ij} là tần suất bộ đôi tuần tự, giá trị các tham số sẽ là:

$$\mu_{S_2} = -5{,}8887; \quad \sigma^2_{S_1} = 1{,}5812.$$

Vì $\hat{S}_2$ có phân phối tiệm cận phân bố chuẩn, chọn mức ý nghĩa α =0,05 ta có $u(\alpha) = 1{,}96$. Như vậy, $|\hat{S}_2| \geq 1{,}96$ thì bản rõ đang được xem xét không phải là bản rõ tiếng Anh với xác suất đúng là 95%.

Chúng tôi cũng áp dụng một tiêu chuẩn nhận dạng nữa vào bài toán này. Đó là Phi Test, công thức như sau:

$$\Delta IC = \frac{26\sum_{i=1}^{m} f_i(f_i-1)}{N(N-1)}$$

với f_i là tần số xuất hiện của ký tự i trong đoạn ký tự có độ dài là N. Với những đoạn ký tự có độ dài không nhỏ hơn 50 ký tự, nếu kết quả ΔIC lớn hơn 1,25 thì ta có thể kết luận ngôn ngữ của đoạn ký tự đang xem xét là tiếng Anh. Ngược lại, nếu kết quả tính toán là nhỏ hơn 1,25 ta kết luận mẫu đang xét là xâu ngẫu nhiên (xem [1]).

2.2. Các tần suất chuẩn

Trên cơ sở giảm số trạng thái của xích Markov, chúng tôi đã tiến hành tính tần suất P_B cho văn bản tiếng Anh. Bản văn có độ dài là 10.000 ký tự A-Z, được lấy từ nhiều nguồn tư liệu có tính điển hình cho văn bản tiếng Anh. Việc tính toán tần suất bộ đôi dựa trên cơ sở mô hình hai trạng thái 0, 1 như mô tả ở trên. Kết quả tần suất được trình bày ở bảng 2.

Bảng 2. Tần suất bộ đôi ký tự đối với tiếng Anh

Tần suất bộ đôi tuần tự

	0	1
0	0.0214	0.1429
1	0.1428	0.6929

Tần suất bộ đôi móc xích

	0	1
0	0.0198	0.1488
1	0.1400	0.6914

2.3. Các kết quả thử nghiệm

Chúng tôi đã thử nghiệm các tiêu chuẩn nói trên nhiều lần với cùng một độ dài là 64 ký tự trên mỗi mẫu. Số lượng mẫu cho mỗi lần thử là 200, trong đó 100 đoạn văn tiếng Anh khác nhau được lấy trong lĩnh vực thương mại điện tử được trộn lẫn với 100 xâu ký tự ngẫu nhiên được tạo bằng thuật toán:

```
1. srand((unsigned) time(&t));     // Chỉ gọi một lần cho một thử nghiệm
2. for (int i = 0; i < len; i++)   // len = 64 ký tự
3. temp[i] = rand() % 26 + 97;     // Ký tự ngẫu nhiên từ 'a' đến 'z'
4. temp[len] = '\0';               // Kết thúc xâu bằng NULL
5. return temp;
```

Thuật toán trên đảm bảo bộ sinh số ngẫu nhiên của máy tính được khởi tạo lại với nhân (t) mới theo thời gian. Do đó các xâu được tạo sẽ ngẫu nhiên hơn.

Thử nghiệm cho cả hai cách chấm tần suất (tuần tự và móc xích), chúng tôi thu được các kết quả trình bày dưới đây.

2.3.1 Thử nghiệm Bài toán 4

- Số mẫu mỗi lần thử = 100 đoạn tiếng Anh + 100 đoạn ngẫu nhiên.
- Mức ý nghĩa $\alpha = 0{,}005$

Từ kết qủa được đưa ra ở bảng 3 sau có thể thấy ngay được rằng với tần suất bộ đôi *móc xích,* phép thử Λ_4 rất mạnh đối với bài toán nhận dạng ngôn ngữ, đặc biệt đối với những bài toán cần tránh loại sai lầm do bỏ sót bản văn đọc được. Với cách chấm tần suất bộ đôi tuần tự phép thử Λ_4 cũng có thể được áp dụng trong một số trường hợp có nhu cầu ngược lại, vì xác suất nó chấp nhận nhầm một xâu ngẫu nhiên là bản văn đọc được cũng là rất nhỏ. Cũng xin chú ý thêm rằng, đối với Bài toán 4, chúng tôi chỉ xét hai giả thiết: hoặc dãy ký tự là ngẫu nhiên, hoặc nó là dãy ký tự của tiếng Anh. Trường hợp cần phân biệt nhiều ngôn ngữ cùng lúc chúng ta sẽ xem xét sau.

Bảng 3. Kết quả thử nghiệm Bài toán 4

Sai lầm trong thử nghiệm Λ_4		Lần thử								Sai số trung bình
		1	2	3	4	5	6	7	8	
Tần xuất móc xích	Chấp nhận nhầm: xâu ngẫu nhiên được coi là bản rõ tiếng Anh	1	1	0	1	0	0	0	1	4/1600 mẫu thử (4/800 xâu ngẫu nhiên)
	Bác bỏ sai: bản rõ tiếng Anh được coi là xâu ngẫu nhiên	0	1	0	0	0	0	1	0	2/1600 mẫu thử (2/800 bản rõ)
Tần xuất tuần tự	Chấp nhận nhầm: xâu ngẫu nhiên được coi là bản rõ tiếng Anh	0	0	0	0	0	1	0	1	4/1600 mẫu thử
	Bác bỏ sai: bản rõ tiếng Anh được coi là xâu ngẫu nhiên	25	15	13	12	17	14	16	9	121/1600 mẫu thử

2.3.2 Thử nghiệm với bài toán 2

- Số mẫu mỗi lần thử = 100 đoạn tiếng Anh + 100 đoạn ngẫu nhiên.
- Mức ý nghĩa $\alpha = 0{,}05$

Bảng 4. Kết quả thử nghiệm Bài toán 2

Sai lầm trong thử nghiệm Λ_2		Lần thử								Sai số trung bình
		1	2	3	4	5	6	7	8	
Tần xuất móc xích	Chấp nhận nhầm: xâu ngẫu nhiên được coi là bản rõ tiếng Anh	0	0	0	0	0	0	0	1	1/1600 mẫu thử (1/800 xâu ngẫu nhiên)
	Bác bỏ sai: bản rõ tiếng Anh được coi là xâu ngẫu nhiên	15	23	14	7	17	19	21	10	126/1600 mẫu thử (126/800 bản rõ)
Tần xuất tuần tự	Chấp nhận nhầm: xâu ngẫu nhiên được coi là bản rõ tiếng Anh	0	0	0	0	0	0	0	0	0/1600 mẫu thử
	Bác bỏ sai: bản rõ tiếng Anh được coi là xâu ngẫu nhiên	0	12	4	1	7	6	6	1	37/1600 mẫu thử

Phép kiểm định Λ_2 "bỏ sót" khá nhiều đoạn văn tiếng Anh. Tuy nhiên phép kiểm định này lại ít sai lầm nhất khi bác bỏ tính ngẫu nhiên của các xâu ký tự đặc biệt là khi sử dụng tần số bộ đôi tuần tự.

2.3.3 Thử nghiệm Phi Test

- Số mẫu mỗi lần thử = 100 đoạn tiếng Anh + 100 đoạn ngẫu nhiên.
- Mức ngưỡng đánh giá = 1,25

Bảng 5. Kết quả thử nghiệm Phi Test

Sai lầm trong thử nghiệm ΔIC	Lần thử								Sai số trung bình
	1	2	3	4	5	6	7	8	
Chấp nhận nhầm: xâu ngẫu nhiên được coi là bản rõ tiếng Anh	4	2	0	1	0	5	1	4	17/1600 mẫu thử (17/800 xâu ngẫu nhiên)
Bác bỏ sai: bản rõ tiếng Anh được coi là xâu ngẫu nhiên	2	5	3	0	0	3	0	2	15/1600 mẫu thử (15/800 bản rõ)

ΔIC có kết quả "dung hòa" giữa hai loại sai lầm xuất hiện trong bài toán. Điểm mạnh của tiêu chuẩn này chính là mức độ phức tạp tính toán thấp nhất trong số các phép thử mà chúng tôi đã xem xét ở trên.

3. KẾT LUẬN

So sánh kết quả thử nghiệm của cả ba phép kiểm định được trình bày ở trên, chúng tôi thấy không có trường hợp nào mà ở đó cả ba phép thử đều mắc sai lầm. Điều đó chứng tỏ ta có thể dùng kết hợp hai hoặc cả 3 tiêu chuẩn này để giảm thiểu các sai lầm gặp phải khi nhận dạng tự động ngôn ngữ trong điều kiện mẫu thử có độ dài hạn chế.

TÀI LIỆU THAM KHẢO

[1] Department of the Army, *Basic Cryptanalysis*, Aegean Park Press, Washington, DC, 13 Sep 1990.

[2] Ravi Ganesan and Alan T. Sherman, Statistical Techniques for Language Recognition, Cryptologia **XVII** (4) 1993.

Nhận bài ngày 11 - 7 - 1999
Nhận lại sau khi sửa ngày 5 - 8 - 2002

Trần Duy Hưng - 38/119 Hồ Đắc Di, Hà Nội
Nguyễn Ngọc Cương - 8/1/126 Hoàng Văn Thái, Hà Nội