Tạp chí Tin học và Điều khiển học, T.20, S.4 (2004), 319–328

NHẬN BIẾT NGÔN NGỮ VÀ BỘ MÃ SỬ DỤNG TRONG CÁC VĂN BẢN ĐA NGỮ

PHAN HUY KHÁNH[1], VÕ TRUNG HÙNG[2]

[1]*Đại học Đà Nẵng*
[2]*GETA-CLIPS, ENSIMAG, CH Pháp*

Abstract. This article presents our new method in order to automatically identify any languages and coding systems used in a heterogeneous multilingual texts by the calculation of the characteristic coefficient of the language and its coding on the different areas of documents.

Tóm tắt. Bài báo trình bày một giải pháp mới để nhận biết tự động các ngôn ngữ và bộ mã sử dụng trong các văn bản đa ngữ không thuần nhất bằng cách tìm hệ số đặc trưng cho ngôn ngữ và bộ mã sử dụng trên các vùng văn bản khác nhau.

1. MỞ ĐẦU

Cách đây không lâu, trong giai đoạn đầu của Tin học, hầu hết phần mềm đều mới chỉ xử lí được dữ liệu tiếng Anh (hoặc tiếng Nga). Người sử dụng (NSD) bắt buộc có thói quen làm việc với tiếng Anh như là ngôn ngữ giao tiếp chủ yếu và máy tính chỉ sử dụng một số bộ mã thông dụng như EBCDIC, ASCII.... Đây là điều trở ngại rất lớn cho NSD khi cần làm việc trong các ngôn ngữ, hay hệ viết (writing system), không phải là tiếng Anh. Ngày nay, khi nhu cầu xử lí thông tin bằng nhiều thứ tiếng khác nhau, khi máy tính và mạng Internet được sử dụng rộng rãi, thì việc nghiên cứu, phát triển và ứng dụng các hệ thống tin học đa ngữ (multilinguality), dùng ngôn ngữ tự nhiên (natural language), đã trở thành một nhu cầu tất yếu và ngày càng được nhiều người quan tâm. Ngay từ những năm 1980, người ta bắt đầu nghiên cứu phát triển các hệ thống xử lí văn bản đa ngữ, không những trên các máy tính chuyên dụng đặc biệt của một số nhà sản xuất (Xerox chẳng hạn [7]), mà ngày càng phổ biến trên những máy tính thường dùng (PC, Macintosh, các máy Unix...) [9]. Nhờ những tiến bộ đạt được, NSD đã có thể làm việc cùng lúc với nhiều ngôn ngữ khác nhau và sử dụng nhiều bộ mã khác nhau trên cùng một máy tính, trên cùng một ứng dụng.

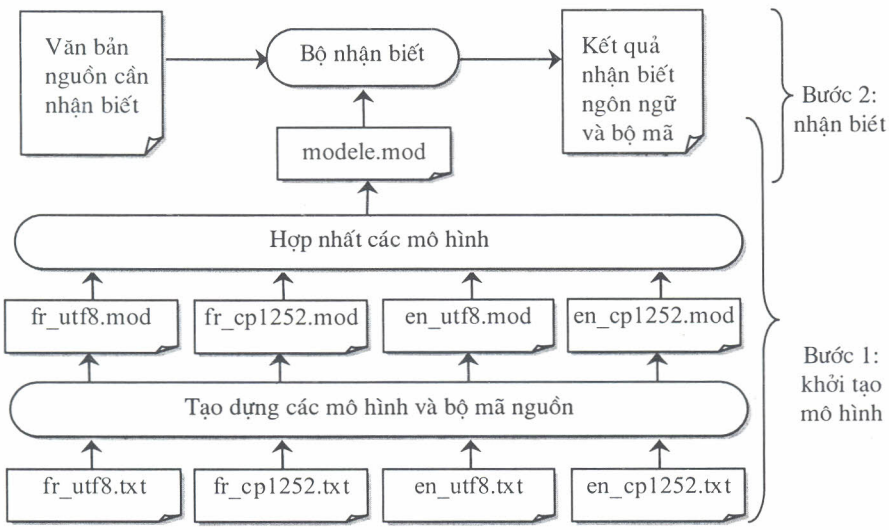
Để thao tác trên các dữ liệu dạng văn bản, gọi chung là các trang văn bản, viết trong một ngôn ngữ hoặc trong một nhóm ngôn ngữ nào đó, người ta có thể chỉ cần sử dụng một bộ mã nhưng cũng có thể sử dụng nhiều bộ mã khác nhau. Ví dụ bộ mã chuẩn ISO8859-1 (hoặc một số bộ mã khác như ISO8879, CP1252, CP1258,...) được dùng cho tiếng Anh, tiếng Đức... và một số hệ viết sử dụng chữ cái LaTin ở các nước Châu Âu, như Pháp, Ý, Bồ Đào Nha, Tây Ban Nha, Ru-ma-ni.... Tiếng Hoa có các bộ mã như GB3212-80 được sử dụng ở lục địa, JIS C6226 ở Nhật Bản, BIG-5 ở Đài Loan. Riêng tiếng Việt, đã có rất nhiều bộ mã đã được đề xuất và sử dụng phổ biến như VNI, TCVN3-ABC, Vietware, VPS, BK HCM, VIQR, v.v.. Hiện nay, Unicode là bộ mã đang được nhiều người khuyến khích tiêu chuẩn hóa và sử dụng đại trà cho tất cả các hệ viết sử dụng trên máy tính.

Tính trạng có nhiều bộ mã, mỗi bộ mã có thể sử dụng cho nhiều ngôn ngữ, một ngôn ngữ sử dụng nhiều bộ mã khác nhau và tính phong phú về yếu tố ngôn ngữ trong nội dung các trang văn bản xử lí trên máy tính đã gây ra những khó khăn rất lớn cho NSD khi nghiên cứu và phát triển các ứng dụng đa ngữ, đặc biệt là trong lĩnh vực xử lí ngôn ngữ tự nhiên (natural language processing). Do đó, việc nhận biết ngôn ngữ và bộ mã sử dụng trong mọi kiểu trang văn bản đã đóng một vai trò quan trọng trong hầu hết các thao tác xử lí thông tin, như đưa vào - đưa ra thông tin, trao đổi thông tin giữa các ứng dụng, kiểm tra sửa lỗi chính tả, sửa lỗi ngữ pháp, tìm kiếm, chuyển mã, dịch tự động đa ngữ, v.v.. Khi cần nhận biết ngôn ngữ và bộ mã sử dụng, người ta thường phân biệt hai loại văn bản: loại văn bản thuần nhất (homogeneous) chỉ sử dụng một ngôn ngữ và một bộ mã, và loại văn bản không thuần nhất hay văn bản hỗn tạp (heterogeneous) sử dụng đồng thời nhiều ngôn ngữ và nhiều bộ mã khác nhau.

Trong Mục 2 của bài báo này, chúng tôi giới thiệu hai phương pháp tiêu biểu ứng dụng cho các trang văn bản thuần nhất đang được sử dụng hiện nay, là thống kê trên các dãy kí tự có độ dài xác định (n-gram method) và thống kê các từ ngữ pháp đặc trưng (grammatical words method). Trong Mục 3, chúng tôi đề xuất giải pháp mới cho phép nhận biết tự động các trang văn bản đa ngữ không thuần nhất bằng cách tìm một hệ số tương quan (correlative coefficient) từ các hệ số đặc trưng (characteristic coefficient) cho ngôn ngữ và bộ mã sử dụng trên các vùng văn bản.

2. NHẬN BIẾT NGÔN NGỮ VÀ BỘ MÃ TRONG VĂN BẢN THUẦN NHẤT

Để nhận biết những ngôn ngữ nào và những bộ mã nào đã được sử dụng trong văn bản thuần nhất đang xét, người ta tiến hành nhận biết qua hai bước [4, 5, 6, 13]: bước đầu tiên là khởi tạo các mô hình ngôn ngữ (linguistic models), bước tiếp theo là sử dụng các mô hình ngôn ngữ đã khởi tạo này để thực hiện nhận biết trên văn bản. Sơ đồ trong hình 1 dưới đây biểu diễn hai bước của quá trình nhận biết.



Hình 1. Sơ đồ biểu diễn quá trình nhận biết ngôn ngữ và bộ mã

Bước khởi tạo, còn được gọi là bước "dạy máy học", bao gồm việc tạo dựng mô hình và hợp nhất mô hình. Nội dung việc tạo dựng mô hình là quá trình thống kê tần suất xuất hiện của dãy các kí tự trong các tệp văn bản mẫu đóng vai trò "bài học" đã được chuẩn bị trước. Hiện nay, người ta đã đề xuất nhiều phương pháp "dạy máy học" khác nhau căn cứ vào cách nhìn nhận sự xuất hiện liên tiếp của các kí tự trong văn bản. Điển hình là phương pháp thống kê trên các dãy các kí tự có độ dài xác định và phương pháp thống kê các từ ngữ pháp đặc trưng cho một ngôn ngữ.

Các tệp dữ liệu văn bản "bài học" lưu giữ thông tin về một ngôn ngữ và bộ mã xác định để xây dựng mô hình ngôn ngữ tương ứng. Ví dụ tệp fr-utf8.txt lưu giữ thông tin tiếng Pháp (French) sử dụng mã UTF-8, tệp en-cp1252.txt lưu giữ thông tin tiếng Anh (English) sử dụng mã CP1252, v.v.. Sau khi "dạy máy học", mỗi một mô hình được tạo ra sẽ chứa nội dung là các lớp kí tự và tần suất xuất hiện tương ứng của chúng, đó là các tệp fr-utf8.mod, en-cp1252.txt, v.v.. Việc tiếp theo là hợp nhất các mô hình này để nhận được một mô hình ngôn ngữ duy nhất, chẳng hạn đó là tệp modele.mod, dành cho tất cả các ngôn ngữ và các bộ mã.

Bước nhận biết sử dụng mô hình đã khởi tạo để đoán nhận một văn bản đưa vào bất kỳ, gọi là văn bản nguồn, đã được viết trong ngôn ngữ nào và đã sử dụng những bộ mã nào. Trong bước này, người ta gọi lại phương pháp đã sử dụng trong bước khởi tạo để xây dựng mô hình (thống kê theo độ dài hay theo từ ngữ pháp đặc trưng).

2.1. Phương pháp thống kê theo độ dài của từ

Ý tưởng của phương pháp là nhận biết sự lặp lại của một dãy các kí tự có độ dài cố định nào đó trong một văn bản. Tùy theo ngôn ngữ mà số lần xuất hiện của một dãy kí tự như vậy là nhiều hơn hay ít hơn. Ví dụ, trong tiếng Anh, các từ chứa dãy kí tự tận cùng là *ck* nhiều hơn trong tiếng Pháp, nhưng trong tiếng Pháp, các từ kết thúc bởi dãy kí tự *ez* lại nhiều hơn trong tiếng Anh. Vì vậy, phương pháp này thống kê tần suất xuất hiện của các dãy kí tự được phân theo lớp có độ dài cố định n khác nhau, gọi là mô hình n-gram, $n = 1$, $n = 2$, $n = 3$, v.v.. Mô hình n-gram có thể áp dụng cho một giá trị n xác định hoặc sử dụng kết hợp nhiều giá trị n cho việc nhận biết.

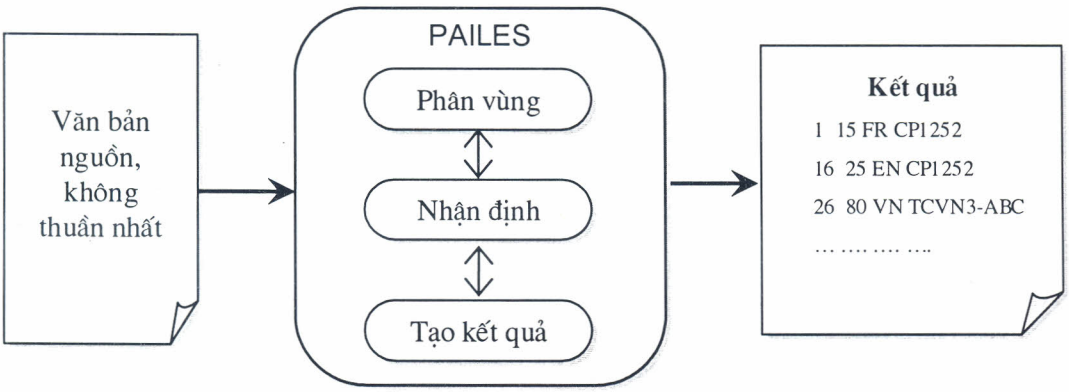
Ví dụ, câu tiếng Pháp "Les chiens et les chats sont des animaux" (dịch ra tiếng Việt: chó và mèo đều là những con vật), người ta thu được các mô hình n-gram tương ứng như sau (chú ý dấu _ trong bộ là dấu cách giữa các từ trong câu).

Bảng 1. Thống kê tần suất xuất hiện theo độ dài n trong mô hình n-gram

Lớp độ dài $n = 1$		Lớp độ dài $n = 2$		Lớp độ dài $n = 3$	
Dãy kí tự	Tần suất	Dãy kí tự	Tần suất	Dãy kí tự	Tần suất
_	7	*s_*	4	*es_*	3
s	6	*es*	3	*les*	2
e	5	*le*	2	*s_c*	2
a	3	*_c*	2		
n	3	*ch*	2		
t	3				

Trong thuật toán "dạy máy học", người ta sử dụng một vòng lặp để thống kê (đếm) tần suất xuất hiện của các dãy kí tự thuộc các lớp kí tự độ dài lần lượt $n = 1, 2, 3...$, từ một tệp

tiến hành nhận biết mã và ngôn ngữ.



Hình 2. Công cụng nhận dạng văn bản không thuần nhất

PAILES có ba khối chức năng chính là phân vùng, nhận định và tạo kết quả:

• Khối phân vùng có chức năng cắt văn bản nguồn ra thành từng vùng nhỏ hơn để xem xét. Mỗi vùng được xác định bởi vị trí của kí tự đầu vùng và vị trí của kí tự cuối vùng. Cách tính vị trí theo kiểu lũy tiến kể từ 1 trở lên. Ví dụ vùng đầu tiên của văn bản có cặp vị trí là $(1, n_{v1})$, vùng 2 là $(n_{v1} + 1, n_{v2})$, v.v...

• Khối nhận định hoạt động như sau:

- Kiểm tra vùng được cắt ra có là thuần nhất hay không?

- Nếu thuần nhất thì tiến hành xác định vùng này đã sử dụng bộ mã nào cho ngôn ngữ nào nhờ mô hình ngôn ngữ. Tiếp tục xác định vùng tiếp theo.

- Nếu không thuần nhất thì quay lên khối phân vùng để tiếp tục cắt thành các vùng nhỏ hơn nữa để sau đó nhận dạng lại. Quá trình tiếp tục cho đến khi không còn văn bản để nhận dạng.

• Khối tạo kết quả tạo ra một bảng liệt kê. Mỗi dòng của bảng, tương ứng với một vùng văn bản thuần nhất đã cắt ra, cho biết vị trí kí tự đầu vùng, vị trí kí tự cuối vùng, tên của ngôn ngữ và tên bộ mã sử dụng cho vùng văn bản này.

Ví dụ: Giả sử ta có đoạn văn bản song ngữ sau đây:

Tổng thống Pháp G. Si-răc khi phát biểu trên Đài truyền hình TF1 về cuộc chiến tranh tại I-răc đã nhận định rằng vấn đề này đã được biết đến từ lâu (nguyên văn tiếng Pháp: "C'est un problème qui date de longtemps"). Ông khẳng định Pháp giữ vững lập trường phản đối chiến tranh dưới bất kỳ hình thức nào.

Khi thực hiện, PAILES đã cắt đoạn văn bản nguồn (tổng cộng 304 kí tự) ra thành ba vùng thuần nhất, lần lượt là: {Tổng thống... tiếng Pháp:}, {"C'est ... longtemps").} và {Ông ... hình thức nào.}.

Sau khi phân tích, PAILES tạo ra bảng liệt kê kết quả như sau.

Bảng 2. Kết quả phân tích bằng phương pháp tìm hệ số đặc trưng theo vùng

Vị trí đầu vùng	Vị trí cuối vùng	Ngôn ngữ	Bộ mã
1	173	Tiếng Việt	TCVN3-ABC
174	217	Tiếng Pháp	CP1252
218	304	Tiếng Việt	TCVN3-ABC

3.3. Tìm hệ số tương quan từ các hệ số đặc trưng

Trong PAILES, khối nhận định có nhiệm vụ nhận biết vùng văn bản đang xét sử dụng bộ mã nào và được viết trong ngôn ngữ nào. Để có thể nhận biết, ta cần phải tìm hệ số đặc trưng l phản ánh *độ tin cậy* (certainty) cho mỗi ngôn ngữ và bộ mã tương ứng. Hệ số đặc trưng l được xác định dựa trên tần suất xuất hiện của các lớp kí tự trong mô hình ngôn ngữ của văn bản cần đánh giá.

Sử dụng hệ số đặc trưng, chúng ta tính hệ số tương quan q giữa hai ngôn ngữ để có được giá trị cao nhất theo công thức (2) như sau:

$$q = \frac{l_1 - l_2}{l_1}. \tag{2}$$

Trong đó:
l_1 là hệ số đặc trưng cao nhất, được tính trong công thức (1) đối với mô hình ngôn ngữ đang xét có giá trị lớn nhất;
l_2 là hệ số đặc trưng thứ cấp, được tính trong công thức (1) đối với mô hình ngôn ngữ đang xét có giá trị lớn thứ hai.

PAILES sẽ sử dụng hệ số tương quan để đánh giá một vùng văn bản đang xét có thuần nhất hay không. Nếu hệ số tương quan của một vùng văn bản nhỏ hơn một giá trị xác định λ nào đó thì phải tiếp tục chia cắt vùng này để nhận được những vùng nhỏ hơn, mà mỗi vùng có thể là thuần nhất. Giá trị λ được chọn theo công thức tương ứng theo công thức (1) và tùy thuộc vào khả năng chính xác khi đánh giá một đoạn văn bản có độ dài tối thiểu là bao nhiêu (đoạn văn bản đánh giá càng dài thì độ chính xác càng cao), trong PAILES, chúng tôi chọn $\lambda = 0{,}25$.

Ví dụ trên một đoạn văn bản đánh giá, giả sử ta tính được $l_1 = 0{,}7$, $l_2 = 0{,}3$, khi đó:

$$q = \frac{0{,}7 - 0{,}3}{0{,}7} = 0{,}57,$$

do $q > \lambda$, kết quả đưa ra chính là ngôn ngữ và bộ mã trong mô hình ngôn ngữ đang xét tương ứng với l_1. Nhưng nếu $l_1 = 0{,}7$ và $l_2 = 0{,}6$, lúc đó tính được $q = 0{,}14 < \lambda$, ta nhận định đoạn văn bản đang xét là không thuần nhất (vì có thể chứa nhiều hơn một ngôn ngữ hoặc chứa nhiều hơn một bộ mã). Lúc này, cần phải chia đoạn văn bản này thành các đoạn nhỏ hơn để đánh giá hoặc buộc phải kết luận theo l_1 nếu không thể chia nhỏ hơn được nữa.

3.4. Thuật toán nhận biết

Sau đây là thuật toán chính để xây dựng công cụ nhận biết ngôn ngữ và bộ mã trong các văn bản đa ngữ không thuần nhất PAILES.

Input: Văn bản nguồn không thuần nhất cần nhận biết.
Chọn giá trị λ.

Output: Kết quả phân vùng cùng với kết quả nhận biết ngôn ngữ và bộ mã sử dụng tương ứng.

Begin
Khởi tạo các mô hình ngôn ngữ
Repeat
Gọi thủ tục phân vùng để lấy ra một vùng văn bản cần đánh giá
Tính giá trị hệ số tương quan $q = (l_1 - l_2)/l_1$
If $q > \lambda$ **Then**
Chọn ngôn ngữ và bộ mã theo hệ số đặc trưng cao nhất l_1
Else
If Độ dài của vùng được cắt đủ lớn để phân chia được
Then
Tiếp tục gọi thủ tục phân vùng để lấy ra một vùng văn bản nhỏ hơn
Else
Chọn ngôn ngữ và bộ mã tương ứng với l_1
EndIf
EndIf
Until Cho đến khi xử lý hết các vùng trong văn bản
Gọi thủ tục tạo bảng liệt kê kết quả
End

Trong thủ tục phân vùng, chúng ta có thể sử dụng nhiều phương pháp khác nhau để cắt văn bản thành các vùng văn bản nhỏ hơn, như cắt theo câu (mỗi câu kết thúc bởi một dấu chấm câu), cắt đều văn bản thành các lớp có độ dài bằng nhau, hay có độ dài biến đổi. Mặt khác, có thể sử dụng kết hợp nhiều phương pháp nhận biết khác nhau tùy thuộc vào độ dài của các vùng văn bản cần được nhận biết.

3.5. Đánh giá kết quả sử dụng công cụ PAILES

Sau đây là bảng kết quả cho biết độ tin cậy bằng cách sử dụng một số công cụ nhận biết so sánh với công cụ PAILES của chúng tôi cho văn bản đồng nhất trên một số ngôn ngữ quen thuộc có độ dài câu từ 20 đến 200 chữ.

Bảng 3. So sánh độ tin cậy (%) sử dụng các công cụ nhận biết văn bản đồng nhất. Các dấu * cho biết cặp ngôn ngữ và bộ mã không tồn tại trong công cụ đang xét hay cần chuyển mã văn bản trước khi nhận biết

Ngôn ngữ (tiếng)	*Bộ mã sử dụng*	*Độ tin cậy*				
		SILC	Xerox	Textcat	Stochastic	PAILES
Anh	CP 1252	100,00	98,50	65,00	98,00	96,50
Pháp	CP 1252	87,00	88,50	92,50	88,00	93,00
Đức	CP 1252	90,00	92,00*	87,00*	90,00*	92,00
Ả Rập	CP 1256	91,00	88,00	92,00	*	85,00
Ý	CP 1252	88,00	90,00*	90,00*	93,00*	90,00
Bồ Đào Nha	CP 1252	85,00	90,00*	93,00*	95,00*	91,00
Nga	KOI8-R	80,00	60,00	80,00	*	89,50

Hán	BIG5	0,00*	70,00	85,00	*	75,00
Hán	GB 2312	85,00	80,00	83,00	*	80,00
Nhật	SHIFT-JIS	90,00	77,00	89,00	*	89,00
Nhật	EUC-JP	80,00	92,00	80,00	*	78,00
Việt Nam	VPS	*	*	99,00	*	81,00
Việt Nam	TCVN3	*	*	*	*	76,00
Việt Nam	UTF-8	*	*	*	*	56,00
Việt Nam	VNI	*	*	*	*	66,00

Nhìn vào bảng kết quả, ta nhận thấy công cụ PAILES luôn luôn cho kết quả trong mọi trường hợp và xử lý được các văn bản tiếng Việt mà các công cụ khác không thể thực hiện được. Đối với các văn bản không đồng nhất, chúng tôi nhận được kết quả như sau.

Bảng 4. So sánh độ tin cậy (%) cho các văn bản không đồng nhất.

Ngôn ngữ	*Bộ mã sử dụng*	*Số câu nhận biết*	*Số câu đúng*	*Độ tin cậy*
		1000	998	99,80
Pháp	UTF-8	1000	1000	100,00
Tây Ban Nha	CP 1252	1000	990	99,00
Đức	CP 1252	1000	993	99,30
Bồ Đào Nha	CP 1252	1000	995	99,50
Ý	CP 1252	1000	990	99,00
Nga	KOI-8	1000	1000	100,00
Việt Nam	TCVN3	1000	900	90,00
Việt Nam	UTF-8	1000	900	90,00
Việt Nam	VNI	1000	850	85,00
Vietnamien	VPS	1000	890	89,00

4. KẾT LUẬN

Việc nhận biết ngôn ngữ và bộ mã sử dụng trong văn bản (thuần nhất hay không thuần nhất) có ý nghĩa quan trọng trong các hệ thống xử lí thông tin đa ngữ. Việc nhận biết này giúp hệ thống có được những bước lựa chọn các xử lí thích đáng cho từng ngôn ngữ và bộ mã đang được sử dụng. Hiện nay, vẫn chưa có được những giải pháp triệt để, sẵn dùng và thuận tiện cho NSD khi họ cần làm việc với các trang văn bản đa ngữ. Việc đề xuất xây dựng PAILES đã giúp NSD một phương tiện để nhận biết ngôn ngữ và bộ mã sử dụng trong từng vùng văn bản đa ngữ không đồng nhất đang cần được xử lí. Công cụ PAILES có thể trợ giúp kiểm tra lỗi chính tả và ngữ pháp bằng cách xác định từng vùng được viết trong ngôn ngữ nào để áp dụng từ điển sửa lỗi tương ứng với ngôn ngữ đó. Trong việc dịch tự động đa ngữ, PAILES có thể xác định ngôn ngữ nào hiện đang được sử dụng trên văn bản nguồn để gọi trình dịch tương ứng sang ngôn ngữ đích. Ngoài ra, công cụ PAILES có thể tích hợp vào các hệ thống xử lí văn bản đa ngữ để thực hiện các công việc như xác định sự sai lệch mã để tự động chuyển về một mã thống nhất theo yêu cầu của NSD, cho phép chọn phông chữ thích hợp để hiện văn bản lên màn hình, đưa ra máy in, v.v...

Chúng tôi sẽ tiếp tục phát triển công cụ này để áp dụng vào hệ thống dịch tự động đa ngữ UNL bằng cách nhận dạng từng vùng văn bản được viết trong ngôn ngữ nào, từ đó xác

định cặp ngôn ngữ cần dịch (ngôn ngữ nguồn và ngôn ngữ đích) để sử dụng bộ dịch tương ứng. Hiện nay, chúng tôi đang hợp tác với nhóm GETA-CLIPS, IMAG, INPG-UJF-CNRS, Cộng hòa Pháp để có thể góp phần tham gia dự án quốc tế UNL dịch tự động cho 15 ngôn ngữ (Anh, Pháp, Đức, Ý, Nga, Nhật, Hàn Quốc, Trung Quốc, Thái Lan, v.v.).

TÀI LIỆU THAM KHẢO

[1] C. Manning and H. Schutze, Foundations of statistical natural language, *Processing, MIT Press*, 1999.

[2] Ch. Boitet. "Projet FeV - Réalisation d'un dictionnaire d'usage et d'une base terminologique par acceptions informatisés français-vietnamien via l'anglais". Tài liệu nội bộ Dự án FEV, GETA-CLIPS, IMAG (UJF, CNRS & INPG), CH Pháp.

[3] E. Giguet, The stakes of multilinguality: Multilingual text tokenization in natural language Diagnosis, *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence Workshop "Future issues for Multilingual Text Processing"*, Cairns, Australia, August 27.

[4] G. Benny, *Reconstruction et Utilisation de SILC, Rapport de Stage*, Département d'Informatique et de Recherche Opérationelle, Université de Montréal, 2001.

[5] G. Grefenstette. *Comparing two Language Identification Schemes*, JADT'95, 1995.

[6] G. Russell, *The QUE Language and Encoding Identification Package*, RALI, University of Montreal, 2002.

[7] J. Berker, *Multilingual Word Processing*, Microsystems, February, 1984.

[8] K. R. Beesley, Language identifier: A computer program for automatic natural language identification of on-line text, In Language at Crossroads, *Proceedings of the 29th Annual Conference of the American Translators Association*, 1998.

[9] Phan Huy Khánh, "Contribution à l'informatique multilingue. Extension d'un éditeur de documents structurés". Luận án Tiến sỹ Tin học, CH Pháp, 1991.

[10] Phan Huy Khánh và Võ Trung Hùng, Thiết kế cơ sở dữ liệu đa ngữ ngữ pháp tiếng Việt, *Tạp chí Khoa học Công nghệ* , Số 36, 37 (2002) 19–24.

[11] TCVN (Tiêu chuẩn Việt Nam), Bộ mã chuẩn 8-bit chữ Việt LaTinh dùng trong trao đổi thông tin, *Kỷ yếu Tuần lễ Tin học VI*, Hà Nội, 1996.

[12] V. Bouffard: *Evaluation de SILC, Rapport Scientifique*, Département d'Informatique et de Recherche Opérationelle, Université de Montréal, 2002.

[13] W. Cavnar and J. Trenkle, *N-gram Based Text Categorization, Symposium on Document Analysis and Information Retrieval*, University of Nevada, Las Vegas, 1994.

Nhận bài ngày 13 - 6 - 2003
Nhận lại sau sửa ngày 11 - 10 - 2003