TỔ CHỨC DỮ LIỆU TỪ ĐIỂN ĐIỆN TỬ CHO DỊCH MÁY ANH – VIỆT

Đinh Điền

Khoa Công nghệ Thông tin - trường ĐH Khoa học Tự nhiên - ĐHQG TP.HCM

(Bài nhận ngày 04 tháng 10 năm 2001, hoàn chỉnh sửa chữa ngày 22 tháng 1 năm 2002)

TÓM TẮT: Trong hệ dịch máy, cơ sở dữ liệu quan trọng nhất chính là từ điển điện tử của nó. Vì vậy, nội dung chính của bài báo này sẽ đề cập đến việc tổ chức các dữ liệu trong từ điển điện tử của một hệ dịch máy Anh – Việt. Cách thức chọn lựa và tổ chức dữ liệu trong từ điển điện tử cho hệ dịch máy hoàn toàn phụ thuộc vào phương pháp dịch và chiến lược dịch của hệ thống, và nói chung việc lựa chọn dữ liệu nào cần lưu trong từ điển sẽ hoàn toàn không giống như cách làm đối với từ điển dành cho người.

1. GIỚI THIỆU:

Không giống như từ điển dành cho người, từ điển dành cho dịch máy không cần chứa các thông tin về ngữ âm (phonetics), về từ nguyên (etymology), nhưng nó lại chứa các thông tin được coi là hiển nhiên, những thông tin không cần nêu ra trong từ điển dành cho người sử dụng (vì những thông tin này con người hoàn toàn suy ra được bằng tri thức về thế giới thực hay bằng vốn sống). Từ điển cho hệ dịch Anh – Việt ở đây được xây dựng theo những quy tắc sau:

- Phương ngữ dùng trong từ điển: thông tin *nghĩa tiếng Việt* trong từ điển này được dựa trên phương ngữ Hà Nội (ví dụ : blanket *: mền, chăn,* pig *: lợn, heo,*...).
- Luật chính tả: dựa theo Qui định 240-1984 của Bộ Giáo dục về chuẩn hóa chính tả tiếng Việt trong ngành Giáo dục.
- Bộ mã tiếng Việt: bộ mã ABC (1 byte) theo tiêu chuẩn TCVN-5712.
- Tiêu chí lựa chọn mục từ: tất cả các từ trong từ điển Anh – Việt thông thường, các cụm từ: từ ghép (compound), ngữ cố định (phrasal), thành ngữ (idiom), tục ngữ (proverb), từ viết tắt, tên riêng,...

2. THÔNG TIN TRONG TỪ ĐIỂN CHO HỆ DỊCH MÁY ANH – VIỆT

Cấu trúc vi mô (microstruture) của từ điển dành cho hệ Dịch máy Anh – Việt bao gồm những thông tin được trình bày một cách có hệ thống trong mỗi mục từ, như: thông tin về hình thái, cú pháp, và ngữ nghĩa nhằm điều khiển quá trình dịch.

2.1 THÔNG TIN VỀ HÌNH THÁI:

Bao gồm các thông tin về:

- Dạng của từ (word form), ví dụ: "program", "book",...
- Mã về dạng của từ (word form): từ nguyên gốc, dạng bất quy tắc, ...
- Mã đặc tính hình thái: như có gấp đôi phụ âm hay không, hay kết hợp với phụ tố (affix) nào...
- Mã về vị trí của từ (word order): từ này thường đứng ở vị trí nào trong ngữ, trong câu,...

2.2 THÔNG TIN VỀ CÚ PHÁP

Bao gồm các thông tin về:

- Từ loại (Parts-of-speech) của từ, như: danh từ, động từ, tính từ,...
- Chủng loại (Subcategory): như danh từ thuộc loại con nào (danh từ đếm được, không đếm được,...), động từ loại con nào (tha động từ, tự động từ,...),...
- Đặc tính cú pháp (syntactic feature): về thì (tense): quá khứ, hiện tại, tương lai; thể (voice): bị động, chủ động; giống (gender): đực, cái, trung; số (number): ít, nhiều,...
- Đặc tính cấu trúc (structure): từ này dùng trong cấu trúc nào, mẫu câu nào.
- Ngữ đi kèm (collocation/phrase/idiom): từ này hay đi kèm với những từ nào, dùng trong ngữ (thành ngữ, tục ngữ) nào.

2.3 THÔNG TIN VỀ NGỮ NGHĨA

Bao gồm các thông tin về:

- Nghĩa tiếng Việt (meaning) của từ.
- Đặc điểm tiếng Việt: khi dịch ra tiếng Việt, cần hiệu chỉnh gì về nghĩa (thêm, bớt các tiểu từ, loại từ, định từ,...), về vị trí.
- Các nét nghĩa của từ (semantic feature): như HUMAN, ANIMATE, OBJECT,... các khái niệm này được liên kết với hệ cơ sở tri thức về ngữ nghĩa của từ được rút ra từ mạng WordNet (xin xem [Miller, George A. (1996)]).
- Vai trong văn phạm cách (case role): Agent (Human), Instrument (Object),...
- Thông tin về nhóm đồng nghĩa (synonym): trỏ tới synset trong WordNet.
- Thông tin về nhóm phản nghĩa (antonym): trỏ tới synset trong WordNet.

2.4 THÔNG TIN VỀ NGỮ CẢNH

Bao gồm các thông tin về:

- Lĩnh vực sử dụng (field): từ này thường được dùng trong những lĩnh vực nào, ví dụ: Tin học, toán học, y học,...
- Tần số xuất hiện (frequency): từ này có thường được dùng hay không, ví dụ: xét theo đơn vị 1.000 từ (mã 01 → có mặt trong 1.000 từ thông dụng nhất, 02 → trong 2.000 từ thông dụng nhất,...) (theo [Khang, Nguyễn Văn, 1994]).
- Mã về tình thái (modality): từ này dùng trong cảnh huống nào (trịnh trọng, thân mật, thông tục,...).

3. CÀI ĐẶT TỪ ĐIỂN ĐIỆN TỬ TRONG HỆ DỊCH ANH – VIỆT

3.1 KHẢO SÁT DỮ LIỆU VÀ MÔI TRƯỜNG SỬ DỤNG TỪ ĐIỂN

3.1.1 Khảo sát về tần số xuất hiện của các từ tiếng Anh:

Chúng tôi thử thống kê trên 2000 từ đầu tiên (không kể hình) của phần 1 chương 3 trong một tài liệu tiếng Anh (tài liệu [Smith, George W. 1991] từ trang 33-43), và nhận thấy các từ xuất hiện thường xuyên nhất đều là những từ chức năng (hư từ), như: "the", "of",.. và kết quả này cũng trùng với kết quả khảo sát tổng hợp trong tập ngữ liệu lớn Brown (xin xem [Smith, George W. (1991)], trang 80). Dưới đây là bảng ghi lại kết quả khảo sát 2 tập ngữ liệu trên:

Bảng 1: thống kê tần số xuất hiện của các từ thông dụng:

(đơn vị tính là % trên số từ của văn bản)

Một chương sách		Ngữ liệu Brown		Một chương sách		Ngữ liệu Brown	
4.4%	*the*	*.9*	*word*	*7.8%*	*the*	.7	it
3.7	of	.8	have	4.6	of	.6	are
3.5	and	*.8*	*information*	2.6	and	.6	this
2.4	*words*	*.8*	*lexicon*	2.5	in	.6	on
2.2	a	.8	be	2.5	to	.6	or
2.2	to	.8	which	2.2	a	.5	which
2.1	in	.7	with	1.5	is	.5	from
2.0	as	.7	not	1.0	that	.5	not
1.7	or	*.7*	*Lexical*	1.0	for	.5	at
1.2	for	.6	some	.8	be	.4	were
1.2	are	.6	more	.8	as		
1.1	that	.6	but	.8	by		
1.0	is	*.6*	*spellings*	.7	with		

So sánh kết quả trên, ta thấy: ở cột thứ nhất, khi khảo sát một ngữ liệu cụ thể (một phần của 1 chương sách), ngoài những từ thông thường hay xuất hiện nhất, như: "the" (4.4%) , "of" (3.7%),... còn có những từ đặc biệt (chiếm tỉ lệ khoảng 1%) phan ánh đề tài mà chương sách đang bàn đến, cụ thể ở đây là vấn đề "từ vựng" qua các từ chính như: "words", "word",...Vậy, tần số từ còn phụ thuộc vào lĩnh vực.

Nhưng khi người ta khảo sát tổng hợp trên rất nhiều các ngữ liệu thuộc các lĩnh vực khác nhau, thì các từ chức năng chiếm ưu thế tuyệt đối, nhất là từ "the" (7.8%), và ta có thể nhận thấy, hầu hết các từ thông dụng này có chiều dài nhỏ (cỡ 3 ký tự).

3.1.2 Khảo sát về độ dài của các từ tiếng Anh:

Chúng tôi đã thống kê trên từ điển 109.582 từ tiếng Anh theo độ dài và nhận thấy:

- Tổng số có 995 từ tiếng Anh có chiều dài từ 3 ký tự trở xuống.
- Nếu chỉ căn cứ vào 3 ký tự đầu tiên, thì ta sẽ có 2.991 từ phân biệt.
- Từ có chiều dài ngắn nhất là 1 ký tự (có 2 từ là: "a" và "I"), dài nhất là 28 ký tự (có 1 từ là: "antidisestablishmentarianism"). Nhiều nhất là từ dài 8 ký tự (19.461 từ). Tính trung bình, từ có chiều dài cỡ 8.5 ký tự.

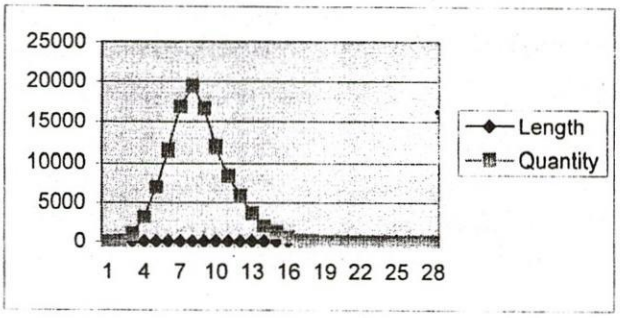
Bảng 2: khảo sát số lượng từ theo chiều dài:

Chiều dài L_i	Số lượng từ Q_i	Chiều dài L_i	Số lượng từ Q_i	Chiều dài L_i	Số lượng từ Q_i
1	**2**	11	8374	21	11
2	**140**	12	5812	22	4
3	**853**	13	3676	23	2
4	3130	14	2102	24	0
5	6919	15	1159	25	1
6	11492	16	583	26	0
7	16882	17	229	27	0
8	**19461**	18	107	28	1
9	16693	19	39	>28	0

10	11882	20	29		

Chiều dài trung bình của từ là (1):

$$\bar{L} = \frac{\sum_{i=1}^{28} L_i * Q_i}{\sum_{i=1}^{28} Q_i} \approx 8.5$$



Hình 1: Số lượng từ theo chiều dài

3.1.3 Khảo sát số ký tự khác biệt trung bình tại mỗi vị trí trong từ:

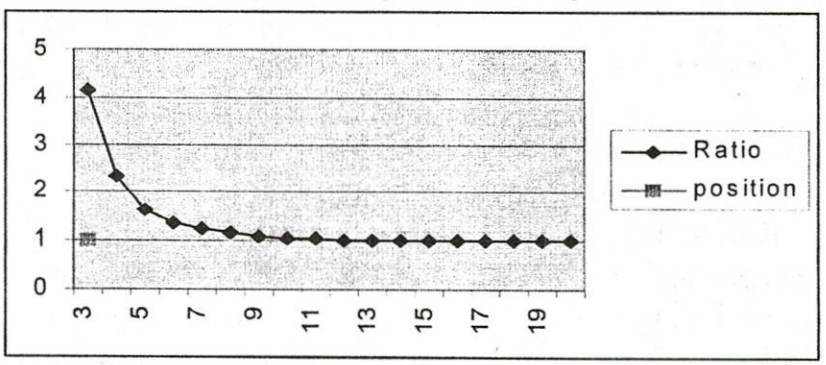
Chúng tôi đã thống kê trên danh sách 109.582 từ tiếng Anh theo số ký tự khác biệt trung bình tại vị trí p tương ứng trong từ tiếng Anh (bảng 3) và nhận thấy:

- Nếu chỉ căn cứ vào 3 ký tự đầu tiên, thì ta sẽ có 2991 mục từ phân biệt.
- Sau vị trí ký tự thứ 3 (gốc cây DST) sẽ có trung bình là 4.16 nhánh.
- Số nhánh (tại một vị trí) sẽ giảm nhanh khi chiều dài tăng. Sau vị trí ký tự thứ 8, hầu như tại mỗi nút trên cây DST ta chỉ còn có trung bình 1 nhánh.

Bảng 3: khảo sát số nhánh ký tự khác nhau theo vị trí trong từ:

Vị trí	Số từ phân biệt	Hệ số nhánh R_i	Vị trí	Số từ phân biệt	Hệ số nhánh R_i
1	26	14.69231	6	46585	1.373253
2	382	7.829843	7	63973	1.237178
3	**2991**	**4.162822**	8	79146	1.148839
4	12451	2.323508	9	90926	1.080769
5	28930	1.610266	10	98270	1.050249

Xét một cây DST (cao 9 bậc) gốc tại vị trí ký tự thứ 4 của từ ,số lá trung bình sẽ là:



90.926 / 2991 = 30

Hình 2: Số nhánh (ký tự) trung bình tại mỗi vị trí

- Tổng số nhánh (tính trung bình) của một cây DST (2):

$$N = \sum_{i=3}^{9} \prod_{i=3}^{9} R_i * R_{i+1} \approx 100$$

3.1.4 Khảo sát môi trường và phương pháp của hệ dịch:

Phương pháp dịch của Hệ dịch máy Anh – Việt được thử nghiệm ở đây là phương pháp chuyển đổi cú pháp kết hợp với bộ phân giải ngữ nghĩa nông (syntactical transfer + shallow semantic analysis). Môi trường hệ điều hành của phần mềm dịch này là môi trường Windows 32 bit (Win95 , 97, 98) và chạy trên máy IBM PC AT 486 trở lên với cấu hình từ 8 Mb RAM trở lên, cần khoảng 50 Mb trống của đĩa cứng.

Với phương pháp dịch nêu trên, nên từ điển sẽ chứa rất nhiều thông tin về ngữ nghĩa, ngữ pháp,.. ở trong phần nội dung của mục từ, và những thông tin này phải tuân theo một cấu trúc chặt chẽ để cho chương trình có thể sử dụng nó trong quá trình dịch.

3.2 GIẢI PHÁP TỐI ƯU CHO TỪ ĐIỂN

Sau khi xem xét các ưu-nhược điểm của các cách thức cài đặt dữ liệu nêu trên cùng với các khảo sát thực tế dữ liệu từ điển cũng như nhu cầu và môi trường hệ dịch nêu trên, chúng tôi đã chọn giải pháp cài đặt từ điển như sau:

- Dùng kỹ thuật bảng băm cho 3 ký tự đầu tiên của từ tiếng Anh: vì chỉ có 3 ký tự đầu, nên không gian lưu trữ không tốn bao nhiêu (ta dành hẳn bảng băm có kích thước cố định là 27*27*27 phần tử * 4 bytes ≈ 76 Kb), nhưng ta sẽ có tốc độ truy cập cực nhanh ($\theta(1)$) cho các từ có chiều dài ngắn dưới 4 ký tự.
- Dùng kỹ thuật Cây tìm kiếm số (DST) cho các ký tự còn lại. Thời gian tìm kiếm phụ thuộc vào chiều dài còn lại của từ (sau khi trừ đi 3 ký tự, $\theta(k-3)$ với $k > 3$). Sau khi đi đến cuối từ của khóa, ta sẽ có con trỏ trỏ đến phần nội dung của từ đó.
- Vì tệp khóa (.IDX) do có dung lượng nhỏ và truy cập thường xuyên nên chúng tôi để trên bộ nhớ chính (RAM) để truy cập nhanh, còn tệp nội dung (.DAT) có khối lượng lớn, sẽ để trên bộ nhớ phụ (đĩa cứng). Đối với những từ vừa mới truy cập trước đó, do đặc tính nội tại của hệ thống máy tính hiện nay, thì nội dung đó sẽ tự động được đệm lại trong bộ nhớ đệm (cache).

Chúng tôi chọn giải pháp trên là vì: theo kết quả thống kê các từ tiếng Anh ở phần trên, những từ có tần suất xuất hiện cao nhất thì thường có chiều dài nhỏ, chính vì vậy, với cách cài đặt trên, thì những từ có tần suất cao sẽ được truy cập nhanh. Khi mở rộng từ điển, tốc độ truy cập giảm không đáng kể và yếu tố không gian lưu trữ vẫn đảm bảo.

3.3 CẤU TRÚC DỮ LIỆU CỦA TỪ ĐIỂN

3.3.1 Cấu trúc bảng băm :

Ta có 26 mẫu tự tiếng Anh (từ a – z), từ ký tự thứ 2 trở đi, ta còn có thể có thêm ký tự khoảng trắng (space), vậy ta dùng tất cả 26*27*27 =18.954 điểm vào (entry) của bảng băm. Trong số 18.954 điểm vào, chỉ có tối đa khoảng 3.000 điểm vào thật sự có trỏ tới gốc cây DST và/hoặc nội dung của từ, còn tất cả số còn lại là không có thật (null). Trong số 3.000 điểm vào khả dĩ đó, sẽ có tối đa khoảng 1.000 từ kết thúc ngay (nghĩa là độ dài từ $k \leq 3$), không cần đi tiếp. Chính vì vậy, ta cần một mã (code) để đánh dấu đặc tính này của mỗi điểm vào. Độ sâu của cây DST sẽ không quá 25 mức (ta căn cứ theo từ tiếng Anh dài nhất trừ đi 3 ký tự đầu).

Hình 3: Cấu trúc bảng Băm cho 3 ký tự đầu của từ

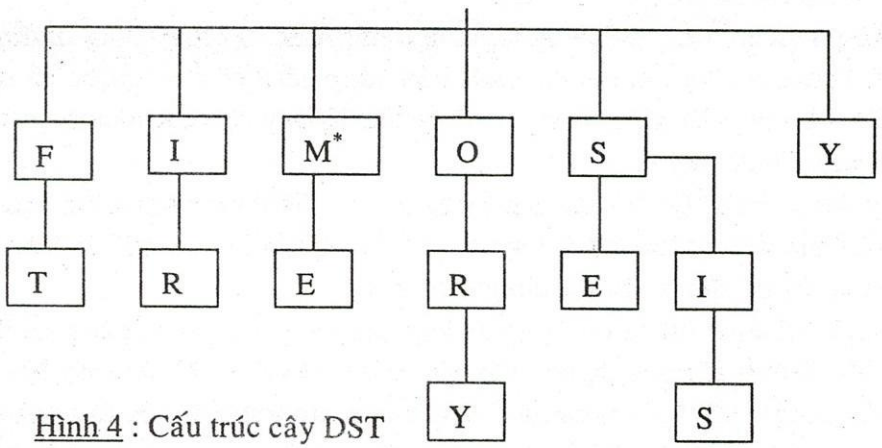
1st	2nd	3rd	code	Pointer
A			0	
A	a		1	null
A	b	z		
T	h	e	1	
Z	z	z		

Nội dung DST của từ "a"

Gốc cây DST của từ "the"

3.3.2 Cấu trúc cây tìm kiếm số (DST):

Cây DST sau chuỗi 3 ký tự "the" sẽ có dạng như dưới đây:



Hình 4 : Cấu trúc cây DST

Các nút lá hay các nút có dấu * cho biết đây là kết thúc của một từ, khi duyệt cây trên ta sẽ có các từ sau : the*ft*, the*ir*, the*m*, the*me*, the*ory*, the*se*, the*sis*, the*y*.

3.3.3 Cài đặt hàm băm (3):

$$f(key) = \sum_{i=0}^{2} (key[i] \,\&\, 0x1F) * 27^i$$

4. KẾT LUẬN

Trên đây chúng tôi đã trình bày về nhu cầu và tổ chức một từ điển điện tử nhằm phục vụ cho hệ dịch máy Anh – Việt được hiệu quả. Trước khi quyết định chọn thông tin nào sẽ được lưu trong từ điển và lưu như thế nào, chúng tôi đã tiến hành khảo sát chi tiết nhu cầu của hệ dịch: vì đây là hệ dịch dựa trên cơ sở cú pháp và tri thức từ, nên trong cấu trúc vi mô của từ, chúng tôi phải chứa đầy đủ những thông tin cần thiết cho quá trình dịch và điều khiển quá trình dịch (lexicon – driven translation). Ngoài ra, chúng tôi cũng khảo sát đặc điểm vốn từ tiếng Anh: về tần số sử dụng, chiều dài, phân bố ký tự, ... và từ đó quyết định chọn phương pháp bảng băm cho 3 ký tự đầu và cây DST cho các ký tự còn lại. Giải pháp này phù hợp về nhu cầu tra cứu nhanh những từ ngắn, những từ hay sử dụng (trung bình 0.02 giây/từ trên máy IBM Pentium-II 350 MHz) và không tốn nhiều bộ nhớ (300 Kb cho tệp khóa của 15.000 từ).

Hạn chế của từ điển hiện nay là chưa cho phép người sử dụng tùy ý cập nhập thông tin vào từ điển hệ thống, chỉ cho phép cập nhập từ điển người sử dụng (chứa các thông tin về tên

riêng, từ viết tắt, từ chuyên ngành). Trong tương lai chúng tôi sẽ có phần giám sát chặt chẽ để cho phép người sử dụng cập nhật từ điển hệ thống ở một mức độ nào đó.

THE ORGANIZATION OF ELECTRONIC DICTIONARY DATA FOR THE ENGLISH-VIETNAMESE MACHINE TRANSLATION

Dinh Dien
Department of Information Technology, University of Natural Sciences – VNU-HCM
(Received 04 November 2001, Revised 22 January 2002)

ABSTRACT: In a Machine Translation system, the most important database is its electronic dictionary. So, the main contents of this paper is to organize the data in an electronic dictionary of the English-to-Vietnamese Machine Translation System. The choice and organization of data in an electronic dictionary for an MT system totally depends on the methodology and strategy of translation of each system, and in generally, this choice of data differs from that of dictionary for human.

TÀI LIỆU THAM KHẢO

[1] Chương, Nguyễn Hồng. *Cấu trúc dữ liệu: ứng dụng và cài đặt bằng C*. NXB TP HCM (1999).
[2] Hutchins, John and Somer, Harold L. *An Introduction to Machine translation*. Academic Press (1992).
[3] Khang, Nguyễn Văn. *Từ điển bậc thang Anh-Việt*. Viện Ngôn ngữ học. NXB Thế Giới (1994).
[4] Miller, George A: *Introduction to WordNet*. http://www.cogsci.princeton.edu/~wn/. Princeton (1996).
[5] Smith, George W. *Computers and Human Language*. Oxford University Press. (1991).
[6] Trực, Nguyễn Trung. *Cấu trúc dữ liệu*. Khoa CNTT, ĐH BK. TPHCM (1998)
[7] Viện Ngôn ngữ học. *Một số vấn đề Từ điển học*. NXB KH xã hội. Hà Nội (1997).