Tạp chí Tin học và Điều khiển học, T. 17, S. 2 (2001), 82–86

CHUẨN HÓA VĂN BẢN CHỮ VIỆT SOẠN THẢO TRONG WORD

CAO ĐÌNH THI

Abstract. In this paper, the algorithms for standardization of documents compiling in Vietnamese by the fonts of ABC are presented.

Tóm tắt. Bài này trình bày những thuật toán về chuẩn hóa văn bản chữ Việt soạn thảo bằng các font chữ thường của bộ ABC trong Word.

1. MỞ ĐẦU

Hiện nay vấn đề xử lý dữ liệu trong văn bản chữ Việt nói chung, chuẩn hóa văn bản chữ Việt soạn thảo trong Word nói riêng đã và đang được rất nhiều người quan tâm. Do ở nước ta hiện nay tồn tại cùng một lúc nhiều bộ chương trình phần mềm không đồng nhất để soạn thảo chữ Việt như ABC, VNI, VietWare,... nên việc tạo ra một chương trình để xử lý chữ Việt (chuẩn hóa, sắp xếp, tìm kiếm...) chung cho tất cả các bộ chương trình này còn gặp nhiều khó khăn. Để thống nhất bộ mã chung cho việc soạn thảo tiếng Việt trên máy tính, Bộ Khoa học Công nghệ và Môi trường đã chọn các font chữ của bộ ABC làm chuẩn Quốc gia. Bộ ABC này đã được sử dụng chính thức trong các cơ quan Đảng và Chính phủ Việt Nam. Từ đó đến nay nó đã được sử dụng để soạn thảo, lưu giữ và in ấn rất nhiều văn bản, tài liệu ở trong và ngoài nước. Vì vậy, việc xây dựng các chương trình phần mềm phục vụ cho việc chuẩn hóa và sắp xếp dữ liệu chữ Việt theo các font chữ của bộ ABC là vô cùng cần thiết.

Vấn đề chuẩn hóa và sắp xếp dữ liệu chữ Việt trong FoxPro và trong Excel đã được giải quyết trong các tài liệu [1]-[5]. Ở bài này, chúng ta sẽ nghiên cứu việc chuẩn hóa văn bản soạn thảo trong Word bằng các font chữ của bộ ABC. Bộ ABC có ưu điểm là dễ soạn thảo, font chữ đẹp nhưng do các chữ thường và các chữ hoa không cùng trong một font chữ như các bộ mã khác (trong bộ ABC tên font chữ dành cho chữ hoa bắt đầu bằng ".Vn" và kết thúc bằng "H", các chữ thường tương ứng ở các font chữ không có chữ H ở cuối, ví dụ: .VnTimeH và .VnTime, .VnArialH và .VnArial,...) nên cũng tạo ra một số bất tiện cho người sử dụng.

Trước hết, nếu ta soạn thảo văn bản ở font chữ thường, ví dụ ".VnTime", mà ta muốn gõ từ bàn phím các chữ hoa đối với các nguyên âm như a, ă, â, e, ê, o, ô, ơ, u, ư, với các dấu thanh, ví dụ, ta muốn gõ chữ hoa đối với các ký tự đầu của các từ "áp", "ất", "ếch", "ót", "ồng", "ửng",... thì chắc chắn chúng ta không đạt được mục đích. Đây lại là thao tác thường xuyên chúng ta phải làm (sau dấu chấm câu bắt buộc chúng ta phải viết hoa ký tự đầu tiên của từ đầu câu). Để khắc phục điều này người ta thường phải đánh dấu (bôi đen) các ký tự đầu của từ đó rồi chọn font chữ ".VnTimeH" để chuyển chúng sang chữ hoa. Rõ ràng, đây là điều rất bất tiện khi soạn thảo bằng font chữ ABC.

Ngoài ra, nếu đối với font chữ thường và font chữ hoa của bộ mã này chúng ta chọn cùng một cỡ chữ thì font chữ hoa sẽ to hơn, khoảng cách ở các dòng có các từ chuyển đổi sang font chữ hoa sẽ bị đẩy đi rộng hơn, làm cho khoảng cách giữa các dòng của văn bản không đồng đều. Thông thường chúng ta phải chọn cỡ chữ cho font chữ hoa nhỏ hơn cỡ chữ ở font chữ thường một đơn vị thì khoảng cách dòng mới tương đối đồng đều.

Việc soạn thảo trong bất kỳ một ngôn ngữ nào cũng phải tuân theo những luật nhất định. Những luật ta thường gặp là: giữa các từ của một câu trong một văn bản phải có một dấu cách; sau dấu chấm phải viết hoa ký tự đầu tiên của từ đầu câu... Trước đây, trong một số ngôn ngữ như tiếng Pháp chẳng hạn, người ta qui định các dấu phẩy (,), dấu chấm phẩy (;), hai chấm (:),

dấu chấm câu (.),... không được viết liền với từ ngay trước đó (trước các dấu này phải có một dấu cách). Ngày nay, khi Tin học phát triển những qui định này lại là bất tiện khi soạn thảo văn bản. Nếu những dấu này không đi liền với từ trước đó thì chúng ta sẽ gặp những dòng chữ bắt đầu bằng một trong các dấu trên vì máy tính đã tự động chỉnh dòng văn bản. Khi đã hết dòng trên, nếu các dấu đi liền với từ trước đó thì máy sẽ chuyển cả từ này và dấu xuống dòng dưới, nếu dấu không đi liền với từ thì chỉ có dấu bị chuyển xuống dòng dưới mà thôi. Trong một số văn bản, bài báo chữ Việt, ta cũng thấy có hiện tượng này. Hiện tượng như vậy không những làm mất ý nghĩa của các dấu này mà còn làm cho người đọc rất khó chịu. Trong tiếng Anh thì các dấu này bắt buộc phải đi liền với từ, nếu không, khi kiểm tra, máy sẽ báo lỗi.

Bài này trình bày một số thuật toán mà theo đó một chương trình đã được lập để chuẩn hóa văn bản chữ Việt soạn thảo bằng các font chữ thường của bộ ABC trong Word. Chúng ta tạm gọi một văn bản chuẩn ở đây là một văn bản mà trong đó

- Trong một câu, mỗi từ cách nhau ít nhất một dấu cách (một ký tự trống);
- Giữa hai câu trong cùng một đoạn văn (pagagraph) được ngăn cách bởi dấu chấm và một dấu cách;
- Ký tự đầu tiên của từ đầu câu phải viết hoa;
- Các loại dấu như dấu phẩy ",", dấu chấm phẩy ";", hai chấm ":", ngoặc đơn đóng ")", ngoặc vuông đóng "]", ngoặc nhọn đóng "}" phải đi liền với từ kề ngay trước đó và sau các dấu này phải có ít nhất một dấu cách.
- Các dấu ngoặc đơn mở "(", ngoặc vuông mở "[", ngoặc nhọn mở "{" (ta tạm gọi là các dấu mở) phải đi liền với từ kề ngay sau đó và trước các dấu này phải có một dấu cách;
- Các dấu chấm ngăn cách phần tỉ, phần triệu, phần nghìn, dấu phẩy ngăn cách phần thập phân trong các chữ số phải đi liền với số đứng trước và số đứng sau, không được có dấu cách.

Chương trình được viết bằng Visual Basic sẽ tự động chuẩn hóa văn bản theo các tiêu chuẩn trên, nghĩa là, khi chạy chương trình này, máy sẽ tự động chuyển các dấu về đúng vị trí của nó và biến đổi những ký tự đầu tiên của những từ tiếng Việt có dấu hoặc không dấu từ font chữ thường sang font chữ hoa nếu các từ này là những từ đầu câu. Như vậy, đối với những từ ở đầu câu có ký tự đầu tiên là các nguyên âm thuần Việt với các dấu thanh hoặc các phụ âm, không cần phải đánh dấu rồi chuyển về font chữ hoa, chúng ta vẫn có được chữ hoa do chương trình đã tự động chuyển đổi từ font chữ thường sang font chữ hoa. Đây là ưu điểm và là mục tiêu chính của chương trình này.

2. NỘI DUNG CHƯƠNG TRÌNH

Chương trình chuẩn hóa được trình bày dưới dạng một macro. Người sử dụng có thể tạo macro rồi soạn thảo chương trình (hoặc copy vào macro nếu chương trình được lưu giữ dưới dạng một tệp văn bản) để sử dụng. Khi chạy chương trình này ta chỉ cần khai báo font chữ thường (có .Vn ở trước) và cỡ chữ rồi làm theo chỉ dẫn là máy sẽ tự động chuẩn hóa văn bản cho chúng ta.

Trong chương trình chúng ta sẽ sử dụng một xâu gồm 103 ký tự (93 ký tự chữ Việt và 10 ký tự dành cho chữ số) như sau:

xau = "aàảãáạăằẳẵắặâầẩẫấậbcdđeèẻẽéẹêềểễếệfghiìỉĩíịjklmnoòỏõóọôồổỗốộơờởỡớợpqrstuùủũúụ
ưừửữứựvwyỳỷỹýỵz0123456789"

Trong xâu ký tự trên có một số ký tự ít được dùng trong chữ Việt như f, j, w, z nhưng ta vẫn đưa vào để dùng trong trường hợp các từ phiên âm tiếng nước ngoài [6].

Kĩ thuật chính được sử dụng trong trương trình này là kỹ thuật tìm kiếm và thay thế ký tự ở font chữ thường bằng font chữ hoa. Riêng việc điều chỉnh các dấu trong văn bản (trừ các trường hợp dấu chấm ngăn cách phần tỉ, phần triệu, phần nghìn, dấu phẩy ngăn cách phần thập phân trong các chữ số) chỉ làm trong font chữ thường.

Các giai đoạn chính của chương trình có thể mô tả như sau:

- Trước hết chúng ta xử lý các dấu chấm, dấu phẩy đối với các chữ số trong trường hợp tồn tại dấu chấm để ngăn cách phần tỉ, phần triệu, phần nghìn, và dấu phẩy để ngăn cách phần thập phân của các số trong văn bản. Các dấu chấm, dấu phẩy ở đây đi liền ký tự trước và ký tự sau, không được có dấu cách. Trong trường hợp này chúng ta chỉ sử dụng các ký tự từ thứ 94 đến ký tự thứ 103 trong xâu ký tự ở trên để lập trình xử lý bằng cách tìm kiếm và thay thế chuỗi ký tự bao gồm ký tự trước dấu chấm (dấu phẩy), dấu chấm (dấu phẩy), ký tự sau các dấu này từ font chữ thường sang font chữ hoa với cỡ chữ ở font chữ hoa nhỏ hơn cỡ ở font chữ thường 1 đơn vị. Ta phải làm việc này trước để khi chỉnh các dấu trong văn bản, các dấu chấm, dấu phẩy đi kèm theo các chữ số sẽ không bị chỉnh, sửa lại vì chúng đã được chuyển sang font chữ hoa mà ở phần sau, khi chỉnh các dấu (bao gồm cả dấu chấm và dấu phẩy) chúng ta chỉ thực hiện trong font chữ thường.
- Tiếp theo chúng ta chỉnh các loại dấu về đúng vị trí của nó. Ở đây cần phân biệt 2 loại dấu. Một loại bao gồm các dấu mở và loại bao gồm các dấu còn lại.

  Đối với loại bao gồm các dấu mở, luật điều chỉnh có khác một chút so với loại bao gồm các dấu còn lại. Các dấu mở phải đi liền với từ kề ngay sau đó (chứ không phải đi liền với từ kề trước như trong các trường hợp của các dấu còn lại) và trước các dấu này phải có một ký tự trống. Do đó nếu các dấu mở chưa đi liền với từ kề ngay sau đó thì ta phải "dồn" chúng lại; nếu trước các dấu này chưa có ký tự trống thì phải "giãn" chúng ra.

  Đối với các dấu thuộc loại không bao gồm các dấu mở ta phải điều chỉnh sao cho các dấu này phải đi liền với từ kề ngay trước đó và sau các dấu này phải có 1 ký tự trống. Do đó nếu các dấu này chưa đi liền với từ kề trước đó thì phải "dồn" chúng lại. Nếu sau dấu này chưa có một ký tự trống thì phải "giãn" chúng ra. Nếu sau các dấu này có hơn 1 ký tự trống thì phải "dồn" lại sao cho chúng chỉ còn 1 ký tự trống mà thôi. Trong suốt quá trình này ta làm trong font chữ thường để khỏi ảnh hưởng tới các dấu chấm, dấu phẩy trong các chữ số ở phần trên.
- Cuối cùng chúng ta sẽ xử lý đối với các từ đầu câu trong văn bản. Ký tự đầu tiên của các từ này bắt buộc phải viết hoa bất kể ký tự này là ký tự Việt có dấu hoặc không dấu. Dựa vào luật chỉnh dấu chấm ở phần trên (sau dấu chấm phải có một dấu cách) và xâu ký tự chữ Việt ta tìm kiếm và thay thế các ký tự ở font chữ thường sang font chữ hoa tương ứng, với cỡ chữ ở font chữ hoa nhỏ hơn cỡ chữ ở font chữ thường 1 đơn vị.

Có thể mô tả tóm tắt nội dung chương trình và các thuật toán như sau:

Khai báo font chữ và cỡ chữ của đoạn văn bản cần chuẩn hóa.

Nhập xâu ký tự

xau = "aàảãáạăằẳẵắặâầẩẫấậbcdđeèẻẽéẹêềểễếệfghiìỉĩíịjklmnoòỏõóọôồổỗốộơờởỡớợpqrstuùủũúụ
ưừửữứựvwyỳỷỹýỵyz0123456789"

Tính độ dài của xâu ký tự: Dd=length(xau).

Trong trường hợp này Dd=103.

Gán các ký tự của xâu này vào một mảng M(i), i = 1, 2,...,103.

```
For i=1 to Dd do
    M(i)=Mid(xau,i,1)
```

Thuật toán xử lý đối với các dấu chấm, dấu phẩy trong các số:

```
For i=94 to 103 do
    For j=94 to 103 do
        {Tìm kiếm và thay thế biểu thức ký tự (M(i)&","&M(j)) ở font chữ thường bằng biểu
         thức (M(i)&","&M(j)) ở font chữ hoa}.
        {Tìm kiếm và thay thế biểu thức ký tự (M(i)&"."&(j)) ở font chữ thường bằng biểu
         thức (M(i)&"."&M(j)) ở font chữ hoa}.
```

Thuật toán xử lý đối với các loại dấu trong văn bản:

Gán các dấu vào một mảng C(i), i = 1, 2,...,12.

C(1)=".", C(2)=",", C(3)=";", C(4)=")", C(5)=":", C(6)="?", C(7)="]",
C(8)="}", C(9)="-", C(10)="[", C(11)= "(", C(12)="{".

Xử lý đối với loại các dấu không bao gồm các dấu mở:

For k=1 **to** 9 **do**

{Tìm kiếm và thay thế C(k) bằng C(k)&" " (giữa hai dấu nháy kép là 1 ký tự trống) để xử lý trường hợp sau các dấu không có ký tự trống. Nếu đã có rồi thì số ký tự trống sẽ tăng thêm 1. Sau đó xử lý các ký tự trống thừa}.

Repeat

{Tìm kiếm và thay thế C(k)&" " (giữa 2 dấu nháy kép là 2 ký tự trống) bằng C(k) &" " (giữa 2 dấu nháy kép là 1 ký tự trống)}

Until {đến khi chỉ còn C(k)&" " (giữa 2 dấu nháp kép là 1 ký tự trống)}.

Repeat

{Tìm kiếm và thay thế " "&C(k) bằng C(k) (giữa 2 dấu nháy kép là 1 ký tự trống)}

Until {đến khi nào kết ký tự trống}.

Xử lý đối với các dấu mở:

For j=10 **to** 12 **do**

{Tìm kiếm thay thế C(j) bằng " "&C(j) (giữa hai dấu nháy kép là 1 ký tự trống)}.

Repeat

{Tìm kiếm và thay thế " "&C(j)&" " (giữa 2 dấu nháy kép trước là 2 ký tự trống, giữa 2 dấu nháy kép sau là 1 ký tự trống) bằng " "&C(j) (giữa 2 dấu nháy kép là 1 ký tự trống)}.

Until {đến khi nào trước các dấu mở chỉ còn 1 ký tự trống, sau các dấu này không còn ký tự trống}.

Việc tìm kiếm và thay thế trong thuật toán này chỉ làm trong font chữ thường.

Đối với các dấu nháy đơn (') và nháy kép (") vì dấu đóng và dấu mở như nhau nên ta phải xác định trong từng trường hợp cụ thể xem đó là dấu gì (đóng hay mở). Tính từ đầu văn bản, nếu lần xuất hiện hiện thời của các dấu này là số lẻ thì đó là các dấu mở, nếu là số chẵn thì đó là các dấu đóng. Phụ thuộc vào trạng thái dấu đó là dấu mở hay dấu đóng mà ta sử dụng các thuật toán trên để đưa chúng về đúng vị trí cần thiết.

Thuật toán xử lý với ký tự đầu tiên của các chữ đầu câu:

Chọn font chữ, cỡ chữ cho tìm kiếm. Đây thực chất là font chữ, cỡ chữ của đoạn văn bản cần chuẩn hóa mà ta phải khai báo từ đầu khi chạy chương trình.

Chọn font chữ thay thế là font chữ hoa tương ứng, cỡ chữ thay thế kém cỡ chữ tìm kiếm 1 đơn vị.

For i=1 **to** Dd-10 **do**

{Tìm kiếm thay thế biểu thức (".”&" "&M(i)) (giữa 2 dấu nháy kép sau là 1 ký tự trống) ở font chữ thường bằng biểu thức (".”&" "&M(i)) (giữa 2 dấu nháy kép sau là 1 ký tự trống) ở font chữ hoa}.

3. CÁCH TẠO VÀ SỬ DỤNG MACRO

Muốn tạo một macro sử dụng chương trình này ta làm như sau:

- Mở một tệp văn bản bất kỳ (kể cả một băn bản trắng khi vào *File, New*);
- Nhấn chuột vào *Tools* trên thanh Menu; chọn *Macro, Record New Macro.* Trong cửa sổ *Record Macro* ở mục *Macro name* gõ từ bàn phím một tên cho macro, ví dụ ChuanWord chẳng hạn; nhấn chuột vào biểu tượng Keyboard ở mục *Assign macro to* để chọn một phím nóng gán cho macro này, ví dụ F11 (ấn vào phím F11) rồi nhấn chuột tiếp vào *Assign, Close*;
- Trở về *Tools, Macro, Stop Reccording*;
- Vào lại *Tools, Macro.* Trong cửa sổ *Macro name* chọn tên macro vừa tạo ra (ở đây là ChuanWord), *Edit.* Khi đó ta sẽ thấy có các dòng lệnh:

```
Sub ChuanWord()
'ChuanWord Macro
'Macro recorded ... by ...
'
End Sub
```

Phụ thuộc vào ngày bạn khởi tạo macro và tên trong máy tính của bạn mà máy sẽ tự động gán thêm các thông số vào các vị trí 3 dấu chấm (...). Nội dung của chương trình bạn phải soạn thảo hoặc copy vào khoảng giữa của 2 dòng lệnh 'Macro recorded ... by ... và End Sub.

Sau khi soạn thảo hoặc copy chương trình xong, nhấn Ctrl S để ghi lại và nhấn chuột vào *File, Close and Return to Microsoft Word* để trở về văn bản.

Muốn tìm hiểu nội dung chi tiết của chương trình, độc giả có thể liên hệ với tác giả theo các địa chỉ E-mail sau:

Cdthi@ioit.ncst.ac.vn hoặc Cdthi@cfvghn.org.vn

Khi đã tạo được macro này rồi thì việc sử dụng nó để chuẩn hóa cho một tệp văn bản bất kỳ (được soạn thảo trong Word bằng các font chữ thường của bộ ABC) trở thành rất đơn giản. Ta chỉ cần mở tệp văn bản cần chuẩn hóa; sau đó nhấn phím F11 rồi làm theo các chỉ dẫn của chương trình. Nếu trong văn bản có nhiều đoạn với các font chữ thường khác nhau, cỡ chữ khác nhau thì mỗi lần chạy chương trình ta sẽ chuẩn hóa được những đoạn có font chữ và cỡ chữ đã khai báo (kể cả trường hợp các đoạn văn bản này không đi liền nhau, nằm lẻ tẻ trong văn bản miễn là chúng có font chữ và cỡ chữ trùng với font chữ và cỡ chữ đã khai báo). Muốn chuẩn hóa tiếp cho các đoạn văn bản với các font chữ và cỡ chữ khác ta chạy lại macro và khai báo lại font chữ và cỡ chữ.

Kỹ thuật lập trình này có thể áp dụng để xây dựng các chương trình chuẩn hóa với các bộ khác của chữ Việt và các ngôn ngữ khác sử dụng ký tự của hệ chữ La tinh. Đối với các bộ chữ mà chữ thường và chữ hoa ở cùng một font chữ thì chương trình sẽ đơn giản hơn rất nhiều.

TÀI LIỆU THAM KHẢO

[1] Cao Đình Thi, Chương trình sắp xếp dữ liệu chữ Việt trong phông (font) chữ thường, *Tin học Ngân hàng* số 6 (1998) 20–22.

[2] Cao Đình Thi, Chuẩn hóa dữ liệu và sắp xếp chữ Việt font chữ hoa, *Tin học Ngân hàng* số 2 (1999) 27–32.

[3] Vũ Văn Thái, Cao Đình Thi, Chuẩn hóa dữ liệu trong Excel, *Tin học Ngân hàng* số 5 (1999) 22–25.

[4] Cao Đình Thi, Vũ Văn Thái, Sắp xếp dữ liệu chữ Việt Excel, *Tin học Ngân hàng* số 4 (2000) 28–30.

[5] Cao Đình Thi, Về chương trình sắp xếp trường Họ và tên trong FoxPro, *Tin học và Đời sống* số 8 (2000) 56–57.

[6] *Từ điển Bách khoa Việt Nam* I, Trụng tâm Biên soạn từ điển Bách khoa Việt Nam, Hà Nôi, 1995.

Nhận bài ngày 8 tháng 9 năm 2000
Nhận bài sau khi sửa ngày 20 tháng 10 năm 2000

Viện Công nghệ thông tin