

NGHIÊN CỨU THU THẬP VÀ XÂY DỰNG CƠ SỞ DỮ LIỆU CHỮ VIẾT TẮT TIẾNG VIỆT

COLLECTING AND BUILDING AN VIETNAMESE ABBREVIATION DATABASE

Huỳnh Công Pháp¹, Nguyễn Văn Huệ²

¹Đại học Đà Nẵng; Email: hcphap@gmail.com

²Công ty phần mềm Enclave; Email: nguyenhuebkd@gmail.com

Tóm tắt - Chữ viết tắt trong tiếng Việt ngày càng tăng lên đáng kể về số lượng, đa dạng về ký hiệu, nhiều chữ viết tắt có nhiều nghĩa khác nhau. Điều này đã dẫn đến một thực trạng là làm cho người đọc văn bản nhiều lúc hiểu nhầm nội dung hoặc khó có thể đoán ra được nghĩa của từ viết tắt. Tuy nhiên, hiện nay chúng ta vẫn chưa tìm thấy một hệ thống tra cứu chữ viết tắt tiếng Việt. Để xây dựng được hệ thống tra cứu chữ viết tắt cũng như công cụ hỗ trợ gõ tắt, bước đầu tiên là cần phải xây dựng được cơ sở dữ liệu chữ viết tắt tiếng Việt. Trong bài báo này chúng tôi tập trung nghiên cứu thu thập tự động và xây dựng một cơ sở dữ liệu tương đối đầy đủ về chữ viết tắt tiếng Việt. Trên cơ sở đó, chúng tôi sẽ tiến đến xây dựng một hệ thống quản lý và tra cứu chữ viết tắt tiếng Việt trực tuyến nhằm đáp ứng nhu cầu của đông đảo người sử dụng. Bên cạnh đó, cũng từ cơ sở dữ liệu này, chúng tôi cũng sẽ tiến hành nghiên cứu xây dựng một công cụ tích hợp trên môi trường Windows (giống Vietkey) nhằm hỗ trợ tra cứu và soạn thảo chữ viết tắt.

Từ khóa - chữ viết tắt; từ điển chữ viết tắt; trích rút văn bản; xử lý tiếng Việt; cơ sở dữ liệu chữ viết tắt; hệ thống tra cứu chữ viết tắt;

Abstract - Vietnamese abbreviations increase very fast, diversify in forms and some of them have multiple meanings. This poses a problem for readers to recognize abbreviations or to understand the relevant meaning in some situation. However, we currently can't still find out a system of vietnamese abbreviation consultation. To have such a system, the first step we should build a vietnamese abbreviation database. In this paper, we focus on the research of acquiring vietnamese abbreviations from documents and the internet to build an abbreviation database. From this database, we aim to propose an online system of abbreviation management and consultation as well as a "hooked" software (like Vietkey) supporting autotext when typing.

Key words - abbreviation; acronym; abbreviation dictionary; text extraction; vietnamese language processing; vietnamese abbreviation consultation system;

1. Đặt vấn đề

Hiện tượng viết tắt trong các văn bản tiếng Việt trở nên rất phổ biến. Số lượng chữ viết tắt tiếng Việt hiện nay tăng lên đáng kể và dưới nhiều hình thức rất đa dạng. Điều này đã dẫn đến một thực trạng là làm cho người đọc văn bản nhiều lúc hiểu nhầm nội dung hoặc khó có thể đoán ra được nghĩa của từ viết tắt.

Thật vậy, do sự phát triển của xã hội và sự hội nhập quốc tế về khoa học kỹ thuật, mỗi ngày trong tiếng Việt xuất hiện thêm nhiều chữ viết tắt mới. Quy tắc hình thành chữ viết tắt cũng rất đa dạng, nhiều chữ viết tắt trong tiếng Việt được hình thành từ các từ có nguồn gốc tiếng nước ngoài (ví dụ: ngành IT - ngành Công nghệ thông tin), nhiều chữ viết tắt được cấu thành từ các chữ cái đầu (ví dụ: ĐHQĐ - Đại học Đà Nẵng), tuy nhiên cũng có nhiều chữ viết tắt được cấu thành từ hai ký tự của một từ (ví dụ: vina - Việt Nam)...

Trước thực trạng vừa nêu như trên thì nhu cầu tra cứu chữ viết tắt đối với người sử dụng là rất cấp thiết và thường xuyên. Tuy nhiên, hiện nay các kết quả nghiên cứu thu thập chữ viết tắt tiếng Việt [1] không được chia sẻ trực tuyến. Do đó chúng ta chưa thật sự có từ điển hay cơ sở dữ liệu đầy đủ về chữ viết tắt tiếng Việt, nhằm phục vụ nhu cầu tra cứu, học tập và hỗ trợ trong vấn đề đọc và soạn thảo văn bản của người sử dụng.

Trong bài báo này, chúng tôi tập trung nghiên cứu thu thập tự động và xây dựng một cơ sở dữ liệu tương đối đầy đủ về chữ viết tắt tiếng Việt. Trên cơ sở đó, chúng tôi sẽ tiến đến xây dựng một hệ thống quản lý và tra cứu chữ viết tắt tiếng Việt trực tuyến nhằm đáp ứng nhu cầu của đông đảo người sử dụng. Bên cạnh đó, cũng từ cơ sở dữ liệu này,

chúng tôi cũng sẽ tiến hành nghiên cứu xây dựng một công cụ tích hợp trên môi trường Windows (giống Vietkey) nhằm hỗ trợ tra cứu và soạn thảo chữ viết tắt.

2. Một số nghiên cứu tổng quan

Liên quan đến vấn đề thu thập, xử lý và quản lý chữ viết tắt tiếng Việt, chúng ta có thể kể đến công trình của tác giả Bạch Như Nguyễn "Nghiên cứu tích hợp môi trường, công cụ và kỹ thuật tìm kiếm, quản lý chữ viết tắt" (2007)[3]. Tuy nhiên, công trình này cũng chỉ dừng lại ở mức nghiên cứu và chưa có dịch vụ trực tuyến hay miễn phí phục vụ cho cộng đồng người sử dụng.

Đối với hỗ trợ tra cứu và xử lý chữ viết tắt, các phần mềm như MS Word cung cấp chức năng "AutoCorrect Options" cho phép định nghĩa chữ viết tắt để phục vụ gõ tắt. Tuy nhiên, dữ liệu chữ viết tắt này không được chia sẻ và sử dụng trong các môi trường khác.

Đối với hệ thống quản lý và tra cứu chữ viết tắt, chúng ta có thể tìm thấy hệ thống tra cứu từ viết tắt tiếng Anh tại địa chỉ <http://www.abbreviations.com> gồm hơn 2000 từ viết tắt, hệ thống tra cứu chữ viết tắt tiếng Pháp tại địa chỉ <http://www.les-abreviations.com>. Tuy nhiên, đối với tiếng Việt, chúng ta vẫn chưa tìm thấy một hệ thống trực tuyến chữ viết tắt như các ngôn ngữ khác.

3. Quy tắc viết tắt trong tiếng Việt

Việc sử dụng chữ viết tắt trong các văn bản tiếng Việt thường xuyên được thực hiện theo quy tắc được mô tả trong tài liệu Ngữ pháp tiếng Việt [1] "Khi tên gọi xuất hiện lần đầu trong văn bản thì không được viết tắt, mà phải viết dạng đầy đủ và ghi chú chữ viết tắt trong ngoặc đơn. Từ lần xuất hiện thứ hai trở đi, ta mới viết tắt".

Khi sử dụng chữ viết tắt trong soạn thảo văn bản, chúng ta phải xem xét hai trường hợp sau:

• Chữ viết tắt đã có sẵn: trường hợp này chữ viết tắt đã được định nghĩa trước hoặc thông dụng, nhiều người biết, không mập mờ, không phản nghĩa khi dùng; hoặc đã có quy định, chẳng hạn như đăng ký nhãn hiệu, quy định bảng các chữ viết tắt, danh mục chữ viết tắt,... thì sử dụng trực tiếp mà không cần định nghĩa lại.

• Chữ viết tắt chưa được định nghĩa: Trong trường hợp này chúng ta cần định nghĩa chữ viết tắt ngay khi xuất hiện lần đầu trong văn bản theo dạng sau:

<Cụm từ đầy đủ> (<Chữ viết tắt>)

Tùy theo các lĩnh vực và ngành nghề khác nhau mà các văn bản hành chính, các chuyên ngành có quy định riêng về việc dùng chữ viết tắt, được ghi rõ trong quyết định, tiêu chuẩn ngành, tài liệu hướng dẫn, triển khai thực hiện...

4. Một số quy tắc cấu tạo chữ viết tắt

Để có thể nhận dạng và thu thập tự động chữ viết tắt, chúng tôi đã tiến hành nghiên cứu các quy tắc cấu tạo chữ viết tắt. Theo kết quả nghiên cứu từ các tài liệu [3][3][4][5], chữ viết tắt được cấu tạo dưới các dạng thức như sau:

4.1. Dùng ký tự đầu tiên của các từ tổ

Dạng phổ biến nhất của chữ viết tắt đó là dạng được cấu tạo bằng các ký tự đầu tiên của các từ tổ trong một cụm từ.

Ví dụ: ĐHĐN = Đại học Đà Nẵng

4.2. Dùng ký tự đầu của các từ kép

Dạng chữ viết tắt thứ hai mà chúng ta cũng thường gặp đó là dạng được cấu tạo từ các ký tự đầu tiên của các từ kép.

Ví dụ: TĐC = Tiêu chuẩn Đo lường Chất lượng

4.3. Dùng các ký tự đầu của các từ tiếng nước ngoài

Một dạng chữ viết tắt trong tiếng Việt mà chúng ta cũng thường gặp đó là dạng cấu thành từ các chữ cái đầu của các từ tiếng nước ngoài.

Ví dụ: VINASA = Hiệp hội phần mềm Việt Nam (Vietnam Software Association)

4.4. Dùng nhiều hơn một ký tự là phụ âm

Chữ viết tắt trong tiếng Việt còn được cấu tạo bằng cách cấu tạo từ nhiều hơn một ký tự là phụ âm để phân biệt các chữ viết tắt khác.

Ví dụ: ThS = Thạc sĩ

4.5. Dùng các ký hiệu khoa học

Trong một số tài liệu tiếng Việt không phải văn bản pháp quy, người ta thỉnh thoảng dùng cả ký hiệu khoa học như hóa học để viết tắt. Đối với đa số nhiều người, những ký hiệu như vậy có thể rất quen thuộc nhưng có một số người chưa bao giờ biết đến. Do đó, trường hợp này cũng cần phải được chú trọng để hỗ trợ giải nghĩa cho người dùng.

Ví dụ: H₂O = Nước

5. Các dạng ký hiệu chữ viết tắt

Để xử lý văn bản nhằm thu thập tự động chữ viết tắt, ngoài việc nắm được quy luật cấu tạo chữ viết tắt chúng ta cũng cần nhận biết các quy luật ký hiệu chữ viết tắt thường

được sử dụng trong các văn bản tiếng Việt.

5.1. Một số dạng ký hiệu chữ viết tắt phổ biến

Dạng 1:

Cụm từ đầy đủ (Chữ viết tắt)

Đây là dạng tương đối phổ biến, ở dạng này chữ viết tắt được đặt trong cặp ngoặc đơn. Ví dụ: Tài liệu tham khảo (TLTK)

Công nghệ thông tin (CNTT)

Dạng 2:

Chữ viết tắt (Cụm từ đầy đủ)

Một dạng ký hiệu chữ viết tắt cũng thường được thấy trong các văn bản tiếng Việt là nghĩa của chữ viết tắt được đặt trong cặp ngoặc đơn.

Ví dụ: ĐHĐN (Đại học Đà Nẵng)

Dạng 3:

Chữ viết tắt dấu cách Cụm từ đầy đủ

Dạng biểu diễn chữ viết tắt thứ 3 là chữ viết tắt và nghĩa của nó được đặt cách nhau bởi một ký tự phân cách như dấu gạch ngang (–) hoặc dấu hai chấm (:).

CNPM – Công nghệ phần mềm

Dạng 4:

Ngoài các dạng phổ biến nêu trên, để thu thập tự động chữ viết tắt chúng ta cũng cần tập trung xử lý các tài liệu tiếng Việt như các luận văn, luận án. Trong các tài liệu dạng này, chữ viết tắt thường được định nghĩa ở dạng các bảng chữ viết tắt như sau:

Bảng 1. Bảng ví dụ chữ viết tắt

Chữ viết tắt	Cụm từ đầy đủ
BTCQ	Biểu thức chính quy
CNTT	Công nghệ thông tin
CSDL	Cơ sở dữ liệu
CVT	Chữ viết tắt

5.2. Một số dấu hiệu nhận biết chữ viết tắt

Để nhận biết chữ viết tắt, ngoài những dạng ký hiệu chữ viết tắt như đã trình bày ở trên, chúng ta cũng có thể rút ra một số dấu hiệu nhận biết chữ viết tắt như sau:

- Chữ viết tắt thường gồm các ký tự là phụ âm,
- Chữ viết tắt đa số là các ký tự in hoa
- Chữ viết tắt có thể có chứa các ký tự phân cách
- Chữ viết tắt tiếng Việt không sử dụng các nguyên âm như *Â, Ă, Ê, Ô, Ơ, U...*

Trên đây là một số dấu hiệu nhận biết chữ viết tắt qua khảo sát, tuy nhiên cho đến nay đối với tiếng Việt chưa có quy tắc thành lập chữ viết tắt, chưa có sự thống nhất về tiêu chí đánh giá chung cho việc thành lập và sử dụng chữ viết tắt. Một trong những vấn đề đối với chữ viết tắt tiếng Việt hiện nay là một chữ viết tắt có thể mang nhiều nghĩa, nhiều chữ viết tắt vay mượn tiếng nước ngoài, nhất là tiếng Anh.

6. Giải pháp nhận biết và thu thập tự động chữ viết tắt

6.1. Sử dụng phương pháp so khớp

Giải pháp thứ nhất mà chúng tôi đề xuất đó là sử dụng phương pháp so khớp tối đa (Maximum Matching - MM):

Forward/Backward hay còn gọi là Left Right Maximum Matching - LRMM. Ở phương pháp này, chúng ta sẽ duyệt một đoạn văn bản từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển và cứ thực hiện lặp lại như vậy cho đến hết câu. Loại trừ các từ đơn và kép tìm thấy trong từ điển, trong số các từ đơn không tồn tại trong từ điển mà khớp với các dấu hiệu nhận biết chữ viết tắt như trình bày ở mục 5.2 là các từ có khả năng là chữ viết tắt.

Nhận xét: Giải pháp tìm kiếm chữ viết tắt theo phương pháp so khớp tối đa không thực sự hiệu quả do phải so khớp với từ điển để loại các từ có nghĩa.

6.2. Sử dụng biểu thức chính quy

Biểu thức chính quy rất quan trọng và thường ứng dụng trong tìm kiếm và xử lý văn bản dựa trên các mẫu được quy định. Kết hợp với các quy luật và cấu tạo chữ viết tắt đã phân tích ở các mục 3, 4, 5 ở trên, chúng ta có thể tạo nên các mẫu tìm kiếm chữ viết tắt rất dễ dàng và xác suất phát hiện chữ viết tắt là rất lớn.

Ví dụ: Biểu thức chính quy `@'\\((.*?))'` định nghĩa mẫu so khớp tất cả các chữ viết tắt theo dạng: *Cụm từ đầy đủ (Chữ viết tắt)*.

Thuật toán thực hiện:

Đầu vào: Tập văn bản

Đầu ra: Danh sách chữ viết tắt

Các bước:

1. Tạo biểu thức chính quy định nghĩa mẫu so khớp: P
2. Tạo đối tượng của lớp Regex với đối số là các mẫu so khớp: R
3. Mở tệp đầu vào
4. Repeat
 - Đọc dòng văn bản từ tệp: S
 - Thực hiện so khớp để thu thập chữ viết tắt: R.Matches(S)
5. Until hết tệp

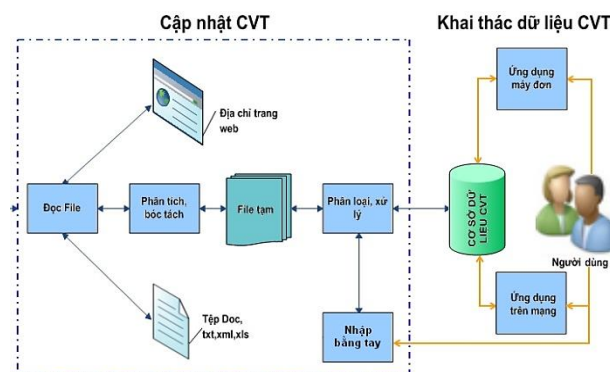
Nhận xét: So với giải pháp thứ nhất, giải pháp này có tốc độ xử lý nhanh hơn và chính xác hơn đối với các chữ viết tắt ở các dạng phổ biến. Tuy nhiên, do chữ viết tắt được cấu thành rất đa dạng, có nhiều chữ viết tắt không theo các quy luật như đã phân tích nên sử dụng phương pháp này không thể nhận biết tất cả các dạng chữ viết tắt.

7. Xây dựng hệ thống và kết quả thực nghiệm

7.1. Mô hình hệ thống tổng quát

Mô hình trên biểu diễn một hệ thống tổng quát cho phép thu thập và khai thác chữ viết tắt. Sau khi thu thập chữ viết tắt từ các tệp văn bản hoặc từ internet, hệ thống tiến hành phân tích và trích rút chữ viết tắt để cập nhật vào cơ sở dữ liệu chữ viết tắt. Quá trình thu thập chữ viết tắt tự động nên có thể không chính xác, do đó hệ thống phải cung cấp môi trường cho phép người dùng xem lại, chỉnh sửa và cập nhật chữ viết tắt. Từ cơ sở dữ liệu chữ viết tắt xây dựng được, chúng tôi sẽ đề xuất hai dịch vụ khai thác và ứng dụng chữ viết tắt đó là: từ điển cho phép tra cứu chữ viết tắt trực tuyến và chương trình tích hợp tra cứu và hỗ trợ gõ tắt.

Trong bài báo này, chúng tôi chỉ tập trung vào việc thu thập và xây dựng cơ sở dữ liệu chữ viết tắt. Đối với các chương trình tích hợp tra cứu và hỗ trợ gõ tắt chúng tôi sẽ đề cập trong các bài báo tiếp theo.



Hình 1. Mô hình hệ thống tổng quát

7.2. Kết quả thực nghiệm

7.2.1. Giao diện chương trình

Trên cơ sở phân tích các quy luật cấu tạo và ký hiệu chữ viết tắt, chúng tôi đã tiến hành xây dựng chương trình trích rút chữ viết tắt sử dụng ngôn ngữ lập trình C#. Sau khi người dùng nạp một tệp văn bản hoặc một trang web, chương trình sẽ tiến hành phân tích và trích rút chữ viết tắt (Hình 1). Giải pháp nhận biết chữ viết tắt mà chúng tôi sử dụng trong chương trình này chính là giải pháp sử dụng biểu thức chính quy như trình bày ở trên.

7.2.2. Giải pháp đánh giá độ chính xác của chương trình

Nhằm hỗ trợ kiểm chứng và đánh giá độ chính xác của chương trình nhận biết và thu thập chữ viết tắt, chúng tôi đã xây dựng một mô đun thống kê chữ viết tắt có giao diện như Hình 3.

Để đánh giá độ chính xác của giải pháp và chương trình trích rút, chúng tôi đã xây dựng và lựa chọn các tệp đầu vào có chứa đựng chữ viết tắt và dùng chương trình trên để thống kê số lượng và hiển thị chữ viết tắt. Qua một số lần thử, chúng tôi nhận thấy kết quả của chương trình trích rút chữ viết tắt tương đối chính xác trong tất cả các trường hợp chữ viết tắt được ký hiệu khớp với các mẫu tìm kiếm mà chúng tôi đã định nghĩa dùng biểu thức chính quy. Tuy nhiên, do chữ viết tắt có cấu tạo rất đa dạng và nhiều lúc ký hiệu tùy tiện trong các văn bản nên các mẫu tìm kiếm bằng biểu thức chính quy đã định nghĩa trong chương trình chưa thể so khớp được với tất cả các dạng chữ viết tắt.

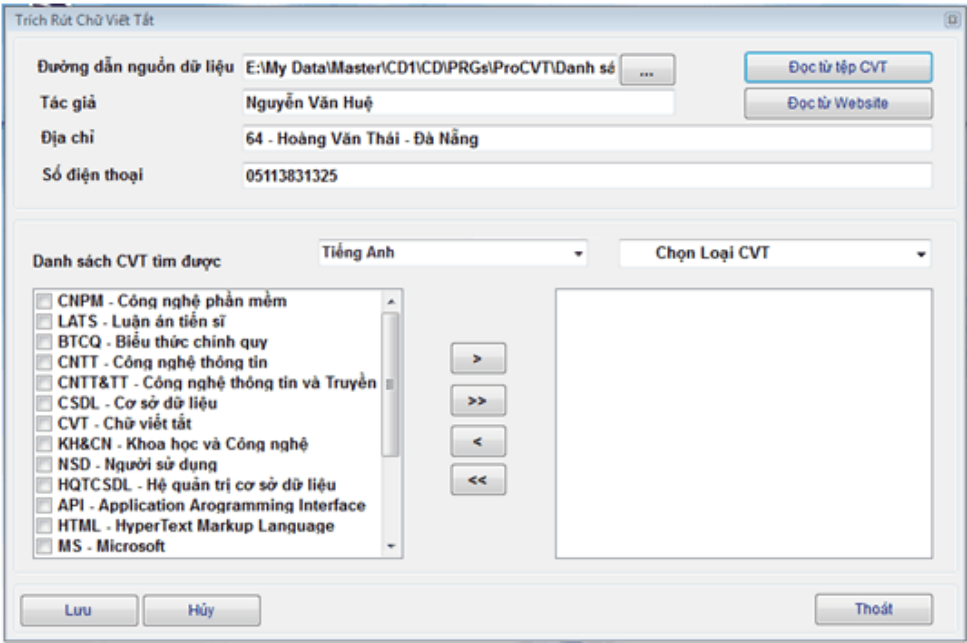
8. Kết luận

Chữ viết tắt trong tiếng Việt ngày càng tăng lên đáng kể về số lượng, đa dạng về ký hiệu, nhiều chữ viết tắt có nhiều nghĩa khác nhau. Chính vì thế người dùng thường gặp khó khăn trong việc đoán chữ viết tắt hoặc hiểu nhầm nghĩa, nhất là người đọc là những người nước ngoài đang nghiên cứu văn bản tiếng Việt hoặc những người không thuộc lĩnh vực mà văn bản đang trình bày.

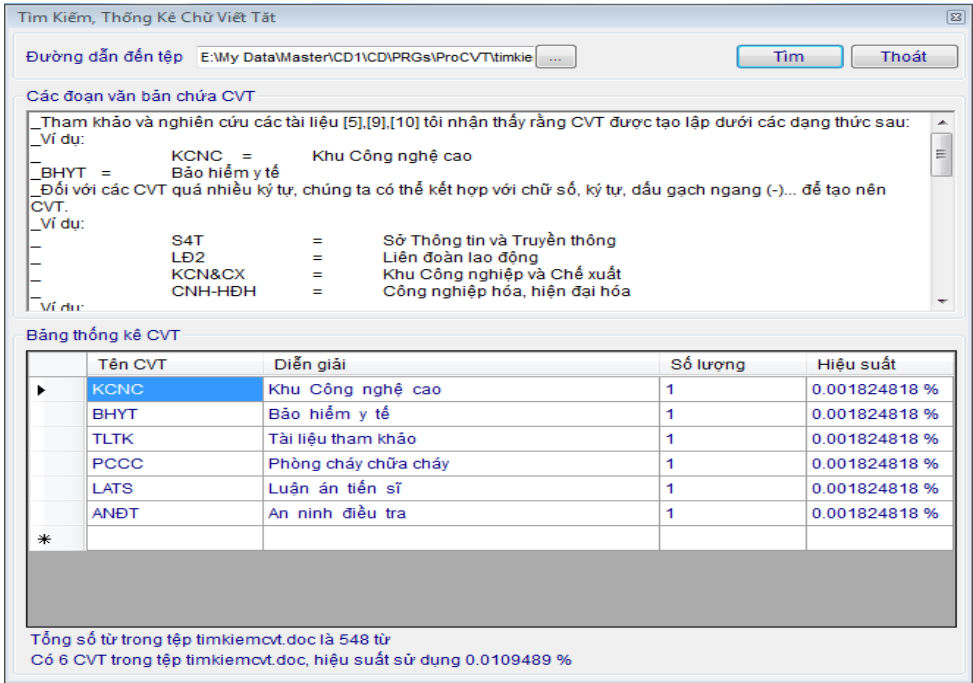
Đối với các ngôn ngữ khác như tiếng Anh, tiếng Pháp..., hiện nay đã có các hệ thống cho phép tra cứu chữ viết tắt. Tuy nhiên, đối với tiếng Việt, chúng ta vẫn chưa tìm thấy các hệ thống tra cứu chữ viết tắt hoặc hỗ trợ gõ tắt. Để xây

dựng được các hệ thống như vậy, việc đầu tiên chúng ta cần phải thực hiện đó là nghiên cứu và xây dựng một cơ sở dữ liệu chữ viết tắt tương đối đầy đủ. Bài báo này chúng tôi đã

đưa ra giải pháp thu thập và xây dựng cơ sở dữ liệu chữ viết tắt. Trong các công bố tiếp theo chúng tôi sẽ trình bày kết quả thu thập cũng như việc khai thác kết quả đó.



Hình 2. Giao diện chương trình trích rút chữ viết tắt



Hình 3. Giao diện chương trình thống kê chữ viết tắt nhằm hỗ trợ kiểm chứng và đánh giá

TÀI LIỆU THAM KHẢO

[1] Phan Huy Khánh, Nguyễn Nho Tuý (2007), *Setting Up the Database of Abbreviation for Service 1080*, Hội thảo RIVF, Hà Nội.

[2] Chim Văn Be (2001), *Ngữ pháp tiếng Việt*, Giáo trình đào tạo Đại học Cần Thơ.

[3] Bạch Như Nguyễn (2007), *Nghiên cứu tích hợp môi trường, công cụ và kỹ thuật tìm kiếm, quản lý chữ viết tắt*, Luận văn thạc sĩ kỹ thuật Khoa học máy tính, Đại học Đà Nẵng.

[4] Xuân Lâm, Minh Tân, Thanh Nghi (1998), *Từ điển tiếng Việt*, NXB Thanh Hóa.

[5] Nguyễn Thị Thu Thủy, Nguyễn Hữu Chinh (2001), *Tổng quan về ngôn ngữ và ngôn ngữ học*, Giáo trình đào tạo Đại học Cần Thơ.

[6] Phương pháp maximum matching, <http://www.cs.dartmouth.edu/~ac/Teach/CS105-Winter05/Notes/kavathekar-scribe.pdf>