XÂY DỰNG HỆ THỐNG RÚT TRÍCH CÁC NỘI DUNG CHÍNH CỦA VĂN BẢN KHOA HỌC TIẾNG VIỆT DỰA TRÊN CẤU TRÚC

Tạ Nguyễn[1], Vũ Đức Lung[2]

[1]*Khoa Công nghệ thông tin, trường Đại học Lạc Hồng*

[2]*Trường Đại học Công nghệ thông tin – ĐHQG TP.HCM*

Email: nguyen@lhu.edu.vn, lungvd@uit.edu.vn

Đến Tòa soạn: 21/8/2013; Chấp nhận đăng: 11/3/2014

TÓM TẮT

Bài báo trình bày cách thức rút trích các câu có nội dung quan trọng trong các văn bản khoa học tiếng Việt dựa trên cấu trúc. Hệ thống rút trích được xây dựng dựa trên một quy trình chặt chẽ mà bài báo đề xuất với việc áp dụng nhiều phương pháp khác nhau trong việc tính toán độ quan trọng thông tin của câu. Kết quả thử nghiệm cho thấy kết hợp phương pháp độ đo cục bộ và toàn cục (TF.IDF) với cách đánh giá câu theo cách cộng dồn trọng số từ cho kết quả tốt nhất. Bước đầu thử nghiệm trên các bài báo khoa học và toàn văn báo cáo thuộc lĩnh vực Công nghệ thông tin đã cho những kết quả có độ chính xác cao so với yêu cầu.

Từ khóa: rút trích, văn bản, ý chính, quy trình, trọng số từ, cấu trúc văn bản.

1. GIỚI THIỆU

Đối với những người làm nghiên cứu thì việc tìm kiếm tài liệu để tham khảo là một vấn đề vô cùng quan trọng, trong khi đó không phải chỉ đọc lướt qua là người ta có thể nắm hết các ý mà tác giả muốn nêu trong tài liệu. Có khi mất khá nhiều thời gian để đọc hết một tài liệu rồi nhận ra tài liệu đó không phù hợp với mục tiêu tìm kiếm của mình. Khác với việc chúng ta đọc rồi tự rút ra cho mình những ý chính trong toàn bộ văn bản như lâu nay mọi người thường làm, điều đó không tránh khỏi sự chủ quan trong chọn lựa ý chính vì mỗi người có những trình độ khác nhau, có chuyên môn khác nhau. Trong khi đặc điểm của văn bản khoa học là trong mỗi văn bản, tác giả – nhà khoa học – luôn mong muốn trình bày, thậm chí là khẳng định một ý tưởng khoa học cụ thể [1].

Với mục đích giúp con người tiết kiệm thời gian hơn trong việc tìm kiếm, sàng lọc và tổng hợp các thông tin một cách khách quan trong kho tri thức khổng lồ của nhân loại – Internet, bài báo muốn đề cập đến một quy trình cho phép máy tính có thể tự động rút trích ý chính từ văn bản tương đối chính xác nhất mà cụ thể là các văn bản khoa học trong ngành công nghệ thông tin như bài báo khoa học và toàn văn báo cáo. Bên cạnh đó bài báo trình bày nhiều phương pháp thực hiện khác nhau trong việc tính độ quan trọng thông tin của câu để đưa ra nhận xét đánh giá phương pháp nào là tối ưu, từ đó đưa vào quy trình thực hiện việc rút trích.

Vấn đề rút trích tự động các ý chính trong văn bản cũng nhận được nhiều sự quan tâm của các nhà công nghệ thông tin trên thế giới. Có thể thấy rõ nhất là qua công cụ AutoSummarize trong phần mềm Microsoft Word của tập đoàn Microsoft. Có thể nói sơ qua cơ chế làm việc của công cụ này là nó sẽ tính điểm cho các câu chứa từ được lặp lại nhiều lần. Những câu được nhiều điểm nhất sẽ được gợi ý đưa ra cho người dùng. Tuy nhiên đối với các văn bản tiếng Việt thì công cụ này cho kết quả không có tính chính xác cao.

Ngoài ra cũng có một số bài báo đề cập đến các công trình nghiên cứu liên quan đến vấn đề xử lí ngôn ngữ tự nhiên trong việc rút trích tự động ý chính trong văn bản như:

- Vấn đề *Extracting Sentence Segments for Text Summarization: A Machine Learning Approach* - tạm dịch là rút trích các phân đoạn câu phục vụ cho việc tóm tắt văn bản: một phương pháp tiếp cận học máy - do Wesley T.Chuang làm việc tại Computer Science Department, UCLA, Los Angeles, CA 90095, USA và Jihoon Yang làm việc tại HRL Laboratories, LLC, 3011 Malibu Canyon Road, CA 90265, USA nghiên cứu [2].

- Đề tài *Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics* - tạm dịch là Đánh giá tự động phần tóm tắt sử dụng N-gram kết hợp với thống kê tần suất - của tác giả Chin-Yew Lin and Eduard Hovy vào năm 2003 [3].

Các đề tài trên đều có ưu điểm nhất định nhưng hầu hết các đề tài đều tập trung xử lí ngôn ngữ tiếng nước ngoài, đa số là các văn bản tiếng Anh. Để áp dụng cho các tài liệu tiếng Việt thì không có được độ chính xác mong muốn do đặc điểm ngôn ngữ tiếng Việt phức tạp và có rất nhiều điểm khác biệt so với ngôn ngữ khác.

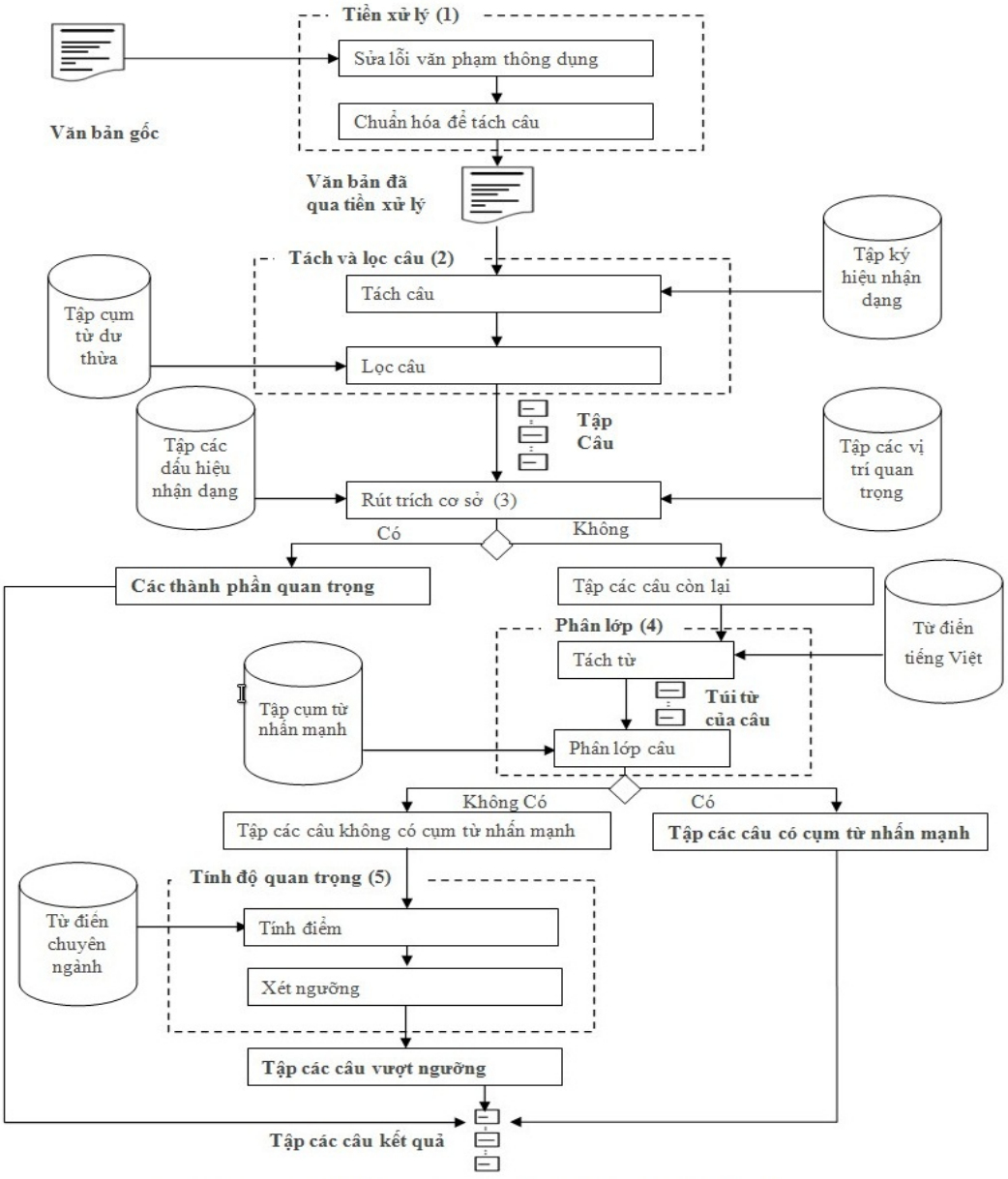
Còn trong nước có công trình nghiên cứu của Hoàng Kiếm và Đỗ Phúc về đề tài *Rút trích ý chính từ văn bản tiếng Việt hỗ trợ tạo tóm tắt nội dung* dựa trên việc sử dụng cây hậu tố để phát hiện các dãy từ phổ biến trong các câu của văn bản, dùng từ điển để tìm các dãy từ có nghĩa để giải quyết vấn đề ngữ nghĩa của các từ. Cuối cùng dùng kĩ thuật gom cụm để gom các câu trong văn bản và hình thành các vector đặc trưng cụm [1].

Các đề tài làm về vấn đề này đều có những ưu điểm nhất định của nó, tuy nhiên phạm vi xử lí văn bản của nó quá rộng, hầu như không xác định cụ thể cho một loại văn bản nào. Nếu đầu vào là một truyện ngắn, một quyển tiểu thuyết hay một bài báo khoa học thuộc những lĩnh vực khác nhau thì kết quả đầu ra có độ chính xác như thế nào? Đó chính là vấn đề mà với đề tài sẽ tập trung tìm hiểu vào một loại hình tài liệu, đó là văn bản khoa học trong ngành công nghệ thông tin nhằm đem lại kết quả có độ chính xác tốt nhất với yêu cầu của người dùng.

2. PHƯƠNG PHÁP RÚT TRÍCH Ý CHÍNH TRONG VĂN BẢN TIẾNG VIỆT

Nghiên cứu trong công trình này áp dụng phương pháp thống kê có cải tiến kết hợp học máy, do thực hiện trên đối tượng là văn bản khoa học cụ thể nên sẽ tập trung khảo sát cấu trúc các loại tài liệu, đưa ra các số liệu thống kê về vị trí thành phần quan trọng, xây dựng tập ngữ cố định dùng phân lớp câu để trích chọn trực tiếp và huấn luyện các từ chuyên ngành phục vụ cho việc tính toán độ quan trọng của câu. Việc tính toán độ quan trọng của câu sẽ sử dụng hai phương pháp khác nhau để từ đó đưa ra nhận xét phương pháp nào cho kết quả tối ưu hơn. Đồng thời cho phép người dùng có thể rút trích ý chính trong văn bản theo tỉ lệ hoặc theo một ngưỡng nào đó, ngưỡng này chính là điểm tối thiểu mà câu được đánh giá tính điểm. Tập các câu kết quả sau khi được trích chọn không sắp xếp theo điểm quan trọng mà sẽ giữ nguyên trật tự như trong văn bản gốc nhằm đảm bảo mạch ý tưởng và trình bày của tác giả văn bản. Bên cạnh đó các kết quả sẽ được huấn luyện bổ sung tập dữ liệu dùng trong công thức tính độ quan trọng của câu.

2.1. Quy trình rút trích ý chính đề xuất



Hình 1. Quy trình tổng quát rút trích ý chính văn bản khoa học.

2.2. Phương pháp tách câu

Câu trong nghiên cứu của chúng tôi được xem như đơn vị văn bản, sự chính xác trong việc tách câu ảnh hưởng nhiều đến việc rút trích hay xử lí văn bản. Chính vì thế module này đóng vai trò quan trọng trong chương trình. Dựa trên tập kí hiệu nhận dạng tách câu chương trình sẽ xử lí

tách câu cho văn bản. Các câu sau khi được tách sẽ được đưa vào một kho chứa dùng để xử lí tiếp tục cho các giai đoạn sau.

2.3. Phương pháp tách từ

Sử dụng mô hình n-gram với n = 2 kết hợp so khớp từ điển rút gọn để tách các từ ghép có nghĩa trong văn bản, huấn luyện tài liệu đồng thời ghi nhận tổng số từ trong văn bản làm tham số đầu vào cho giai đoạn tính toán.

Từ điển rút gọn là từ điển chỉ chứa các từ tiếng Việt có nghĩa bắt đầu bằng từ đầu tiên của cụm từ tách bằng n-gram, đây là một cải tiến nhằm giảm bớt thời gian xử lí trong việc so khớp.

Sau khi đã có túi từ chương trình sẽ huấn luyện các từ đó vào kho ngữ liệu dùng để phục vụ cho phần tính toán sau này.

2.4. Rút trích dựa trên cấu trúc tài liệu

Chương trình sẽ ghi nhận các vị trí quan trọng là mã câu sau khi tách câu, dựa trên các vị trí quan trọng và tập các dấu hiệu nhận dạng cho các phần quan trọng đã khảo sát từ trước. Sau khi có các vị trí đó sẽ nạp các phần đó vào tập các câu kết quả. Lưu ý giai đoạn rút trích cơ sở này chỉ áp dụng cho loại tài liệu là bài báo khoa học, còn đối với toàn văn thì chương trình sẽ không rút phần quan trọng trong toàn văn mà sẽ đánh giá tất cả các câu trong đó.

2.5. Phân lớp câu

Từ tập các câu không rơi vào các thành phần quan trọng sẽ được đưa vào bộ xử lí phân lớp câu. Bộ xử lí này dựa trên tập các ngữ cố định nhấn mạnh sẽ phân lớp các câu thành hai tập câu. Một tập chứa các câu mà trong nó có tồn tại ngữ cố định nhấn mạnh, tập còn lại không chứa ngữ nhấn mạnh đó. Tập các câu chứa ngữ nhấn mạnh sẽ được đưa vào tập câu kết quả.

2.6. Tính độ quan trọng của từ

2.6.1. Công thức kết hợp của độ đo cục bộ và toàn cục

Hiện nay một thuật toán đánh giá từ khóa dựa trên sự kết hợp của độ đo cục bộ và toàn cục là TF.IDF (Term Frequency - Inverse Document Frequency) cho một kết quả khá tốt.

Cách tiếp cận của TF.IDF sẽ ước lượng được độ quan trọng của một từ đối với một văn bản trong danh sách tập tài liệu văn bản cho trước. Nguyên lí cơ bản của TF.IDF là: "độ quan trọng của một từ sẽ tăng lên cùng với số lần xuất hiện của nó trong văn bản và sẽ giảm xuống nếu từ đó xuất hiện trong nhiều văn bản khác" [4]. Lí do đơn giản là vì nếu một từ xuất hiện trong nhiều văn bản khác nhau thì có nghĩa là nó là từ rất thông dụng vì thế khả năng nó là từ khóa sẽ giảm xuống (ví dụ như các từ "vì thế", "tuy nhiên", "nhưng", "và"…). Do đó độ đo sự quan trọng của một từ t trong tài liệu f sẽ được tính bằng: tf * idf, với tf là độ phổ biến của từ t trong tài liệu f và idf là nghịch đảo độ phổ biến của từ t trong các tài liệu còn lại của tập tài liệu. Được tóm tắt trong công thức tổng quát sau:

$$Weight_{wi} = tf * idf$$

với

$$tf = Ns(t)/ \sum w$$

$$idf = \log(\sum d/(d:t \in d))$$

trong đó: Ns(t): Số lần xuất hiện của từ t trong tài liệu f; ∑w: Tổng số các từ trong tài liệu f; ∑d = tổng số tài liệu; d:t∈d: số tài liệu có chứa từ t.

Ví dụ: Có một văn bản gồm 100 từ, trong đó từ "máy tính" xuất hiện 10 lần thì độ phổ biến: tf("máy tính") = 10 / 100 = 0,1.

Bây giờ giả sử có 1000 tài liệu, trong đó có 200 tài liệu chứa từ "máy tính". Lúc này chúng ta sẽ tính được idf("máy tính") = log(1000 / 200) = 0.699. Như vậy chúng ta tính được độ đo TF.IDF = tf*idf = 0.1 * 0.699 = 0.0699.

Độ đo này của từ càng cao thì khả năng là từ khóa càng lớn. Hướng tiếp cận độ đo TF.IDF này rất thông dụng hiện nay.

2.6.2. Công thức tính điểm thông tin quan trọng (Information Significant Score)

Theo [5] thì độ quan trọng của thông tin, ở đây là từ tiếng Việt được thể hiện qua công thức sau :

$$I(w_i) = \frac{N_s(w_i)}{\sum_{w_i \in S} w_i} + \frac{N_D(w_i)}{N_D}$$

trong đó: $N_s(w_i)$: số lần xuất hiện w_i trong văn bản gốc; $\sum w_i$: Tổng số w_i trong câu gốc; $N_D(w_i)$: Tổng số văn bản huấn luyện có mặt w_i; N_D: Tổng số tài liệu được huấn luyện (D).

Trong công thức này độ quan trọng thông tin của từ được xét trên từng câu so với toàn bộ văn bản.

Để kiểm nghiệm tính đúng đắn trong việc tính toán độ quan trọng của từ đề tài sẽ cài đặt cả hai công thức trên vào module đánh giá câu của hệ thống, qua đó đưa ra nhận xét và kết luận về khả năng ứng dụng và kết quả thực hiện của từng công thức.

2.7. Đánh giá câu

Theo Makoto [6] thì độ quan trọng của câu sẽ do trọng số của từng từ trong câu và tổng số từ trong câu quyết định, theo đó công thức mà Makoto đưa ra như sau :

$$Score(W) = \frac{1}{N}\sum_{n=1}^{N} I(w_n)$$

trong đó: *N*: là tổng số từ trong câu; $I(w_n)$: trọng số của từ;.

Với trọng số của từ được tính bằng công thức *TF.IDF* đã nói ở trên. Tuy nhiên công thức Makoto đưa ra áp dụng cho việc xử lí đánh giá câu không phải tiếng Việt.

Và theo đề tài dùng trọng số của từ để tóm tắt văn bản của tác giả R.C. Balabantara và cộng sự được đăng trong International Journal of Computer Applications (0975 – 8887) vào năm 2012 [7] thì cũng có ý tưởng tương tự như tác giả Makoto. Công thức mà đề tài của tác giả R.C. Balabantara [7] đưa ra như sau :

$$wt_s = \sum_{i=1}^{n} (wt_i)/n$$

với Wt_s là điểm của câu, wt_i là trọng số của từng từ được tính bằng công thức tính độ đo cục bộ kết hợp toàn cục và n là số từ có trong câu.

Qua đó chúng ta có thể thấy quan niệm của hai tác giả đề tài [6] và [7] là giống nhau. Điều đó có nghĩa là câu chứa ít từ cũng có thể chứa thông tin quan trọng.

Lại có quan niệm câu càng có nhiều từ quan trọng thì câu đó được xem quan trọng, điều đó có nghĩa là độ quan trọng của câu bằng tổng điểm (*tf*idf*) của các từ trong câu. Sau đây gọi là quan niệm thông thường.

3. KẾT QUẢ VÀ ĐÁNH GIÁ

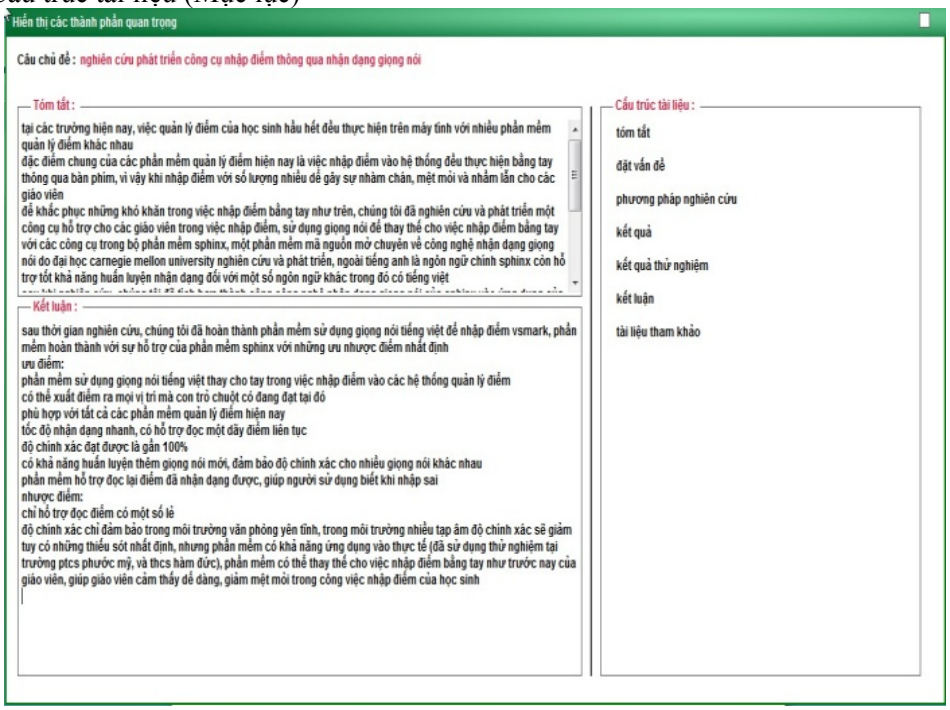
3.1. Thực nghiệm và đánh giá kết quả của EMIS (Extract Main Ideas System)

Chương trình thực nghiệm xử lí một bài báo khoa học có chủ đề "Nghiên cứu phát triển công cụ nhập điểm thông qua nhận dạng giọng nói".

Về các thành phần quan trọng mặc định của bài báo

Chương trình rút trích các phần quan trọng như đã quy định ban đầu là:

- Chủ đề (Tên tài liệu)
- Tóm tắt
- Kết luận
- Cấu trúc tài liệu (Mục lục)



Hình 2. Rút trích các thành phần quan trọng mặc định.

Qua hình 2 chúng ta có thể thấy kết quả xử lí cho tài liệu này là chính xác với các phần được rút trích đầy đủ như quy định.

Về việc xử lí đánh giá câu

Lọc theo tỉ lệ 7 % kết quả cho ra 13 câu có điểm cao nhất (kể cả các câu có ngữ cố định nhấn mạnh).

Bảng 1. Lọc kết quả theo tỉ lệ 7 %.

Mã câu	**Nội dung**
8	Để khắc phục những khó khăn trong việc nhập điểm bằng tay như trên, chúng tôi đã nghiên cứu và phát triển một công cụ hỗ trợ cho các giáo viên trong việc nhập điểm, sử dụng giọng nói để thay thế cho việc nhập điểm bằng tay
9	Với các công cụ trong bộ phần mềm Sphinx, một phần mềm mã nguồn mở chuyên về công nghệ nhận dạng giọng nói do đại học Carnegie Mellon University nghiên cứu và phát triển, ngoài tiếng Anh là ngôn ngữ chính Sphinx còn hỗ trợ tốt khả năng huấn luyện nhận dạng đối với một số ngôn ngữ khác trong đó có tiếng Việt
10	Sau khi nghiên cứu, chúng tôi đã tích hợp thành công công nghệ nhận dạng giọng nói của Sphinx vào ứng dụng của mình, và đã hoàn thành phần mềm VSMark có khả năng chuyển đổi giọng nói thành các từ dạng điểm số và xuất ra các vị trí mong muốn
11	Phần mềm Vsmark có khả năng hỗ trợ nhập điểm cho tất cả các phần mềm quản lí điểm hiện nay với độ chính xác khi nhận dạng giọng nói đạt được gần 100% sẽ giúp giáo viên cảm thấy dễ dàng, đơn giản và đảm bảo chính xác khi nhập điểm vào các hệ thống quản lí điểm khác nhau
18	Vì thế, việc đưa ra một giải pháp để thay thế cho việc nhập điểm bằng tay là một nhu cầu khách quan, chúng tôi đã đưa ra giải pháp sử dụng giọng nói tự nhiên để thay thế cho việc nhập điểm bằng tay như trước nay
22	Đơn giản, dễ sử dụng, việc sử dụng giọng nói tự nhiên để nhập điểm rất gần gũi với cuộc sống hằng ngày, vì vậy người sử dụng sẽ dễ dàng tiếp thu và sử dụng
37	Phần mềm có khả năng hỗ trợ cho hầu hết các phần mềm quản lí điểm hiện nay với độ chính xác khi nhận dạng đạt sắp sĩ 100% và có khả năng thích ứng với nhiều giọng nói khác nhau
47	Chúng tôi đã sử dụng các công cụ Sphinx4-beta6. SphinxTrain-1.0.7. CMUclmtk-0.7 và ngôn ngữ lập trình Java để hoàn thành phần mềm VSMark
55	Xác định các yêu cầu đặt ra trong quá trình nhập điểm của các phần mềm quản lí điểm
56	Tạo khả năng thích ứng với các hệ thống quản lí điểm khác nhau cho phần mềm hỗ trợ nhập điểm
67	Ngôn ngữ lập trình Java với nền Java Runtime JDK1.6.0 với công cụ hỗ trợ lập trình NetBean IDE 6.9.1
75	Tiến hành thử nghiệm phần mềm trên 2 môi trường khác nhau: môi trường văn phòng yên tĩnh và môi trường có nhiều tạp âm (tiếng gió, tiếng trò chuyện)
117	Tiến hành thử nghiệm trên hai đối tượng sử dụng khác nhau, một đối tượng đã thu âm trong cơ sở dữ liệu, một đối tượng chưa thu âm

Thật khó để đánh giá kết quả khi chưa có một ứng dụng đánh giá tóm tắt văn bản tiếng Việt, vì thế để có cái nhìn khách quan hơn về tính đúng đắn của hệ rút trích chúng ta xem xét các tiêu chí với cái nhìn của người đọc như:

- Câu phải chứa thông tin cụ thể
- Lí do thực hiện đề tài
- Phương pháp thực hiện
- Kết quả

Đây cũng là những tiêu chí mà người dùng quan tâm khi muốn tìm ý chính trong một tài liệu khoa học. Qua các tiêu chí trên chúng ta thấy:

- Các câu đều chứa thông tin cụ thể, không mơ hồ.
- Lí do thực hiện đề tài: câu số 8, 18, 22
- Phương pháp thực hiện: câu số 9, 10, 47, 55, 56, 67,75,117
- Kết quả: câu số 11, 37

Như vậy số câu mang các tiêu chí như trên là 13/13 câu, tỉ lệ là 100%. Qua đó chúng ta thấy kết quả trên có thể là cơ sở để người dùng tham khảo đưa ra quyết định, tỉ lệ trên thay đổi theo số lượng câu mà người dùng chọn ban đầu, tỉ lệ này có thể thay đổi để người dùng có thể tham khảo thêm nhiều câu hơn đến khi nào đưa ra quyết định hay nhận biết được nội dung chính của tài liệu.

3.2. Đánh giá kết quả thực nghiệm từ hai công thức sử dụng

Trong đề tài cũng như trong chương trình đã sử dụng cả hai công thức, là công thức *TF.IDF* và công thức Information Significant Score [5] để đánh giá độ quan trọng cho từng câu. Đây là hai công thức đã có từ trước, việc quyết định công thức nào phù hợp với bài toán rút trích này hoặc công thức nào cho độ chính xác cao hơn sẽ được thực nghiệm qua chương trình. Bên cạnh đó với công thức tính độ đo cục bộ và toàn cục đề tài cũng xét kết quả đánh giá câu theo hai quan niệm như đã đề cập ở phần trước là quan niệm thông thường và quan niệm của Makoto [6].

Qua kết quả thực nghiệm đề tài đã nhận thấy để đạt được kết quả tốt hơn thì nên chọn lựa sử dụng phương pháp kết hợp độ đo cục bộ và toàn cục (TF.IDF) với cách đánh giá câu theo quan niệm câu càng chứa nhiều từ có độ quan trọng cao thì câu đó càng có độ quan trọng cao.

3.3. Đánh giá kết quả của con người với kết quả của EMIS (Extract Main Ideas System)

Bảng 2 là kết quả rút trích của 10 người học tập và làm việc trong lĩnh vực công nghệ thông tin và hệ thống rút trích ý chính (EMIS) tham gia xử lí các tài liệu sau:

Tài liệu 1: *Xây dựng hệ thống mô phỏng phòng máy dùng trong quản lí hỏng hóc, sửa chữa* của tác giả Nguyễn Minh Sơn và Phan Thị Hường, Hội nghị nghiên cứu khoa học, trường Đại học Lạc Hồng, 2012

Tài liệu 2: *Hệ thống điều khiển Robot di chuyển tự động theo mục tiêu màu ứng dụng Board DE2* của tác giả Vũ Đức Lung, Trần Ngọc Đức và Lê Phước Phát Đạt Đức. Hội nghị nghiên cứu khoa học, trường Đại học Công nghệ thông tin, Đại học Quốc gia TP.HCM, 2012

Tài liệu 3: *Enrichment Computer Science Bibliography* của tác giả Đỗ Văn Tiến, Nguyễn Phước Cường và Huỳnh Ngọc Tín, Hội nghị khoa học trẻ UIT 2011.

Tài liệu 4: *Build social networking location-based services on Windows Phone 7 environments* của tác giả Đoàn Ngọc Nam, Trần Lễ Nhơn, Phạm Thi Vương, Hội nghị khoa

học trẻ UIT 2011

Tài liệu 5: *Một số vấn đề về xử lí ngữ nghĩa trong dịch tự động ngôn ngữ tự nhiên* của tác giả Trương Xuân Nam và Hồ Sỹ Đàm, công bố năm 2004.

Bảng 2. Chi tiết kết quả rút trích.

	P1 (n(S))	P2 (n(S))	P3 (n(S))	P4 (n(S))	P5 (n(S))	P6 (n(S))	P7 (n(S))	P8 (n(S))	P9 (n(S))	P10 (n(S))	EMIS
D1 (82 câu)	10(12,13,24,32,45,46,47,48,50,53)	11(11,12,13,26,27,28,36,48,53,**63,67)**	12(11,12,14,15,28,29,32,48,52,54,55,56)	11(14,26,27,36,46,47,48,50,54,55,56)	9(14,15,46,47,48,50,54,55,56)	2(27,32)	20(11,12,23,26,32,36,46,47,48,50,53,55,59,**61,62,63,64,65,66,67)**	8(11,23,26,28,32,36,55,**65)**	6(14,36,44,53,**61,67)**	11(11,12,24,29,31,32,36,46,47,48,53)	61,53,48,47,46,36,31,32,11,12,23,13,14,58,26,15,27,10,29,24,16,28,39,44,55,57,56,59,54,52
D2 (168 câu)	14(20,29,30,31,32,36,47,60,70,88,97,107,108,134)	28(16,17,18,19,25,28,29,30,31,32,34,41,74,75,76,80,81,82,83,89,90,91,92,103,104,109,110,111)	11(21,24,34,36,47,48,60,103,134,135)	23(15,18,20,34,36,51,52,53,58,60,61,71,72,73,88,89,90,96,104,107,109,110,123)	19(21,51,52,53,58,60,71,72,73,88,89,90,91,94,104,107,109,110,123)	1(34)	4(21,69,70,**134)**	5(20,24,34,74,79)	5(20,21,70,**134,141)**	9(11,20,21,22,31,32,36,108,134)	19,20,21,22,29,30,31,32,36,41,42,47,60,70,80,108,134,43,11,44,114,98,15,34,100,10,105,45,91,103,51,111,66,109,53,52,40,106,88,59
D3 (209 câu)	10(22,23,35,59,68,71,75,88,90,98)	16(22,23,31,32,33,34,44,46,47,59,75,76,95,98,**136,138)**	12(17,22,23,32,33,44,64,68,71,94,95,98)	20(18,20,22,23,29,42,44,48,50,52,53,59,60,61,64,73,88,91,98,133)	9(18,19,30,64,71,73,90,95,133)	10(30,45,46,47,50,59,73,75,76,90)	9(18,20,22,23,35,44,59,**135,138)**	11(22,24,25,26,44,52,53,64,75,88,102)	7(22,23,35,64,71,**135,138)**	12(22,23,35,44,51,59,60,68,71,73,132,135)	20,22,23,35,44,51,59,60,64,68,71,73,88,90,91,95,98,135,16,132,87,18,17,76,82,131,45,34,103,25,46,39,48,63,42,40,37,75,53,32,43,61,31
D4 (186 câu)	9(29,63,69,72,114,124,127,**142)**	18(29,40,41,42,52,53,60,61,62,69,109,110,114,122,123,**133,134)**	9(41,42,52,53,60,109,110,114,**139)**	14(29,30,42,46,52,53,63,64,69,71,72,106,109,114)	10(29,36,42,52,69,72,109,110,114,115)	16(35,37,40,41,46,47,69,71,106,109,114,122,123,124,127,128)	7(65,69,72,106,110,**133,134)**	8(29,52,69,72,106,109,116,134)	10(29,42,52,63,69,72,106,109,114)	6(23,24,29,40,71,1133)	29,40,41,52,69,71,109,114,123,133,63,64,25,60,24,65,134,139,136,115,147,141,106,26,110,30,116,140,146,126,42,28,48,91,58,36,37,127,35
D5 (235 câu)	14(20,22,23,37,38,56,57,58,65,66,67,77,85,102)	13(16,18,25,39,46,54,63,72,78,86,128,144,**200)**	17(26,28,29,35,36,37,38,54,62,63,75,76,77,85,102,118,200)	16(18,25,26,28,37,38,54,65,71,77,85,86,102,148,149,150)	15(20,22,23,37,38,54,65,66,67,71,77,85,102,148,152)	16(28,35,36,37,38,54,55,56,57,58,62,77,78,85,86,102)	7(26,28,72,75,152,**158,201)**	12(16,26,28,48,54,62,71,75,114,148,149,150)	8(16,26,28,75,117,149,150,**199)**	10(26,28,37,38,54,63,71,72,73,117)	54,55,26,73,117,102,71,72,63,85,86,75,76,77,78,118,114,116,55,142,68,71,69,147,140,139,141,113,105,120,137,125,126,74,,94,103,87,17,91,89,197

Chú thích:

- *n(S)*: *n* là số câu được người dùng rút trích và *S* là tập các câu được rút trích với các số nguyên là mã câu sau khi được EMIS xử lí.

- Tập các câu được nêu ra trong cột "EMIS" bao gồm tất cả các câu được EMIS rút ra và được sắp xếp giảm dần theo điểm quan trọng.

- Các câu được in đậm là các câu nằm trong thành phần quan trọng được EMIS rút ra nên mặc định sẽ được tính là trùng khớp với EMIS.

- Các câu mà EMIS rút ra trong bảng không bao gồm các câu trong phần tóm tắt và kết luận đối với bài báo khoa học – các thành phần đặc biệt quan trọng mặc định được rút trích.

Cách thức đánh giá

- Kết quả được đánh giá theo số lượng câu mà người dùng rút ra để bảo đảm tính khách quan. Ví dụ như người dùng rút ra được 12 câu thì sẽ lấy 12 câu có điểm cao nhất mà EMIS xử lí để so sánh, nếu người dùng rút ra 4 câu thì cũng chỉ lấy 4 câu điểm cao nhất của EMIS để so sánh.

Sau đây bảng 3 là kết quả so sánh giữa người và EMIS.

Bảng 3. Kết quả và tỉ lệ rút trích giữa người và EMIS.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Tổng	Tỷ lệ
D 1	5/10	10/13	5/12	5/11	3/9	0/2	17/20	3/8	4/6	9/11	61/102	59.80%
D 2	11/14	11/28	4/10	10/23	8/19	0/1	2/4	1/5	4/5	6/9	57/118	48.31%
D 3	9/10	8/16	8/12	11/20	5/9	4/10	8/9	3/11	5/7	10/12	71/116	61.21%
D 4	5/8	10/17	7/9	8/14	6/10	7/16	3/7	4/8	6/10	3/6	59/105	56.19%
D 5	4/15	5/9	9/17	7/16	5/14	7/16	4/7	4/12	4/8	7/10	56/124	45.16%
Trung bình	34/57	44/83	33/60	41/84	27/61	18/45	34/47	15/44	23/36	35/48	304/565	53.81%

Chú thích: m/n: *m* là số câu được rút trùng khớp giữa người dùng và EMIS, *n* là tổng số câu dùng so sánh.

Nhận xét

Qua bảng 2 chúng ta có thể thấy giữa những người tham gia khảo sát đã có sự khác biệt rất nhiều về việc rút trích, vì mỗi người mỗi ý, có thể một câu có thể là quan trọng với người này nhưng lại không có ý nghĩa với người khác. Qua đó thấy được sự phức tạp của vấn đề rút trích, ngoài việc đáp ứng gần 100 % các tiêu chí như bài báo này đã đề cập ở phần đánh giá kết quả xử lí tổng quát thì việc đáp ứng về phía người dùng cũng vô cùng quan trọng.

Qua bảng 3 nhận thấy được trong tổng số câu mà người dùng rút ra hay nói cách khác là tổng số câu mà người dùng xem như ý chính là 565 câu thì trong đó có 304 câu trùng khớp với các câu mà EMIS rút trích. Như vậy tỉ lệ của sự trùng khớp này là 53,81 %. Cũng cần nói thêm trong [7] được công bố năm 2012, cách đánh giá của [7] cũng tương tự như tác giả và cho ra kết quả trung bình khoảng 60 % nhưng có hai sự khác biệt lớn so với bài báo này:

- [7] xử lí ngôn ngữ là tiếng Anh.
- Độ nén của [7] thấp hơn nhiều so với bài báo này. Trong khảo sát mà [7] trình bày việc rút trích 1 đoạn văn trong khoảng dưới 10 câu, và rút ra từ 3 - 5 câu, như vậy độ nén trong khoảng 30 – 50 %. Trong khi đó với bài báo này là xử lí các bài báo khoa học và toàn văn thì số lượng câu lớn hơn rất nhiều, đối với bài báo (trung bình khoảng 200 câu) thì độ nén trong khoảng từ 4 – 10 %, còn đối với toàn văn (trung bình khoảng 1800 câu) thì độ nén thấp hơn chỉ từ 1 – 3 %. Chính vì thế xác suất xử lí của bài báo không thể lớn hơn do việc xử lí số lượng câu nhiều như vậy. Hay có thể nói việc chọn 3 câu trong 100 câu thì xác suất trùng khớp khó mà cao hơn được việc chọn 3 câu trong 10 câu.

Cho nên có thể nói với tỉ lệ xử lí 53,81 % là kết quả chấp nhận được và nhóm tác giả vẫn tiếp tục xây dựng thêm kho ngữ liệu qua việc huấn luyện và cập nhật để có thể nâng cao hơn tính chính xác của hệ thống.

4. KẾT LUẬN

Bài toán tóm tắt văn bản không phải là một vấn đề mới trên thế giới, đã có rất nhiều đề tài nghiên cứu về vấn đề này. Nhưng đến nay vẫn chưa có một hệ tóm tắt văn bản tiếng Việt nào

hoàn chỉnh và đạt độ chính xác mong muốn, phần vì sự phức tạp của tiếng Việt, phần vì miền giá trị xử lí của một số đề tài quá rộng không đảm bảo độ chính xác như mong muốn. Với bài báo này, chúng tôi hy vọng sẽ đem đến một quy trình rút trích cho những thể loại văn bản cụ thể dựa trên đặc trưng của ngôn ngữ tiếng Việt, cấu trúc của tài liệu đồng thời thử nghiệm các phương pháp đã áp dụng thành công với tiếng Anh vào việc xử lí tiếng Việt. Từ đó đưa ra những đánh giá và đề xuất một quy trình rút trích ý chính mà trong đó sử dụng phương pháp cho ra kết quả tốt nhất.

Kết quả thực nghiệm và khảo sát cho thấy mức độ chính xác của việc rút trích trên máy dựa trên quy trình đề xuất so với các tiêu chí đề ra là tốt và so với con người có thể chấp nhận được, bước đầu tạo tiền đề xây dựng một hệ tóm tắt văn bản tiếng Việt hoàn chỉnh với độ chính xác cao.

Sau quá trình nghiên cứu và thực hiện, bài báo đã đạt được những kết quả sau:

- Tìm hiểu một hệ thống rút trích các ý chính trong văn bản tiếng Việt dựa trên bài toán tóm tắt văn bản tự động.
- Tìm hiểu các bài toán tách từ, tách câu tiếng Việt từ đó xây dựng module tách từ sử dụng mô hình n-gram kết hợp so khớp từ điển rút gọn đem lại kết quả tách từ chính xác, tham gia vào việc huấn luyện tài liệu phục vụ cho việc tính toán độ quan trọng của từ và câu.
- Xây dựng bộ xử lí tính toán độ quan trọng của câu dựa trên nhiều phương pháp khác nhau, so sánh đánh giá kết quả để chọn ra phương pháp tốt nhất.
- Xây dựng kho dữ liệu các ngữ cố định nhấn mạnh, các ngữ cố định dư thừa phục vụ cho việc lọc và phân lớp câu.
- Xây dựng quy trình rút trích ý chính trong văn bản tiếng Việt với những giai đoạn chặt chẽ để cho ra các kết quả rút trích với độ chính xác tốt nhất.
- Xây dựng chương trình rút trích ý chính văn bản khoa học thể hiện đúng quy trình đã đề xuất.

Hướng phát triển tiếp của nhóm tác giả bài báo này:

- Phát triển thêm kho ngữ liệu ngữ cố định nhấn mạnh, ngữ cố định dư thừa và từ ghép chuyên ngành để tăng thêm độ chính xác trong việc tính toán độ quan trọng của câu.
- Cải thiện thuật toán phân lớp và tính toán câu để tăng tốc độ xử lí cho hệ thống.
- Mở rộng xử lí rút trích thêm các lĩnh vực khác.

TÀI LIỆU THAM KHẢO

1. Đỗ Phúc và Hoàng Kiếm - Rút trích ý chính từ văn bản tiếng Việt hỗ trợ tạo tóm tắt nội dung, Tạp chí Bưu Chính Viễn thông, Chuyên san các Công trình nghiên cứu triển khai Viễn thông và Công nghệ Thông tin **13** (2004).
2. Wesley T. Chuang and Jihoon Yang - Extracting Sentence Segments for Text Summarization: A Machine Learning Approach, SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (2000) 152-159.
3. Chin-Yew Lin and Eduard Hovy - Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics, NAACL '03 Proceedings of the 2003 Conference of the North

American Chapter of the Association for Computational Linguistics on Human Language Technology **1** (2003) 71-78.

4. Nguyễn Quý Minh - Xây dựng công cụ quảng cáo theo ngữ cảnh tiếng Việt, Luận văn thạc sĩ ngành Khoa học máy tính – Trường Đại học Khoa học tự Nhiên, TP. Hồ Chí Minh, 2009. tr 78
5. Ha Nguyen Thi Thu and Quynh Nguyen Huu - Concatenate the Most Likelihood Substring for Generating Vietnamese Sentence Reduction, IACSIT International Journal of Engineering and Technology **3** (3) (2011) 203-207.
6. Makoto Hirohata et al. - Sentence extraction-based presentation summarization techniques and evaluation metrics, Acoustics, Speech, and Signal Processing, (ICASSP '05) IEEE International Conference **1** (2005) 1065-1068.
7. Balabantara R. C. et al. - Text Summarization using Term Weights, International Journal of Computer Applications **38** (1) (2012) 0975-8887, 10-14.

ABSTRACT

EXTRACTING THE MAIN CONTENT OF VIETNAMESE SCIENTIFIC DOCUMENTS BASED ON THE STRUCTURE

Ta Nguyen[1], Vu Duc Lung[2]

[1]*Department of Information Technology,Lac Hong Universit, 10 Huynh Van Nghe Street, Buu Long Ward, Bien Hoa City, Dong Nai Province*

[2]*University of Information Technology, Vietnam National University - Ho Chi Minh city, Ward 6, Thu Duc District, Ho Chi Minh City*

Email: nguyen@lhu.edu.vn, lungvd@uit.edu.vn

This paper presents how to extract the main content in Vietnamese scientific documents based on their structure. In order to build this extraction system we proposed a strict process using different methods to evaluate the importance of the information of each sentence. The experimental results show that combining Term Frequency - Inverse Document Frequency method (TF.IDF) and Makoto Hirohata method gives us the best results. Our initiative tests only on full-text scientific papers and reports in information technology field, which are usually very long, offer a comparative extraction accuracy.

Keywords: extract, main content, extracting process, word weight, document structure.