ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN VĂN NGHIỆP

TÓM TẮT VĂN BẢN TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP TEXTRANK

LUẬN VĂN THẠC SĨ

HÀ NỘI – 2015

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN VĂN NGHIỆP

TÓM TẮT VĂN BẢN TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP TEXTRANK

Ngành: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Mã số: 60.48.01.04

LUẬN VĂN THẠC SĨ

Hướng dẫn khoa học: PGS. TS. NGUYỄN PHƯƠNG THÁI

HÀ NỘI - 2015

Lời cảm ơn

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành đến PGS, TS Nguyễn Phương Thái, người thầy đã hướng dẫn và chỉ dạy tận tình trong suốt quá trình tôi nghiên cứu khoa học và thực hiện luận văn thạc sỹ này.

Tôi cũng xin chân thành cảm ơn sự giúp đỡ nhiệt tình của bạn bè đã động viên, giúp đỡ trong thời gian học tập và nghiên cứu. Xin cảm ơn ThS Vũ Huy Hiển đã có những góp ý để tôi hoàn thiện hơn luận văn này.

Cuối cùng, tôi xin gửi lời cảm ơn đến gia đình, người thân và đồng nghiệp đã động viên, giúp đỡ và khuyến khích tôi vượt qua những lúc khó khăn trong cuộc sống, học tập và công việc.

Xin chân thành cảm ơn!

Tác giả

Nguyễn Văn Nghiệp

Lời cam đoan

Tôi xin cam đoan luận văn này được hoàn thành trên cơ sở nghiên cứu, tổng hợp và phát triển các nghiên cứu tóm tăt văn bản trong nước và trên thế giới do tôi thực hiện.

Luận văn này là mới, các đề xuất trong luận văn do chính tôi thực hiện qua quá trình nghiên cứu, thực nghiệm kết quả đưa ra và không sao chép nguyên bản từ bất kỳ một nguồn tài liệu nào khác.

Tác giả

Nguyễn Văn Nghiệp

Danh sách ký hiệu, viết tắt

Kí hiệu	**Giải thích**
w_{ij}	Trọng số giữa hai đỉnh V_i và V_j
$S(V_i)$	Trọng số của đỉnh V_i trong đồ thị
$In(V_i)$	Số cạnh vào đỉnh V_i
$Out(V_j)$	Số cạnh ra từ đỉnh V_j
$Similarity(S_i,S_j)$	Độ tương tự giữa câu S_i và câu S_j
w_k	Từ thứ k trong câu S
DUC	Document Understanding Conferences (Hội nghị chuyên về hiểu văn bản)
TAC	Text Analysis Conference
ACL	Association for Computational Linguistics
BLEU	BiLingual Evaluation Understudy
ROUGE	Recall Oriented Understudy of Gisting Evaluation

Danh sách hình vẽ

Hình 1 Đường cong hội tụ của phương pháp xếp hạng dựa trên đồ thị với đồ thị có hướng - vô hướng, có trọng số - không trọng số, 250 đỉnh và 250 cạnh 13

Hình 2 Đồ thị thể hiện mối quan hệ giữa các đơn vị từ vựng trong văn bản 17

Hình 3 Đồ thị mô phỏng các kết nối giữa các cập câu trong văn bản 23

Hình 4 Mô hình tóm tắt văn bản Tiếng Việt sử dụng TextRank 28

Hình 5 Mô hình tóm tắt văn bản Tiếng Việt sử dụng Cosine 28

Hình 6 Đồ thị mô phỏng quan hệ giữa các câu trong văn bản mẫu sử dụng TextRank 33

Hình 7 Đồ thị mô phỏng quan hệ giữa các câu trong văn bản mẫu sử dụng Cosine 34

Hình 8 Biểu đồ phân bố điểm đánh giá văn bản tóm tắt 6 tập mẫu 40

Hình 9 Biểu đồ phân bố điểm đánh giá văn bản tóm tắt của 13 tập dữ liệu 43

Hình 10 Giao diện chương trình tóm tắt văn bản tự động 47

Hình 11 Giao diện hiển thị đồ thị quan hệ giữa các câu trong văn bản 47

Danh sách bảng biểu

Bảng 1 So sánh kết quả trích xuất từ khoá giữa TextRank và Hulth 2003 20

Bảng 2 Kết quả so sánh tóm tắt đơn giữa TextRank và các hệ thống khác 25

Bảng 3 Danh sách chủ đề và số lượng văn bản tương ứng 37

Bảng 4 Kết quả đánh giá hệ thống tóm tắt tự động sử dụng độ đo Cosine 38

Bảng 5 Thời gian tóm tắt và đánh giá các bộ dữ liệu dùng Cosine 39

Bảng 6 Kết quả đánh giá hệ thống tóm tắt tự động sử dụng TextRank 39

Bảng 7 Thời gian tóm tắt và đánh giá các bộ dữ liệu dùng TextRank 41

Bảng 8 Kết quả đánh giá 13 bộ dữ liệu sau khi đã phân tích 43

Mục lục

Lời cảm ơn i

Lời cam đoan ii

Danh sách ký hiệu, viết tắt iii

Danh sách hình vẽ iv

Danh sách bảng biểu v

Mở đầu 1

Chương 1 Tổng quan bài toán tóm tắt văn bản 3

1.1. Tổng quan tóm tắt văn bản 3

1.2. Một số khái niệm cơ bản 4

1.3. Phân loại bài toán tóm tắt 4

1.4. Tóm tắt đơn văn bản 7

1.4.1. Tóm tắt theo trích xuất 7

1.4.2. Tóm tắt theo tóm lược 8

1.5. Đánh giá văn bản tóm tắt 8

Chương 2 Tóm tắt văn bản sử dụng TextRank 11

2.1. Mô hình TextRank 11

2.1.1. Đồ thị vô hướng 12

2.1.2. Đồ thị có trọng số 13

2.1.3. Đồ thị hoá văn bản 14

2.2. Sử dụng TextRank trích xuất từ khoá 15

2.3. Sử dụng TextRank trích rút câu 20

2.4. Tóm tắt văn bản Tiếng Việt sử dụng TextRank 26

2.4.1. Một số đặc trưng của Tiếng Việt 26

2.4.2. Xây dựng hệ thống tóm tắt tự động văn bản Tiếng Việt 27

Chương 3 Thực nghiệm và đánh giá kết quả 37

3.1. Dữ liệu thực nghiệm 37

3.2. Thực nghiệm và đánh giá với độ đo Cosine 38

3.3. Thực nghiệm và đánh giá với độ đo TextRank 39

3.4. Khuyến nghị tăng cường độ chất lượng văn bản tóm tắt 44

3.4.1. Khuyến nghị tăng cường độ liên quan giữa các câu 44

3.4.2. Khuyến nghị tăng cường chất lượng văn bản tóm tắt 45

Tổng kết 46

Phụ lục 48

Tài liệu tham khảo 51

Mở đầu

Hiện nay, công nghệ thông tin đang phát triển mạnh mẽ kèm theo với nó là sự bùng nổ của internet đã mang đến một lượng thông tin khổng lồ cho con người. Rất nhiều người có nhu cầu tổng hợp và tóm tắt lại các thông tin để thuận lợi cho việc tổng hợp các thông tin đó. Xuất phát từ nhu cầu đó, các phương pháp tóm tắt tự động được nghiên cứu và phát triển. Tóm tắt dữ liệu tự động là một lĩnh vực rất quan trọng, nó bao gồm trong đó là học máy và khai phá dữ liệu. Bài toán tóm tắt dữ liệu tự động không chỉ dừng lại ở tóm tắt văn bản mà nó còn mở rộng ra các loại dữ liệu đa phương tiện như hình ảnh, âm thanh và video. Một ví dụ điển hình cho việc ứng dụng của tóm tắt dữ liệu tự động là các máy tìm kiếm, trong đó nổi bật nhất là bộ máy tìm kiếm Google.

Hiện nay trên thế giới, nhiều nhà khoa học và các công ty tỏ ra rất quan tâm đến bài toán tóm tắt văn bản tự động. Tại các hội nghị nổi tiếng như: DUC 2001 - 2007, TAC 2008 – 2011, ACL 2001-2015, tóm tắt văn bản tự động đã được đề cập đến nhiều trong các bài báo. Ngoài ra, có nhiều hệ thống tóm tắt văn bản độc lập hoặc tích hợp được phát triển như: MEAD, LexRank, chức năng tự động tóm tắt trong Microsoft Word.

Trên thế giới có hai cách tiếp cận bài toán tóm tắt: Tóm tắt trích rút và tóm tắt rút gọn. Đối với tóm tắt trích rút, chương trình tóm tắt tự động sẽ trích rút ra các thành phần của văn bản mà không chỉnh sửa nội dung của nó rồi ghép lại thành một văn bản hoàn chỉnh. Loại tóm tắt này bao gồm trích rút câu và trích rút cụm từ. Như vậy, tóm tắt trích rút chỉ sử dụng các thông tin có sẵn trong văn bản như: từ, cụm từ, câu để tạo ra văn bản tóm tắt. Đối với tóm tắt rút gọn, cách tiếp cận này sử dụng ngữ nghĩa của các thành phần trong văn bản, các kỹ thuật trong xử lý ngôn ngữ tự nhiên để tạo ra văn bản tóm tắt gần giống với văn bản được tóm tắt bởi con người.

So sánh với các phương pháp tóm tắt văn bản tự động khác, TextRank có ưu điểm không cần thiết phải có các kiến thức sâu về ngôn ngữ, đồng thời có thể chuyển đổi phù hợp với nhiều bài toán khác nhau và nhiều ngôn ngữ khác nhau. Từ các ưu điểm đó ta sẽ dễ dàng áp dụng phương pháp này đối với bài toán tóm tắt văn bản Tiếng Việt và có thể kết hợp với các phương pháp khác để cho kết quả tốt hơn. Với các phân tích và lý do trên, tác giả lựa chọn đề tài luận văn thạc sĩ "**Tóm tắt văn bản Tiếng Việt sử dụng phương pháp TextRank**" để nghiên cứu.

Ngoài phần mở đầu và kết luận, cấu trúc luận văn bao gồm 3 chương như sau:

- **Chương 1: Tổng quan bài toán tóm tắt văn bản**, chương này giới thiệu tổng quan về bài toán tóm tắt văn bản tự động, tóm tắt văn bản sử dụng phương pháp trích rút.

- **Chương 2: Tóm tắt văn bản sử dụng Text Rank**, trình bày chi tiết về phương pháp tóm tắt văn bản Text Rank. Đồng thời áp dụng phương pháp này vào tóm tắt văn bản tiếng Việt tự động.

- **Chương 3: Thực nghiệm và đánh giá** trình bày chi tiết quá trình thực nghiệm trong khi thực hiện luận văn và đưa ra các đánh giá về các kết quả đạt sau thực nghiệm. Đồng thời đưa ra một số kiến nghị nâng cao hiệu năng và chất lượng của văn bản tóm tắt của văn bản tiếng Việt.

Chương 1 Tổng quan bài toán tóm tắt văn bản

1.1.Tổng quan tóm tắt văn bản

Trong những năm thập niên 50 – 60 của thế kỷ XX, các nhà khoa học đã bắt đầu nghiên cứu về tóm tắt văn bản tự động. Tháng 4/1958, H. P. Luhn đã công bố bài báo trình bày phương pháp tóm tắt tự động sử dụng thống kê tần suất và phân bố từ trong văn bản. Đến năm 1969, H. P. Edmundson đã công bố nghiên cứu về phương pháp mới trong việc tóm tắt tự động văn bản. Phương pháp này dựa trên tổng hợp của bốn thành phần: vai trò, khoá, tiêu đề và vị trí. Các phương pháp tiếp cận của hai nhà khoa học trên đều thuộc dạng trích rút câu. Các nghiên cứu về tóm tắt văn bản tự động sau một thời gian không có nhiều tiến triển thì đến cuối thế kỷ XX, đầu thế kỷ XXI, với sự bùng nổ mạnh mẽ của CNTT và internet, lượng thông tin được con người sinh ra và lưu trữ vô cùng lớn. Vấn đề được đặt ra là làm sao để thu nhận thông tin quan trọng nhanh nhất, hiệu quả nhất. Từ đó, bài toán tóm tắt văn bản trở nên cấp thiết và được quan tâm hơn đúng với tầm quan trọng của nó.

Theo Inderjeet Mani, tóm tắt văn bản tự động nhắm đến mục đích: "*Tóm tắt văn bản tự động nhằm mục đích trích xuất nội dung từ một nguồn thông tin và trình bày các nội dung quan trọng nhất cho người sử dụng theo một khuôn dạng súc tích và gây cảm xúc đối với người sử dụng hoặc một chương trình cần đến*".

Kết quả của quá trình tóm tắt văn bản tự động thường không cho kết quả chất lượng như văn bản tóm tắt bởi con người do bị giới hạn bởi nhiều yếu tố. Chúng ta rất khó khăn để nâng cao chất lượng văn bản tóm tắt tự động mà không bị giới hạn bởi miền ứng dụng. Vì vậy, trong tóm tắt văn bản tự động, các hướng giải quyết thường hướng đến các bài toán cụ thể với một phương pháp cụ thể.

1.2.Một số khái niệm cơ bản

- **Tỷ lệ nén (Compression Rate)**: là độ đo giữa thông tin văn bản tóm tắt và văn bản gốc được tính bằng công thức:

$$CompressionRate = \frac{SummaryLength}{SourceLength}$$

Trong đó:

- SummaryLength: Độ dài văn bản tóm tắt
- SourceLength: Độ dài văn bản gốc

- **Độ liên quan (Relevance)**: là độ đo cho mức độn quan trọng của thông tin mà văn bản tóm tắt có được so với văn bản gốc.
- **Sự mạch lạc (Coherence)**: là thước đo cho sự mạch lạc, tuân theo thể thống nhất của văn bản, không có sự trung lặp các thành phần.

1.3.Phân loại bài toán tóm tắt

Hiện tại có nhiều cách phân loại tóm tắt văn bản khác nhau, việc phân loại phụ thuộc vào cơ sở để tóm tắt[1]. Luận văn đề cập đến phân loại tóm tắt dựa trên các cơ sở:

- Định dạng văn bản, nội dung đầu vào
- Định dạng, nội dung đầu ra
- Mục đích tóm tắt

Chi tiết các phân loại dựa trên định dạng, nội dung đâu vào như sau:

- Kiểu văn bản (bài báo, bản tin, thư, báo cáo ...). Với cách phân loại này, tóm tắt văn bản là bài báo sẽ khác với tóm tắt thư, tóm tắt báo cáo khoa học do những đặc trưng văn bản quy định.

[1] Trần Mai Vũ (2009), Tóm tắt đa văn bản dựa vào trích xuất câu, Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2009

- Định dạng văn bản: dựa vào từng định dạng văn bản khác nhau, tóm tắt cũng chia ra thành các loại khác nhau như: tóm tắt văn bản không theo cấu trúc nhất định và tóm tắt văn bản có cấu trúc. Đối văn bản có cấu trúc, tóm tắt văn bản thường sử dụng một mô hình học dựa vào mẫu cấu trúc đã xây dựng từ trước để tiến hành tóm tắt.
- Số lượng dữ liệu đầu vào: Tóm tắt đơn văn bản khi đầu vào chỉ là một văn bản đơn, trong khi đó đầu vào của tóm tắt đa văn bản là một tập các tài liệu có liên quan đến nhau như: các tin tức có liên quan đến cùng một sự kiện, các trang web cùng chủ đề hoặc là cụm dữ liệu được trả về từ quá trình phân cụm.
- Miền dữ liệu: tùy theo miền của dữ liệu về cụ thể về một lĩnh vực nào đó, ví dụ như: y tế, giáo dục... hay miền dữ liệu tổng quát, có thể chia tóm tắt ra thành từng loại tương ứng.
- Tóm tắt trên cơ sở mục đích thực chất là làm rõ cách tóm tắt, mục đích tóm tắt là gì, tóm tắt phục vụ đối tượng nào …
    - Nếu phụ thuộc vào đối tượng đọc tóm tắt thì tóm tắt cho chuyên gia khác cách tóm tắt cho các đối tượng đọc thông thường.
    - Tóm tắt sử dụng trong tìm kiếm thông tin (IR) sẽ khác với tóm tắt phục vụ cho việc sắp xếp.
    - Dựa trên mục đích tóm tắt, còn có thể chia ra thành tóm tắt chỉ thị và tóm tắt thông tin. Tóm tắt chỉ thị chỉ ra loại của thông tin, ví dụ như là loại văn bản chỉ thị “tuyệt mật”. Còn tóm tắt thông tin chỉ ra nội dung của thông tin.
    - Tóm tắt trên cơ sở truy vấn (Query-based) hay tóm tắt chung. Tóm tắt chung có mục đích chính là tìm ra đoạn tóm tắt cho toàn bộ văn bản mà nội dung của đoạn văn bản sẽ bao quát toàn bộ nội dung của văn bản đó. Tóm tắt trên cơ sở truy vấn thì nội dung của văn bản tóm tắt sẽ dựa trên truy vấn của người dùng hay chương trình đưa vào, loại tóm

tắt này thường được sử dụng trong quá trình tóm tắt các kết quả trả về từ máy tìm kiếm.

Tóm tắt trên cơ sở đầu ra cũng có nhiều cách phân loại.

- Dựa vào ngôn ngữ: Tóm tắt cũng có thể phân loại dựa vào khả năng tóm tắt các loại ngôn ngữ:
  - Tóm tắt đơn ngôn ngữ (Monolingual): hệ thống có thể tóm tắt chỉ một loại ngôn ngữ nhất định như: tiếng Việt hay tiếng Anh…
  - Tóm tắt đa ngôn ngữ (Multilingual): hệ thống có khả năng tóm tắt nhiều loại văn bản của các ngôn ngữ khác nhau, tuy nhiên tương ứng với văn bản đầu vào là ngôn ngữ gì thì văn bản đầu ra cũng là ngôn ngữ tương ứng.
  - Tóm tắt xuyên ngôn ngữ (Crosslingual): hệ thống có khả năng đưa ra các văn bản đầu ra có ngôn ngữ khác với ngôn ngữ của văn bản đầu vào.
- Dựa vào định dạng đầu ra của kết quả tóm tắt: như bảng, đoạn, từ khóa.

Ngoài hai cách phân loại trên, phân loại tóm tắt trên cở sở đầu ra còn có một cách phân loại được sử dụng phổ biến là: tóm tắt theo trích xuất (Extract) và tóm tắt theo tóm lược (Abstract).

- Tóm tắt theo trích xuất: là tóm tắt có kết quả đầu ra là một tóm tắt bao gồm toàn bộ các phần quan trọng được trích ra từ văn bản đầu vào.
- Tóm tắt theo tóm lược: là tóm tắt có kết quả đầu ra là một tóm tắt không giữ nguyên lại các thành phần của văn bản đầu vào mà dựa vào thông tin quan trọng để viết lại một văn bản tóm tắt mới.

Hiện nay, các hệ thống sử dụng tóm tắt theo trích xuất được sử dụng phổ biến và cho kết quả tốt hơn tóm tắt theo tóm lược. Nguyên nhân tạo ra sự khác biệt này là do các vấn đề trong bài toán tóm tắt theo tóm lược như: biểu diễn ngữ nghĩa, suy luận và sinh ra ngôn ngữ tự nhiên được đánh giá là khó và chưa có

nhiều kết quả nghiên cứu khả quan hơn so với hướng trích xuất câu của bài toán tóm tắt theo trích xuất. Trong thực tế, theo đánh giá của Dragomir R. Radev (Đại học Michigan, Mỹ) chưa có một hệ thống tóm tắt theo tóm lược đạt đến sự hoàn thiện, các hệ thống tóm tắt theo tóm lược hiện nay thường dựa vào thành phần trích xuất có sẵn. Các hệ thống này thường được biết đến với tên gọi tóm tắt theo nén văn bản.

Tóm tắt theo nén văn bản (Text Compaction): là loại tóm tắt sử dụng các phương pháp cắt xén (truncates) hay viết gọn (abbreviates) đối với các thông tin quan trọng sau khi đã được trích xuất.

Mặc dù tính trên cơ sở phân loại có nhiều loại tóm tắt khác nhau nhưng hai loại tóm tắt là tóm tắt đơn văn bản và tóm tắt đa văn bản vẫn được sự quan tâm lớn của các nhà nghiên cứu về tóm tắt tự động.

1.4.Tóm tắt đơn văn bản

Bài toán tóm tắt văn bản đơn cũng giống như các bài toán tóm tắt khác, là một quá trình tóm tắt tự động với đầu vào là một văn bản, đầu ra là một đoạn văn bản ngắn gọn mô tả nội dung chính của văn bản đầu. Văn bản đơn có thể là một trang Web, một nội dung đăng trên mạng xã hội, một bài báo, một tài liệu dạng văn bản (ví dụ: .doc, .txt)... Tóm tắt văn bản đơn là bước làm cơ sở cho việc xử lý tóm tắt đa văn bản và các bài toán tóm tắt phức tạp hơn. Đó là nguyên nhân lý giải cho việc những phương pháp tóm tắt văn bản ra đời đầu tiên đều là các phương pháp tóm tắt đơn văn bản.

Các phương pháp nhằm giải quyết bài toán tóm tắt văn bản đơn cũng tập trung vào hai loại tóm tắt là: tóm tắt theo trích xuất và tóm tắt theo tóm lược.

1.4.1. Tóm tắt theo trích xuất

Đa số các phương tóm tắt loại này tập trung vào việc trích xuất ra các câu hay các ngữ nổi bật từ các đoạn văn bản và kết hợp chúng lại thành một văn bản tóm tắt. Một số nghiên cứu giai đoạn đầu thường sử dụng các đặc trưng như vị

trí của câu trong văn bản, tần số xuất hiện của từ, ngữ hay sử dụng các cụm từ khóa để tính toán trọng số của mỗi câu, qua đó chọn ra các câu có trọng số cao nhất cho văn bản tóm tắt [Lu58, Ed69]. Các kỹ thuật tóm tắt gần đây sử dụng các phương pháp học máy và xử lý ngôn ngữ tự nhiên nhằm phân tích để tìm ra các thành phần quan trọng của văn bản. Sử dụng các phương pháp học máy có thể kể đến phương pháp của Kupiec, Penderson and Chen năm 1995 sử dụng phân lớp Bayes để kết hợp các đặc trưng lại với nhau [PKC95] hay nghiên cứu của Lin và Hovy năm 1997 áp dụng phương pháp học máy nhằm xác định vị trí của các câu quan trọng trong văn bản [LH97]. Bên cạnh đó việc áp dụng các phương pháp phân tích ngôn ngữ tự nhiên như sử dụng mạng từ Wordnet của Barzilay và Elhadad vào năm 1997 [BE97].

1.4.2. Tóm tắt theo tóm lược

Các phương pháp tóm tắt không sử dụng trích xuất để tạo ra tóm tắt có thể xem như là một phương pháp tiếp cận tóm tắt theo tóm lược. Các hướng tiếp cận có thể kể đến như dựa vào trích xuất thông tin (information extraction), ontology, hợp nhất và nén thông tin... Một trong những phương pháp tóm tắt theo tóm lược cho kết quả tốt là các phương pháp dựa vào trích xuất thông tin, phương pháp dạng này sử dụng các mẫu đã được định nghĩa trước về một sự kiện hay là cốt truyện và hệ thống sẽ tự động điền các thông tin vào trong mẫu có sẵn rồi sinh ra kết quả tóm tắt. Mặc dù cho ra kết quả tốt tuy nhiên các phương pháp dạng này thường chỉ áp dụng trong một miền nhất định [MR95].

1.5.Đánh giá văn bản tóm tắt

Hiện tại, việc đánh giá kết quả văn bản tóm tắt tự động là việc làm khó khăn. Cách đánh giá tốt nhất là sử dụng ý kiến đánh giá của các chuyên gia ngôn ngữ. Nhưng đây là một phương pháp tốn kém. Vì vậy, ngoài các phương pháp đánh giá thủ công, vấn đề đánh giá tự động kết quả tóm tắt cũng nhận được

nhiều sự chú ý. Từ năm 2000, NIST[2] tổ chức hội nghị DUC hàng năm để thực hiện việc đánh giá các hệ thống tóm tắt văn bản. Việc đánh giá tự động nhằm mục đích là tìm ra được một độ đo đánh giá văn bản tóm tắt giống với đánh giá của con người nhất.

Độ hồi tưởng (recall) tại các tỷ lệ nén khác nhau là thước đo đánh giá hợp lý, cho nó không chỉ ra được sự khác nhau về hiệu suất. Độ đo này được tính theo công thức:

$$C = R \times E$$

Ở đây, R là độ hồi tưởng câu theo công thức:

$$R = \frac{Số\ đơn\ vị\ bao\ phủ}{Tổng\ số\ đơn\ vị}$$

E là tỷ lệ hoàn thành nằm trong khoảng từ 0 đến 1 (1 là hoàn thành tất cả, ¾ là một phần, ½ là một số, ¼ là khó, 0 là không có) DUC 2002 đã sử dụng một công thưc khác để đánh giá, C':

$$C' = \alpha * C + (1 - \alpha) * B$$

Trong đó B là sự ngắn gọn và α là tham số phản tầm quan trọng. Các loại nhãn cho E thay đổi tương ứng thành 100%, 80%, 60%, 40%, 20%, và 0%.

Phương pháp ROUGE

BiLingual Evaluation Understudy (BLEU) [KST02] là một phương pháp đưa ra để đánh giá các hệ thống dịch tự động. Phương pháp này có nhanh, độc lập với ngôn ngữ và sự liên quan với các đánh giá của con người. Recall Oriented Understudy of Gisting Evaluation (ROUGE) [LH03] được Lin và Hovy đưa ra vào năm 2003 dựa trên khái niệm tương tự BLEU. ROUGE sử dụng n-gram để đánh giá sự tương quan giữa các kết quả của văn bản tóm tắt và

[2] National Institute of Standards and Technology. http://nist.gov

tập dữ liệu đánh giá. Phương pháp này cho ra kết quả tốt và được đánh giá cao trong cộng đồng các nhà khoa học trong cùng lĩnh vực.

Công thức đánh giá ROUGE với n-gram được xác định như sau:

$$ROUGE - N = \frac{\sum_{S\epsilon\{Văn\ bản\ tham\ chiếu\}} \sum_{gram_n \epsilon S} Count_{match}(gram_n)}{\sum_{S\epsilon\{Văn\ bản\ tham\ chiếu\}} \sum_{gram_n \epsilon S} Count(gram_n)} \quad (1)$$

Trong đó:

- *S*: là câu trong văn bản
- *n*: Độ dài của gram đang xét
- $Count_{match}(gram_n)$: là số gram *n* trùng nhau lớn nhất của văn bản cần đánh giá và văn bản tham chiếu
- $Count(gram_n)$: Số gram *n* có trong văn bản tham chiếu

Như vậy, độ đo ROUGE-N thuộc dạng độ đo hồi tưởng (recall-related).

Chương 2 Tóm tắt văn bản sử dụng TextRank

Các thuật toán xếp hạng dựa trên đồ thị đã được đưa ra và sử dụng rộng rãi trong những năm trong thế kỷ XX. Trong số đó có thuật toán HITS của Kleinberg và Page rank của Google do hai nhà đồng sáng lập phát triển (Brin và Page). Chúng được sử dụng trong việc phân tích mạng xã hội, cấu trúc liên kết của các trang web, … Thực tế thì thuật toán xếp hạng dựa trên đồ thị xác định đỉnh nào là quan trọng trong đồ thị bằng cách tính toán đệ quy các thông tin trên toàn đồ thị thay vì chỉ sử dụng thông tin trên từng đỉnh. Quá trình này làm cho việc xác định mức độ quan trọng chính xác hơn.

Từ cách tiếp cận trên, ta có thể áp dụng sang các đồ thị từ vựng và đồ thị ngữ nghĩa trích xuất được từ các tài liệu trong ngôn ngữ tự nhiên. Kết quả của việc sử dụng mô hình xếp hạng dựa trên đồ thị có thể ứng dụng trong nhiều chương trình xử lý ngôn ngữ tự nhiên. Ví dụ như mô hình xếp hạng hướng văn bản được ứng dụng trong các vấn đề như tự động trích xuất từ khoá đến tóm tắt văn bản và xác định từ nhập nhằng ý nghĩa (Mihalcea et al., 2004).

Trong chương này ta sẽ tìm hiểu mô hình TextRank, các thuật toán và ứng dụng của nó trong việc trích xuất từ khoá và xếp hạng các câu trong một văn bản. Đây là tiền đề cho tóm tắt văn bản tiếng Việt tự động sử dụng phương pháp TextRank.

2.1. Mô hình TextRank

Như trình bày ở trên, thuật toán xếp hạng dựa trên đồ thị là cách đưa ra cách chọn định quan trọng trong đồ thị dựa trên các thông tin toàn cục của các đỉnh trong đồ thị. Ý tưởng của thuật toán này dựa trên hai yếu tố: bỏ phiếu và đề cử. Mỗi một liên kết đến đỉnh đang xét thì nó được 1 phiếu bầu. Như vậy, càng nhiều phiếu bầu thì đỉnh đó càng quan trọng. Từ cách xác định trên thì trọng số của một đỉnh chính là số phiếu bầu cho đỉnh đó.

Ta có đồ thị $G = (V, E)$ là đồ thị có hướng. Trong đó:

V: là tập các đỉnh

E: là tập các cạnh của đồ thị, E là tập con của $V \times V$ ($E \subseteq V \times V$)

Với mỗi đỉnh V_i thì ta có:

- In(V_i) là tập các đỉnh trỏ đến V_i

- Out(V_i) là tập các đỉnh mà V_i trỏ đến.

Trọng số của đỉnh V_i được xác định như sau (Brin and Page, 1998):

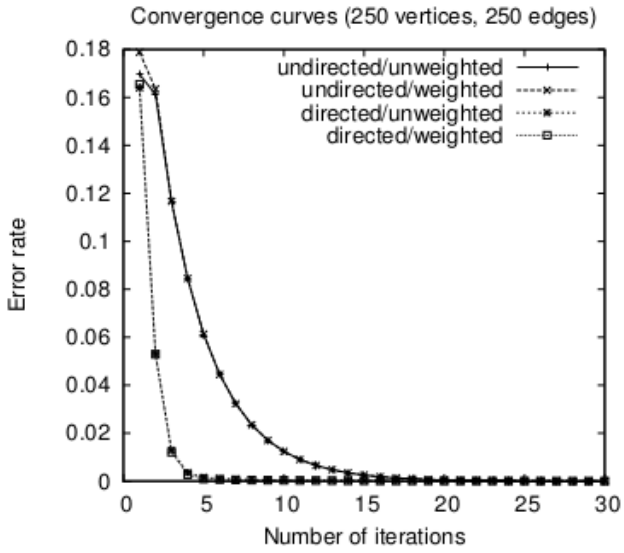
$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \qquad (2)$$

Trong đó d là nhân tố giảm, có giá trị từ 0 đến 1. Nó là xác xuất mà một đỉnh có liên kết đến một đỉnh bất kỳ trong đồ thị. Đối với các trang web thì d là xác suất người dùng nhấn vào một liên kết bất kỳ và xác suất để người dùng vào một trang web hoàn toàn mới là 1 – d. Theo Pagerank thì d = 0.85. Đây cũng là xác suất sẽ đươc sử dụng trong TextRank.

Ban đầu gán cho tất cả các đỉnh trong đồ thị các giá trị khởi tạo và tính toán lặp lại cho đến khi kết quả hội tụ lại đạt ngưỡng xác định. Sau quá trình tính toán thì trọng số của mỗi đỉnh chính là mức độ quan trọng của đỉnh đó trong toàn đồ thị. Có điều cần lưu ý, đó là giá trị trọng số của mỗi đỉnh sẽ không phụ thuộc vào giá trị khởi tạo ban đầu được gán cho mỗi đỉnh. Ngoài ra thì số lượng các vòng lặp tính toán để ra được trọng số là khác nhau.

2.1.1. Đồ thị vô hướng

Việc áp dụng thuật toán TextRank vào đồ thị vô hướng cũng giống như với đồ thị có hướng. Có một điểm cần lưu ý, đó là trong đồ thị vô hướng thì số đỉnh vào bằng số đỉnh ra.



Hình 1 Đường cong hội tụ của phương pháp xếp hạng dựa trên đồ thị với đồ thị có hướng - vô hướng, có trọng số - không trọng số, 250 đỉnh và 250 cạnh

Trong hình 1 thì đường cong hội tụ cho đồ thị được sinh ngẫu nhiên với 250 đỉnh và 250 cạnh, với ngưỡng dừng là 10^{-5} (ngưỡng này được xác định đủ nhỏ để thuật toán dừng tính toán) cho thấy số lần lặp của quá trình tính toán không cao mặc dù số lượng đỉnh và cạnh lớn. Bên cạnh đó thì đường cong độ tụ của đồ thị có hướng và vô hương gần như trùng nhau. Điều đó cho thấy đồ thị vô hướng hay có hướng đều cho kết quả giống nhau, chỉ khác nhau ở số lần tính toán lặp lại.

2.1.2. Đồ thị có trọng số

Gần như không có tình huống một trang web có nhiều liên kết đến một trang nào đó trong môi trường web. Vì vậy mà thuật toán Pagerank ban đầu chỉ sử dụng đồ thị không trọng số. Tuy nhiên đối với các văn bản trong ngôn ngữ tự nhiên thì việc một văn bản nào đó có nhiều thành phần tham chiếu đến một văn bản khác là hoàn toàn xảy ra. Do đó, để cải tiến Pagerank cho phù hợp với ngôn ngữ tự nhiên, thuật toán TextRank sử dụng đồ thị có trọng số. Trọng số ở đây được định nghĩa là độ dài kết nối giữa hai đỉnh V_i và V_j, ký hiệu w_{ij}. Từ đó suy ra,

công thức (2) cần phải được thay đổi để phù hợp với đồ thị có trọng số trong thuật toán TextRank. Ta được công thức mới như sau:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_i) \qquad (3)$$

Như vậy, cũng theo hình 1 ở trên thì số lần lặp lại tính toán để có độ tụ đạt ngưỡng 10^{-5} của đồ thị có trọng số và đồ thị không trọng số là tương đương nhau.

2.1.3. Đồ thị hoá văn bản

Văn bản là một chuỗi các ký tự / từ được sắp xếp với nhau. Vậy nên, để áp dụng được vào thuật toán dùng đồ thị thì cần phải đồ thị hoá văn bản. Việc đồ thị hoá văn bản là xây dựng một đồ thị để đại diện cho văn bản, các liên kết giữa các từ, cụm từ, câu hoặc các quan hệ ngữ nghĩa. Tuỳ thuộc vào các ứng dụng mà kích thước văn bản, các đặc trưng được đưa vào đồ thị là từ, cụm từ hay là cả câu. Cũng giống việc xác định các đỉnh trong đồ thị như trên thì việc xác định các cạnh trong đồ thị là gì cũng phụ thuộc vào miền ứng dụng. Quan hệ được xác định có thể là từ vựng, ngữ nghĩa hoặc ngữ cảnh.

Tùy vào các loại và đặc trưng để đưa vào đồ thị mà có các cách thức làm việc. Nhưng cách thức hoạt động của thuật toán xếp hạng dựa vào đồ thị áp dụng cho ngôn ngữ tự nhiên có các bước như sau:

- Xác định đơn vị văn bản dùng tốt nhất cho từng công việc, thêm vào là đỉnh của đồ thị.
- Xác định quan hệ kết nối giữa các đơn vị văn bản đã xác định ở trên để vẽ các cạnh giữa các đỉnh trong đồ thị. Các cạnh này có thể là vô hướng hoặc có hướng, có trọng số hoặc không trọng số.
- Lặp lại thuật toán xếp hạng cho đến khi độ tụ thoả mãn ngưỡng.
- Sắp xếp các đỉnh dựa trên các trọng số đã được tính toán trong bước trên.

Như vậy, thuật toán này giúp cho chúng ta làm được hai việc: trích rút từ khoá và trích rút câu trong văn bản ngôn ngữ tự nhiên. Vấn đề được đề cập ngay sau đây.

2.2. Sử dụng TextRank trích xuất từ khoá

Mục đích của việc trích xuất từ khoá tự động là tìm ra các cụm từ mô tả văn bản tốt nhất. Các từ khoá này có thể dùng cho nhiều mục đích khác nhau như phân lớp văn bản hay tóm tắt văn bản tự động. Trong các cách để trích xuất từ khoá thì cách trích xuất các từ khoá có tần suất xuất hiện nhiều nhất là dễ nhất. Mặc dù vậy thì kết quả của phương pháp này không tốt. Điều này đã thúc đẩy các nhà khoa học tìm ra các phương pháp khác hiệu quả hơn. Trong số đó có phương pháp sử dụng học máy có giám sát để trích xuất từ khoá dựa trên các đặc trưng về từ vựng và cú pháp. Phương pháp này lần đầu tiên được biết đến vào năm 1999, trong đó việc kết hợp tham số hoá các nguyên tắc phỏng đoán và thuật toán di truyền vào hệ thống trích xuất từ khoá sẽ tự động nhận dạng các từ khoá trong tài liệu. Một thuật toán khác cũng được đưa ra trong năm 1999 sử dụng phương pháp học máy Naive Bayes đã nâng cao chất lượng từ khoá trích rút được.

Năm 2003, Hulth đã dùng hệ thống học máy giám sát để trích xuất từ khoá kết hợp cả các đặc trưng về từ vựng và cú pháp. Trong nghiên cứu của mình, Hulth chỉ sử dụng bản tóm lược để trích xuất ra từ khoá thay vì toàn văn vì theo bà, văn bản trên Internet tồn tại chủ yếu ở dạng tóm lược. Đối với thuật toán TextRank, việc trích xuất từ khoá cũng được thực hiện đối với văn bản tóm lược. Mặc dù vậy thì việc áp dụng cho toàn văn hoàn toàn khả thi.

Đơn vị để xếp hạng trong thuật toán TextRank đối với quá trình trích xuất từ khoá là chuỗi của một hoặc nhiều từ vựng được rút ra từ văn bản và chúng là các đỉnh trong đồ thị. Bất kỳ quan hệ nào nữa 2 đơn vị từ vựng hữu ích cho việc đánh giá thì đều được thêm vào là cạnh của đồ thị. Ở đây ta sử dụng quan hệ đồng xuất hiện, nó được xác định bởi khoảng cách giữa các từ đồng xuất hiện

trong văn bản; hai đỉnh được xác định là nối với nhau khi khoảng cách đồng xuất hiện của hai đơn vị từ vựng không quá N từ với $2 \leq N \leq 10$. Các liên kết đồng xuất hiện thể hiện mối quan hệ giữa các yếu tố cú pháp, nó cũng tương tự như các liên kết ngữ nghĩa để tìm ra từ có nghĩa nhập nhằng, chúng đại diện cho các chỉ số của một văn bản.

Các đỉnh được thêm vào đồ thị bị giới hạn bởi các bộ lọc ngữ nghĩa, nó chỉ chọn các đơn vị từ vựng phù hợp, ví dụ như chọn danh từ, động từ và tạo các cạnh kết nối giữa các danh từ và động từ đó. Từ đó, ta tạo ra nhiều bộ lọc ngữ nghĩa để cho kết quả tốt hơn.

Thuật toán trích xuất từ khoá TextRank là thuật toán hoàn toàn không giám sát. Cách thức hoạt động như sau:

- Tách từ và gán nhãn, có các bộ lọc ngữ nghĩa. Để tránh gia tăng kích thước đồ thị thì áp dụng các đơn vị từ vựng phải có độ dài nhất định (n-gram).

- Đưa tất cả các đơn vị từ vựng có ở bước trên vào đồ thị. Các cạnh được đưa vào để liên kết các đơn vị từ vựng đồng xuất hiện với khoảng cách N từ. Sau khi dựng xong đồ thị (vô hướng, không trọng số) thì khởi tạo trọng số cho các đỉnh giá trị là 1. Và theo hình 1 thì số lần lặp lại từ 20 – 30 của thuật toán sẽ cho kết quả đạt ngưỡng 10^{-5}.

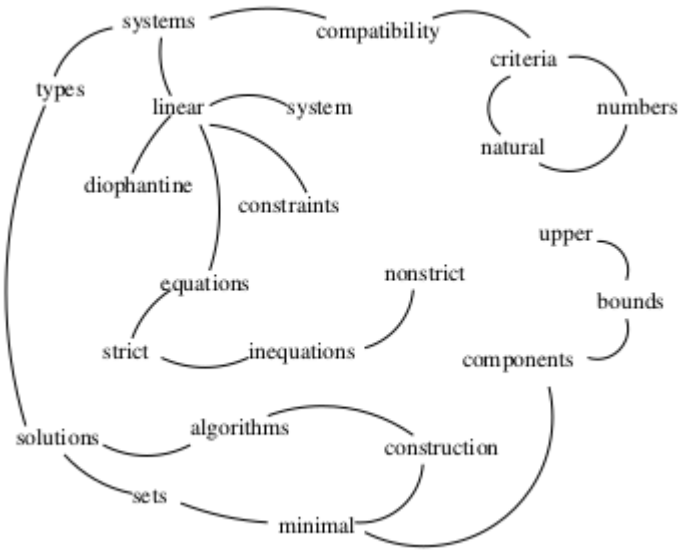
- Sau khi có kết quả cho mỗi đỉnh thì thực hiện quá trình sắp xếp ngược trọng số. *T* đỉnh đầu tiên sẽ được đưa vào quá trình tiếp theo, $5 \leq T \leq 20$. Ở đây thì *T* được lấy theo kích thước văn bản đầu vào.

- Sau bước trên ta được một tập các đơn vị từ vựng. Các đơn vị liền kề nhau thì được ghép lạ với nhau để tạo thành từ khoá dài.

Ta có ví dụ văn bản sau:

"Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types."

Đồ thị của nó sẽ có dạng:



Hình 2 Đồ thị thể hiện mối quan hệ giữa các đơn vị từ vựng trong văn bản

Từ khoá đưa ra bởi TextRank:

linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Từ khoá do con người đưa ra thủ công:

linear constraints; linear diophantine equations; minimal generating sets; nonstrict inequations; set of natural numbers; strict inequations; upper bounds

Như ví dụ trên, với bản tóm tắt có độ dài 120 từ thì số từ khoá thuật toán đưa ra không quá nhiều. Các đơn vị từ vựng có điểm số cao khi áp dụng TextRank là:

numbers (1.46)	inequations (1.45)	linear (1.29)	diophantine (1.28)
upper (0.99)	bounds (0.99)	strict (0.77)	

Ở đây cần chú ý là, điểm số của TextRank khác với tần suất xuất hiện của đơn vị từ vựng trong vă bản. Các từ xuất hiện nhiều là: systems (4), types (3), solutions (3), minimal (3), linear (2), inequations (2), algorithms (2).

Đánh giá

Tập dữ liệu được dùng để đánh giá bao gồm 500 văn bản tóm lược từ CSDL Inspec và bao gồm cả các từ khoá được con người xác định. Đây là tập dữ liệu giống với tập dữ liệu mà Hulth sử dụng trong báo cáo của mình năm 2003. Tập dữ liệu tóm lược Inspec là các bài báo được lấy từ tạp chí Khoa học máy tính và Công nghệ thông tin. Mỗi văn bản kèm theo là hai tập từ khoá: tập từ khoá được kiểm soát, giới hạn trong từ điển; tập thứ hai không bị giới hạn.

Hulth sử dụng 2000 văn bản, trong đó 1000 văn bản dùng để học, 500 văn bản dùng để phát triển và 500 văn bản dùng để kiểm tra. Đối với thuật toán TextRank, do là phương pháp không giám sát hoàn toàn nên không yêu cầu chia thành các tập dữ liệu học và phát triển.

Kết quả được thể hiện trong bảng 1. Đánh giá kết quả thu được trong bảng 1, ta nhận thấy TextRank cho các kết quả tốt hơn nhiều. Hulth sử dụng tập bốn đặc trưng: tần suất xuất hiện trong văn bản, tần suất xuất hiện tập, vị trí tương đối của lần xuất hiện đầu tiên, chuỗi thành phần của các tag ngôn từ. Các từ khoá được chia vào trong các nhóm: N-grams, NP-chunks (cụm danh từ), các mẫu. Với TextRank, thuật toán sử dụng sự đồng xuất hiện của các đơn vị từ vựng với các khoảng cách là 2, 3, 5 và 10 từ.

Như vậy, dựa trên các kết quả, ta có một số nhận xét về TextRank:

- Kết quả tốt hơn trong về độ chính xác và điểm F
- Độ hồi tưởng lại không cao bằng phương pháp học giám sát. Nguyên nhân có thể do áp đặt số lượng từ khoá được chọn.
- Khoảng cách từ trong thước đo đồng xuất hiện không phải càng lớn càng tốt.

Các bộ lọc được sử dụng trong ví dụ trên:

- Tất cả lớp từ mở
- Danh từ và tính từ
- Chỉ danh từ

Trong đó bộ lọc danh từ và tính từ cho kết quả tốt nhất.

Tóm lại, TextRank cho điểm F (F-score) cao hơn các phương pháp khác đã có trước đây. Điều quan trong hơn đó là, TextRank là thuật toán không giám sát, nó sử dụng các thông tin được trích rút ra từ chính văn bản được áp dụng. Chính điều này làm cho thuật toán này có thể dễ dàng áp dụng cho các tập dữ liệu, miền ứng dụng và ngôn ngữ khác nhau.

Phương pháp	Trích xuất		Chính xác				
	Tổng	Tỷ lệ	Tổng	Tỷ lệ	Precision	Recall	F-measure
TextRank							
Undirected, Co-occ.window=2	6784	13.7	2116	4.2	**31.2**	43.1	**36.2**
Undirected, Co-occ.window=3	6715	13.4	1897	3.8	28.2	38.6	32.6
Undirected, Co-occ.window=5	6558	13.1	1851	3.7	28.2	37.7	32.2
Undirected, Co-occ.window=10	6570	13.1	1846	3.7	28.1	37.6	32.2
Directed, forward, Co-occ.window=2	6662	13.3	2081	4.1	31.2	42.3	35.9
Directed, backward, Co-occ.window=2	6636	13.3	2082	4.1	31.2	42.3	35.9
Hulth – 2003							
Ngram with tag	7815	15.6	1973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4788	9.6	1421	2.8	29.7	37.2	33
Pattern with tag	7012	14	1523	3.1	21.7	39.9	28.1

Bảng 1 So sánh kết quả trích xuất từ khoá giữa TextRank và Hulth 2003

2.3. Sử dụng TextRank trích rút câu

Đặc điểm của TextRank là sử dụng đồ thị, nên để áp dụng được TextRank thì cần phải đồ thị hoá văn bản. Muốn trích rút được câu thì cần phải xếp hạng được các câu trong văn bản trên toàn đồ thị. Vì thế, mỗi câu sẽ là một đỉnh của đồ thị. Có một điểm cần lưu ý, quan hệ đồng xuất hiện không thể áp dụng trong trường hợp này. Đơn giản vì không tồn tại các câu giống nhau 100% trong toàn văn bản. Thay vào đó, một định nghĩa quan hệ khác được đưa ra để xác định kết nối giữa các câu với nhau. Đó là độ tương tự giữa các câu. Ở đây, độ tương tự được xác định bằng độ bao phủ về mặt nội dung giữa các câu với nhau. Mối quan hệ giữa hai câu đó được xem là một "đề cử": một câu đề cập đến một khái niệm nào đó trong văn bản sẽ "đề cử" cho độc giả một câu khác trong văn bản cũng đề cập đến khái niệm đó. Do đó xuất hiện một liên kết giữa các câu có chung nôi dung.

Độ bao phủ của hai câu có thể đo bằng số lượng từ trùng nhau giữa hai câu hoặc có thể chạy chung một hoặc nhiều bộ lọc ngữ nghĩa, cú pháp. Để giảm giá trị của độ tương đồng giữa các câu, tạo thuận lợi trong quá trình tính toán dotồn tại các câu dài thì TextRank sử dụng hệ số chuẩn hoá là chia số lượng nội dung bao phủ cho độ dài của từng câu.

Với hai câu S_i và S_j với một câu được đại diện bởi một tập N_i các từ xuất hiện trong câu: $S_i = w_1^i, w_2^i, \ldots, w_{N_i}^i$

Độ tương đồng giữa hai câu S_i và S_j được tính theo công thức:

$$Similarity(S_i, S_j) = \frac{\left|\{w_k | w_k \in S_i \wedge w_k \in S_j\}\right|}{log(|S_i|) + log(|S_j|)}$$

Trong đó:

- $Similarity(S_i, S_j)$: là độ tương đồng giữa câu S_i và câu S_j
- w_k: từ thuộc cả câu S_i và câu S_j
- $log(|S_i|) + log(|S_j|)$: hệ số chuẩn hóa

Có nhiều độ đo độ tương tự giữa các câu với nhau như chuỗi nhân, cô-sin, chuỗi con chung dài nhất, … đều có thể được sử dụng. Trong luận văn, độ đo cô-sin được dùng để đối sánh kết quả.

Dưới đây là một văn bản mẫu được đưa vào thử nghiệm:

1. "Ông lớn" hàng điện tử Nhật chịu án phạt nghìn tỷ vì thao túng giá

2. Hãng Panasonic của Nhật cùng với công ty con Sanyo của tập đoàn này hôm qua đã bị Bộ tư pháp Mỹ kết án tham gia thao túng giá thiết bị ô tô và pin máy tính, buộc "ông lớn" này nộp phạt 56,5 triệu USD, tương đương 1200 tỷ đồng.

3. Thông tin được hãng tin AFP đăng tải.

4. Một công ty nữa cũng phải nhận án phạt là LG Chem LTD của Hàn Quốc do bị kết luận thao túng giá pin, và phải nộp khoản tiền phạt hơn 1 triệu USD, thông báo của Bộ tư pháp Mỹ cho biết.

5. Theo các bản cáo trạng chống lại Panasonic, trong khoảng thời gian từ tháng 9/2003 đến tháng 2/2010, tập đoàn của Nhật này "đã tham gia vào một âm

mưu gian lận các cuộc đấu thầu, thao túng, giữ ổn định và duy trì giá của" các phụ tùng ô tô.

6. Các phụ tùng này - bao gồm các công tắc tay lái và cảm biến góc lái – được bán cho hãng xe Toyota tại Mỹ và một số nơi khác, Bộ tư pháp Mỹ cho biết.

7. Kể từ năm 1998, công ty này cũng tham gia thao túng giá các thiết bị kiểm soát hiệu điện thế đèn ô tô, được bán cho các hãng xe Honda, Mazda và Nissan.

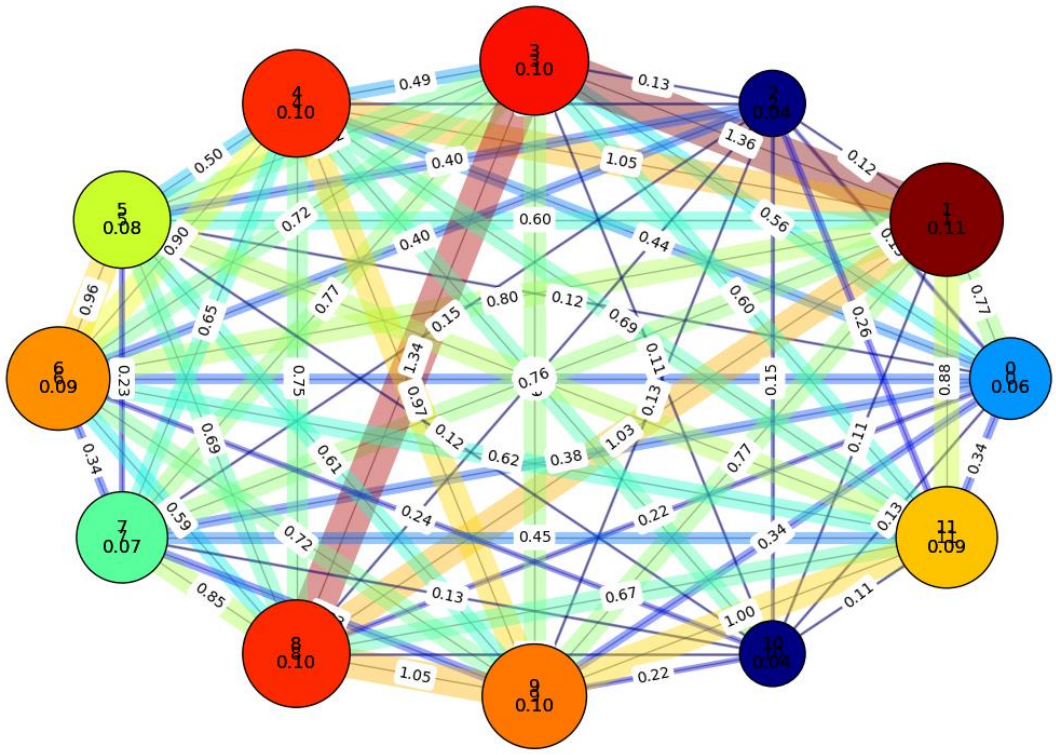
8. Panasonic sẽ phải nộp khoản phạt 45,8 triệu USD vì tham gia vào âm mưu này.

9. "Cùng với Panasonic, 11 công ty và 15 lãnh đạo đã bị xác định có tội hoặc chấp thuận với kết luận có tội và đồng ý nộp số tiền phạt tổng cộng hơn 874 triệu USD sau cuộc điều tra các phụ tùng ô tô", thông báo cho biết thêm và khẳng định các lãnh đạo này đã bị hoặc sẽ bị kết án phạt tù.

10. Trong một vụ án khác liên quan đến Sanyo và LG Chem, hai công ty này bị xác định đã thỏa thuận "trong các cuộc họp và đối thoại", sẽ đặt ra các mức giá cho các loại pin sạc sử dụng trong máy tính xách tay.

11. Sự vi phạm xảy ra trong quãng thời gian từ tháng 4/2007 – 9/2008.

12. "Sanyo, LG Chem và các đồng phạm đã thu thập và trao đổi thông tin vì mục đích theo dõi và áp đặt các mức giá đã được họ thỏa thuận trước và có những động thái để che giấu âm mưu này", Bộ tư pháp Mỹ khẳng định.



Hình 3 Đồ thị mô phỏng các kết nối giữa các cập câu trong văn bản

Kết quả tóm tắt tự động với độ dài 4 câu:

2. Hãng Panasonic của Nhật cùng với công ty con Sanyo của tập đoàn này hôm qua đã bị Bộ tư pháp Mỹ kết án tham gia thao túng giá thiết bị ô tô và pin máy tính, buộc "ông lớn" này nộp phạt 56,5 triệu USD, tương đương 1200 tỷ đồng.

4. Một công ty nữa cũng phải nhận án phạt là LG Chem LTD của Hàn Quốc do bị kết luận thao túng giá pin, và phải nộp khoản tiền phạt hơn 1 triệu USD, thông báo của Bộ tư pháp Mỹ cho biết.

9. "Cùng với Panasonic, 11 công ty và 15 lãnh đạo đã bị xác định có tội hoặc chấp thuận với kết luận có tội và đồng ý nộp số tiền phạt tổng cộng hơn 874 triệu USD sau cuộc điều tra các phụ tùng ô tô", thông báo cho biết thêm và khẳng định các lãnh đạo này đã bị hoặc sẽ bị kết án phạt tù.

8. Theo các bản cáo trạng chống lại Panasonic, trong khoảng thời gian từ tháng 9/2003 đến tháng 2/2010, tập đoàn của Nhật này "đã tham gia vào một âm mưu gian lận các cuộc đấu thầu, thao túng, giữ ổn định và duy trì giá của" các phụ tùng ô tô.

Trong đồ thị ở hình 3, trọng số trên các cạnh là chỉ số đo mức độ kết nối giữa các cặp câu với nhau. Đồ thị thể hiện các câu có kết nối với nhau như thế nào. Câu nào có càng nhiều câu khác kết nối với thì nó có trọng số càng lớn. Điểm TextRank được tính trên mỗi đỉnh trong đồ thị được tạo ra. Từ đó, các câu có điểm cao nhất sẽ được lựa chọn để tạo thành văn bản tóm tắt. Việc lựa chọn số câu được lấy ra phụ thuộc vào yêu cầu của độ dài văn bản tóm tắt. Trong ví dụ trên, với yêu cầu độ dài văn bản tóm tắt là 4 câu nên các câu số 2, 4, 9, 8 được lựa chọn để tạo thành văn bản tóm tắt. Với 4 câu được trích xuất làm văn bản tóm tắt thì tỉ lệ nén ($CompressionRate$) = 1/3. Ở trong luận văn chỉ sử dụng đến tỉ lệ nén để đánh giá mức độ rút gọn của văn bản tóm tắt so với văn bản gốc.

Đánh giá

Theo Rada Mihalcea và Paul Tarau, cả hai đánh giá thuật toán TextRank bằng cách sử dụng 567 bài báo trong bộ DUC 2002. Mỗi một bài báo, hai tác giả thuật toán sẽ dùng TextRank để lấy ra bản tóm tắt có độ dài 100 từ. Để đánh giá, tác giả đã sử dụng bộ công cụ ROUGE với phương pháp sử dụng là thống kê n-grams. Kết quả đánh giá của TextRank được so sánh với 5 hệ thống tóm tắt văn bản có kết quả tốt nhất trong số 15 hệ thống tóm tắt văn bản được đưa vào đánh giá. Kết quả so sánh được thể hiện trong bảng dưới đây.

Hệ thống	**Điểm ROUGE - Ngram(1,1)**		
	Cơ bản	**Từ gốc**	**Từ gốc bỏ từ dừng**
S27	0.4814	0.5011	0.4405

S32	0.4715	0.4914	0.4160
TextRank	**0.4708**	**0.4904**	**0.4229**
S28	0.4703	0.4890	0.4346
S21	0.4683	0.4869	0.4222
Baseline	0.4599	0.4779	0.4162
S29	0.4502	0.4681	0.4019

Bảng 2 Kết quả so sánh tóm tắt đơn giữa TextRank và các hệ thống khác

Như vậy, TextRank đã thành công khi nhận dạng được các câu quan trọng trong văn bản mà chỉ dựa vào các thông tin có trong văn bản. Không giống như các hệ thống học có giám sát cần phải học để cải thiện chất lượng bản tóm tắt dựa trên tập dữ liệu đã tóm tắt của các văn bản khác, TextRank hoàn toàn không giám sát, chỉ dựa vào thông tin có được trong văn bản để đưa ra văn bản tóm tắt. Văn bản tóm tắt của TextRank cũng gần với văn bản tóm tắt của người làm.

Cũng cần lưu ý là TextRank vượt ra ngoài việc sử dụng kết nối câu trong văn bản. Chính vì vậy mà một số câu trong văn bản có lượng kết nối đến các câu khác ít nhưng vẫn được TextRank lựa chọn để trích xuất đưa vào văn bản tóm tắt, những câu đó cũng thường xuất hiện trong các văn bản do người tóm tắt. Đây là điểm làm cho TextRank trở nên giống người hơn khi đưa ra các văn bản tóm tắt.

Bên cạnh đó, TextRank còn xếp hạng các câu trong toàn bộ văn bản. Vì vậy, việc điều chỉnh độ dài của văn bản tóm tắt rất đơn giản. Điểm đặc biệt khác của TextRank là phương pháp này có thể áp dụng được cho nhiều ngôn ngữ khác nhau mà việc tùy chỉnh đơn giản và không quá phức tạp.

2.4. Tóm tắt văn bản Tiếng Việt sử dụng TextRank

2.4.1. Một số đặc trưng của Tiếng Việt

Trước khi trình bày chi tiết quá trình tóm tắt văn bản Tiếng Việt sử dụng TextRank, tác giả xin trình bày về các đặc điểm của ngôn ngữ Tiếng Việt. Các đặc điểm của Tiếng Việt như sau:

❖ Tiếng Việt thuộc loại hình ngôn ngữ đơn lập không biến đổi hình thái.

- Về mặt ngữ âm, đơn vị trong Tiếng Việt là "tiếng" hoặc "chữ" tùy theo ngữ âm hoặc văn tự. "tiếng" ở đây khi được phát âm là một âm tiết.
- Đơn vị để cấu tạo từ là "hình vị". Đối với Tiếng Việt, "hình vị" chính là "tiếng". Về ngữ pháp, "tiếng" cũng được xem là đơn vị cơ sở của ngữ pháp học.
- Từ trong Tiếng Việt không biết đổi hình thái trong khi sử dụng. Ví dụ: "Tôi yêu cô ấy" và "Cô ấy yêu tôi" là hai câu trong Tiếng Việt. Các từ "tôi" và "cô ấy" dù đứng ở vị trí chủ ngữ hay bổ ngữ đều không biến đổi hình thái. Ngay cả động từ "yêu" cũng không biến đổi hình thái theo ngôi hoặc theo số ít hoặc số nhiều của chủ ngữ.
- Cách sắp xếp các từ theo trật tự nhất định dùng để biểu thị quan hệ cú pháp. Khi trật tự từ thay đổi thì bản chất ngữ pháp cũng thay đổi. Nhà thơ Phan Thị Thanh Nhàn có viết:

"Người tôi yêu đã đi xa

Người yêu tôi lại ở nhà... chán không!"

Từ "tôi" và "yêu" đã đổi vị trí của nhau trong hai câu làm cho ý nghĩa hoàn toàn thay đổi.

- Từ trong Tiếng Việt về mặt cấu tạo bao gồm từ đơn âm tiết và từ đa âm tiết. Từ đa âm tiết được hình thành từ việc ghép các đơn âm tiết với nhau. Ví dụ: xe + máy ➔ xe máy, trường + học ➔ trường học, …

❖ Chữ viết của Tiếng Việt là chữ ghi âm vị

- Chữ ghi âm vị là chữ gồm các con chữ ghi từng đơn vị ngữ âm nhỏ nhất, nghĩa là mỗi kí hiệu biểu thị một âm vị.

❖ Hiện tượng đồng âm trong Tiếng Việt khá phổ biến. Ví dụ, cùng âm tiết "gấu" nhưng trong từ "gấu áo" và "con gấu" lại biểu thị hai nghĩa hoàn toàn khác nhau.

❖ Tiếng Việt không có dấu hiệu về hình thái để nhận biết từ loại. Ví dụ: "Tôi lấy cân để cân gạo". Từ "cân" trong hai vị trí là hai từ loại khác nhau. Từ "cân" đầu tiên là danh từ, từ "cân" tiếp theo lại là động từ.

❖ Hiện tượng gần âm khác nghĩa cũng xảy ra khá phổ biến. Ví dụ: bàng quan và bàng quang; cao tần và cao tầng; bàn bạc và bàng bạc; …

❖ Hiện tượng gần nghĩa, đồng nghĩa giữa các từ nhưng lại không thể thay thế cho nhau trong từng hoàn cảnh sử dụng cũng xảy ra phổ biến.

Ví dụ:

- đề bạt, đề cử, đề đạt, đề xuất, đề nghị;

- chủ tịch, chủ trì, chủ toạ, chủ nhiệm;

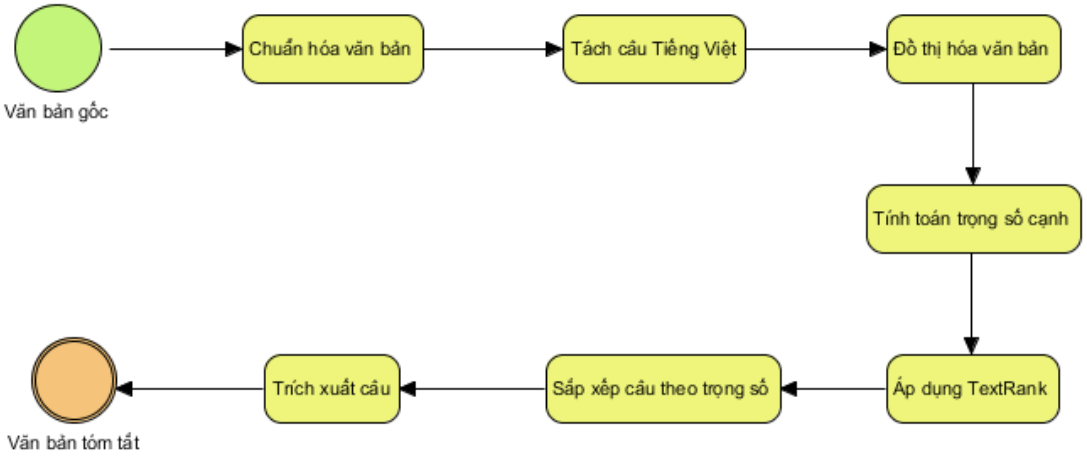
- hội đàm, hội nghị, hội thảo, toạ đàm;

2.4.2. Xây dựng hệ thống tóm tắt tự động văn bản Tiếng Việt

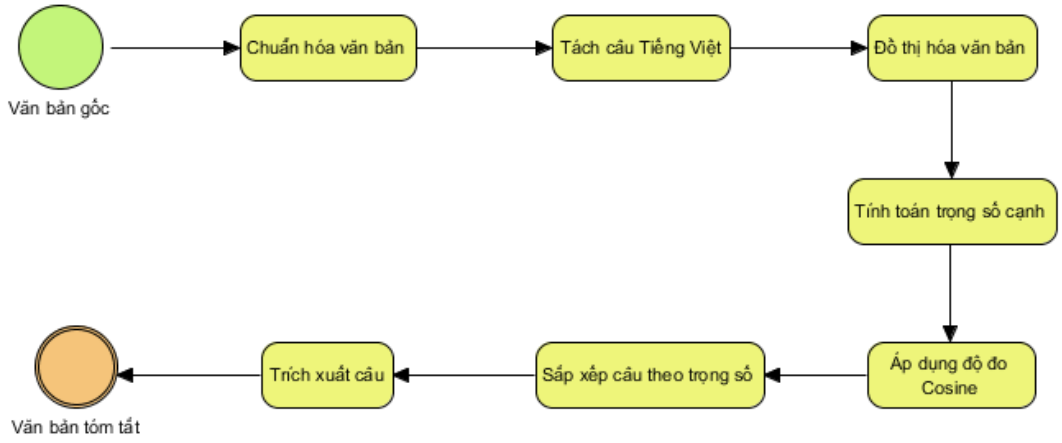
Như đã trình bày ở phần 2.1 đến 2.3, TextRank chỉ sử dụng các thông tin có trong văn bản tóm tắt mà không cần quan tâm để cấu trúc văn bản, các thành phần của văn phạm, ngôn ngữ. Đây là điểm mấu chốt để tác giả lựa chọn thuật toán này cho bài toán tóm tắt văn bản.

Để có cái nhìn khách quan và đánh giá chính xác hơn chất lượng văn bản tóm tắt được sinh ra bởi thuật toán TextRank, tác giả sẽ đưa vào một hệ thống đối sánh sử dụng độ đo mức độ tương tự giữa các câu sử dụng Cosine (Cô-sin). Mô hình của hệ thống tóm tắt văn bản sử dụng TextRank và Cosine là giống nhau. Điểm khác biệt quan trọng của 2 mô hình này là khi sử dụng độ đo xác định mức độ tương đồng giữa các câu.

Mô hình tóm tắt văn bản Tiếng Việc dựa trên thuật toán TextRank và Cô-sin được xây dựng như hình dưới.



Hình 4 Mô hình tóm tắt văn bản Tiếng Việt sử dụng TextRank



Hình 5 Mô hình tóm tắt văn bản Tiếng Việt sử dụng Cosine

Ở đây, tác giả sẽ mô tả chi tiết cách thức xây dựng hệ thống tóm tắt tự động văn bản Tiếng Việt sử dụng TextRank (hệ thống). Hình 4 cho biết mô hình chuẩn của hệ thống. Hệ thống tóm tắt văn bản tự động được xây dựng trên ngôn ngữ lập trình Python, phiên bản 3.4. Các thư viện được sử dụng trong chương trình được liệt kê trong phụ lục 2. Hệ thống bao gồm các thành phần:

i. Bộ chuẩn hóa văn bản Tiếng Việt, có nhiệm vụ chuẩn hóa văn bản đầu vào của văn bản. Loại bỏ các ký tự thừa, các thành phần không thuộc văn bản tiếng Việt.

ii. Bộ tách câu tiếng Việt, bộ này có nhiệm vụ nhận dạng câu Tiếng Việt và tách thành các câu riêng biệt. Có một số ký tự được ưu tiên khi tách câu: ký tự dấu chấm ".", ký tự lùi về đầu dòng "LF", ký tự xuống dòng "CR",… Tuy nhiên, trong một số trường hợp ký tự dâu ba chấm "…" không phải dùng để kết thúc câu.

iii. Chương trình đồ thị hóa văn bản Tiếng Việt. Chương trình này dùng để chuyển đổi văn bản sau khi tách các câu sang dạng đồ thị vô hướng và có trọng số. Trọng số của đồ thì được tính toán trong bước tiếp theo của hệ thống. Mỗi đỉnh trong đồ thị là một câu trong văn bản. Cạnh nối giữa 2 đỉnh của đồ thị thể hiện mức độ tương đồng của cặp câu tương ứng.

iv. Bộ tính toán TextRank và trọng số. Bộ này được cài đặt thuật toán TextRank để xác định mức độ tương đồng giữa các câu trong văn bản. Từ đó bổ sung vào đồ thị văn bản đã được dựng lên từ bước trước các trọng số cạnh. Trọng số của cạnh càng lớn thì mức độ tương đồng càng cao. Một đỉnh có nhiều đỉnh khác nối đến không có nghĩa là đỉnh đó quan trọng hơn trong đồ thị. Độ quan trọng được tính toán theo thuật toán PageRank dựa trên trọng số cạnh nối đến đỉnh đó. Các thuật toán được cài đặt theo công thức (2) và (3).

v. Bộ trích rút câu và hợp thành văn bản. Bộ này sẽ lấy kế quả từ bước tính toán trước đó – đồ thị vô hướng có trọng số cạnh và trọng số đỉnh. Bộ trích rút sẽ sắp xếp cách đỉnh – tương ứng với các câu - theo thứ tự giá trị trọng số tại đỉnh đó giảm dần. Trọng số tại đỉnh càng cao thì mức độ quan trọng của đỉnh đó càng cao trong đồ thị (văn bản). Từ đó, tùy theo yêu cầu của bài toàn tóm tắt văn bản với độ dài bao nhiêu hoặc tỉ lệ nén là bao nhiêu mà bộ hợp thành văn bản

sẽ lấy ra số lượng câu phù hợp để hợp thành văn bản tóm tắt. Đơn vị đo độ dài của văn bản trong hệ thống tóm tắt tự động văn bản là "câu".

Thử nghiệm với văn bản mẫu:[3]

Tăng cường ngăn chặn gia cầm nhập lậu

Với chức năng là cơ quan thường trực đề án "Phòng ngừa, ngăn chặn vận chuyển và kinh doanh gia cầm, sản phẩm gia cầm nhập khẩu trái phép", Phòng Cảnh sát phòng chống tội phạm về môi trường (PCTPMT) Công an tỉnh Lạng Sơn đã xây dựng kế hoạch mở cao điểm, tập trung đấu tranh ngăn chặn hoạt động nhập lậu gia cầm, vận chuyển, kinh doanh gia cầm không rõ nguồn gốc.

Đơn vị đã thành lập 4 tổ công tác, huy động 50% quân số tăng cường kiểm soát liên tục trên các tuyến, góp phần cùng nhiều lực lượng khác đấu tranh ngăn chặn nên đến thời điểm này, thực trạng buôn bán, vận chuyển gia cầm nhập lậu đã được Công an tỉnh Lạng Sơn kiểm soát chặt chẽ. Giám đốc Công an tỉnh Lạng Sơn đã chỉ đạo công an các huyện biên giới đẩy mạnh các biện pháp nghiệp vụ, lập hồ sơ những đối tượng có biểu hiện hoạt động buôn bán gia cầm nhập lậu để quản lý giáo dục, tuyên truyền đến các hộ dân thường xuyên tham gia vận chuyển gia cầm ký cam kết không vi phạm.

Thượng tá Hoàng Văn Nguyên, Phó Trưởng phòng Cảnh sát PCTP về môi trường Công an tỉnh Lạng Sơn cho biết: "Do các lực lượng đấu tranh mạnh, nên hoạt động của các đối tượng mua bán, vận chuyển gà nhập lậu qua các tuyến biên giới Cao Lộc, Văn Lãng, Tràng Định (Lạng Sơn) cơ bản được kiểm soát. Tuy nhiên, vẫn còn dấu hiệu lén lút vận chuyển qua một số đường mòn khu vực biên giới Việt - Trung thuộc địa bàn huyện Lộc Bình".

[3] Văn bản có tên XH26.txt trong tập dữ liệu đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt", TS. Lê Thanh Hương, Hà Nội, 2014

Đáng chú ý, khi các loại gia cầm sống khó lọt vào nội địa bởi nhiều tầng kiểm soát của các lực lượng chức năng, thì các đối tượng đã thay đổi phương thức vận chuyển bằng cách đưa chim bồ câu được thịt sẵn đóng vào thùng xốp, bọc bên ngoài bằng thùng carton, rồi theo xe khách để vận chuyển vào sâu trong nội địa. Ngày 16-4 vừa qua, tại km số 38 - Quốc lộ 1A, Công an tỉnh Lạng Sơn kiểm tra xe ôtô khách (BKS: 12B - 000.04) do Nguyễn Xuân Quý, SN 1971, trú tại thị trấn Lộc Bình, tỉnh Lạng Sơn điều khiển, phát hiện trên xe có 2 thùng xốp chứa 74kg bồ câu thịt sẵn và một số hàng hóa khác có xuất xứ từ Trung Quốc, không có giấy tờ chứng minh nguồn gốc và không có giấy tờ kiểm dịch theo quy định.

Nhằm đấu tranh ngăn chặn có hiệu quả hơn nguồn gia cầm nhập lậu, được xác định là nguyên nhân lây nhiễm chủng virus cúm A, có xuất xứ từ Trung Quốc, Công an tỉnh Lạng Sơn tiếp tục đẩy mạnh các biện pháp kiểm soát chặt chẽ ngay từ biên giới. Tham mưu cho cấp ủy, chính quyền các xã biên giới tổ chức tuyên truyền, ký cam kết không vi phạm chứa chấp, vận chuyển gia cầm nhập lậu trong nhân dân.

Văn bản đã tách câu và tách từ:

1. Tăng cường ngăn_chặn gia_cầm nhập lậu

2. Với chức_năng là cơ_quan thường_trực đề_án “Phòng_ngừa, ngăn_chặn vận_chuyển và kinh_doanh gia_cầm, sản_phẩm gia_cầm nhập_khẩu trái_phép”, Phòng Cảnh_sát phòng_chống tội_phạm về môi_trường (PCTPMT) Công_an tỉnh Lạng_Sơn đã xây_dựng kế_hoạch mở cao_điểm, tập_trung đấu_tranh ngăn_chặn hoạt_động nhập lậu gia_cầm, vận_chuyển, kinh_doanh gia_cầm không rõ nguồn_gốc.

3. Đơn_vị đã thành_lập 4 tổ công_tác, huy_động 50 % quân_số tăng_cường kiểm_soát liên_tục trên các tuyến, góp_phần cùng nhiều lực_lượng khác

đấu_tranh ngăn_chặn nên đến thời_điểm này, thực_trạng buôn_bán, vận_chuyển gia_cầm nhập lậu đã được Công_an tỉnh Lạng_Sơn kiểm_soát chặt_chẽ.

4. Giám_đốc Công_an tỉnh Lạng_Sơn đã chỉ_đạo công_an các huyện biên_giới đẩy_mạnh các biện_pháp nghiệp_vụ, lập hồ_sơ những đối_tượng có biểu_hiện hoạt_động buôn_bán gia_cầm nhập lậu để quản_lý giáo_dục, tuyên_truyền đến các hộ dân thường_xuyên tham_gia vận_chuyển gia_cầm ký cam_kết không vi_phạm.

5. Thượng_tá Hoàng_Văn_Nguyên, Phó_Trưởng_phòng Cảnh_sát PCTP về môi_trường Công_an tỉnh Lạng_Sơn cho biết: "Do các lực_lượng đấu_tranh mạnh, nên hoạt_động của các đối_tượng mua_bán, vận_chuyển gà nhập lậu qua các tuyến biên_giới Cao_Lộc, Văn_Lãng, Tràng_Định (Lạng_Sơn) cơ_bản được kiểm_soát.

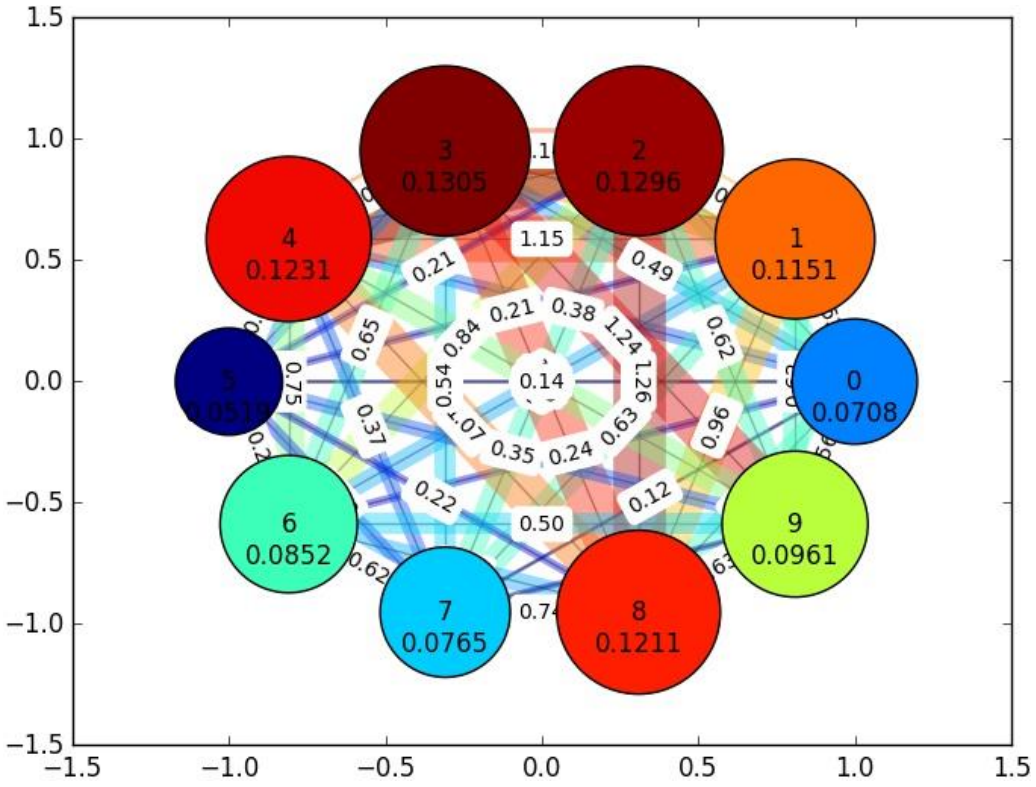
6. Tuy_nhiên, vẫn còn dấu_hiệu lén_lút vận_chuyển qua một_số đường_mòn khu_vực biên_giới Việt - Trung thuộc địa_bàn huyện Lộc_Bình".

7. Đáng chú_ý, khi các loại gia_cầm sống khó lọt vào nội_địa bởi nhiều tầng kiểm_soát của các lực_lượng chức_năng, thì các đối_tượng đã thay_đổi phương_thức vận_chuyển bằng cách đưa chim bồ_câu được thịt sẵn đóng vào thùng xốp, bọc bên ngoài bằng thùng carton, rồi theo xe_khách để vận_chuyển vào sâu trong nội_địa.

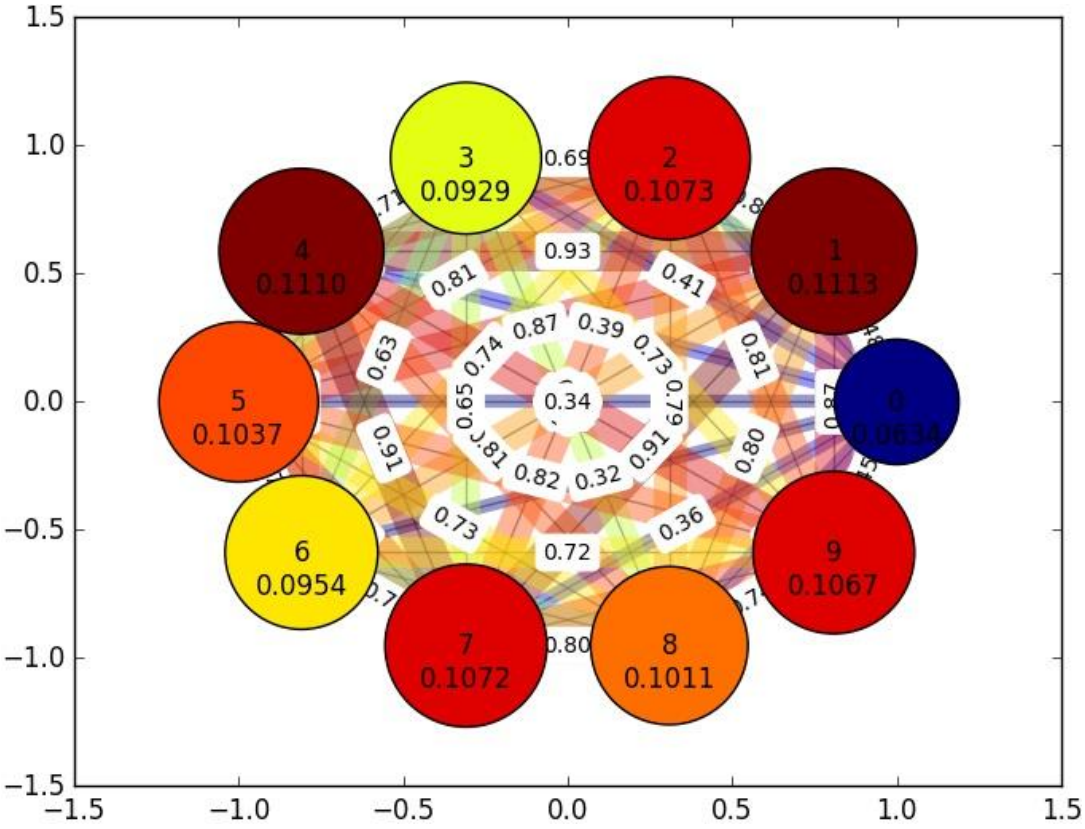
8. Ngày 16-4 vừa_qua, tại km số 38 - Quốc_lộ 1A, Công_an tỉnh Lạng_Sơn kiểm_tra xe ôtô khách (BKS: 12B - 000.04) do Nguyễn_Xuân_Quý, SN 1971, trú tại thị_trấn Lộc_Bình, tỉnh Lạng_Sơn điều_khiển, phát_hiện trên xe có 2 thùng xốp chứa 74kg bồ_câu thịt sẵn và một_số hàng_hoá khác có xuất_xứ từ Trung_Quốc, không có giấy_tờ chứng_minh nguồn_gốc và không có giấy_tờ kiểm_dịch theo quy_định.

9. Nhằm đấu_tranh ngăn_chặn có hiệu_quả hơn nguồn gia_cầm nhập lậu, được xác_định là nguyên_nhân lây_nhiễm chủng virus cúm A, có xuất_xứ từ Trung_Quốc, Công_an tỉnh Lạng_Sơn tiếp_tục đẩy_mạnh các biện_pháp kiểm_soát chặt_chẽ ngay từ biên_giới.

10. Tham_mưu cho cấp_uỷ, chính_quyền các xã biên_giới tổ_chức tuyên_truyền, ký cam_kết không vi_phạm chứa_chấp, vận_chuyển gia_cầm nhập lậu trong nhân_dân.



Hình 6 Đồ thị mô phỏng quan hệ giữa các câu trong văn bản mẫu sử dụng TextRank



Hình 7 Đồ thị mô phỏng quan hệ giữa các câu trong văn bản mẫu sử dụng Cosine

Từ đồ thị trên, ta tiến hành trích xuất câu để hợp thành văn bản tóm tắt. Ở đây tác giả lấy độ dài văn bản tóm tắt là 4 câu (tỉ lệ nén = 0.4). Ta có kết quả như sau:

TextRank	**Cosine**
Giám đốc Công an tỉnh Lạng Sơn đã chỉ đạo công an các huyện biên giới đẩy mạnh các biện pháp nghiệp vụ, lập hồ sơ những đối tượng có biểu hiện hoạt động buôn bán gia cầm nhập lậu để quản lý giáo dục, tuyên truyền đến các hộ dân thường xuyên tham gia vận chuyển gia cầm ký cam kết không vi phạm.	Với chức năng là cơ quan thường trực đề án "Phòng ngừa, ngăn chặn vận chuyển và kinh doanh gia cầm, sản phẩm gia cầm nhập khẩu trái phép", Phòng Cảnh sát phòng chống tội phạm về môi trường (PCTPMT) Công an tỉnh Lạng Sơn đã xây dựng kế hoạch mở cao điểm, tập trung đấu tranh ngăn chặn hoạt động nhập lậu gia cầm, vận

Đơn vị đã thành lập 4 tổ công tác, huy động 50 % quân số tăng cường kiểm soát liên tục trên các tuyến, góp phần cùng nhiều lực lượng khác đấu tranh ngăn chặn nên đến thời điểm này, thực trạng buôn bán, vận chuyển gia cầm nhập lậu đã được Công an tỉnh Lạng Sơn kiểm soát chặt chẽ. Thượng tá Hoàng Văn Nguyên, Phó Trưởng phòng Cảnh sát PCTP về môi trường Công an tỉnh Lạng Sơn cho biết: "Do các lực lượng đấu tranh mạnh, nên hoạt động của các đối tượng mua bán, vận chuyển gà nhập lậu qua các tuyến biên giới Cao Lộc, Văn Lãng, Tràng Định (Lạng Sơn) cơ bản được kiểm soát. Nhằm đấu tranh ngăn chặn có hiệu quả hơn nguồn gia cầm nhập lậu, được xác định là nguyên nhân lây nhiễm chủng virus cúm A, có xuất xứ từ Trung Quốc, Công an tỉnh Lạng Sơn tiếp tục đẩy mạnh các biện pháp kiểm soát chặt chẽ ngay từ biên giới.	chuyển, kinh doanh gia cầm không rõ nguồn gốc. Thượng tá Hoàng Văn Nguyên, Phó Trưởng phòng Cảnh sát PCTP về môi trường Công an tỉnh Lạng Sơn cho biết: "Do các lực lượng đấu tranh mạnh, nên hoạt động của các đối tượng mua bán, vận chuyển gà nhập lậu qua các tuyến biên giới Cao Lộc, Văn Lãng, Tràng Định (Lạng Sơn) cơ bản được kiểm soát. Đơn vị đã thành lập 4 tổ công tác, huy động 50 % quân số tăng cường kiểm soát liên tục trên các tuyến, góp phần cùng nhiều lực lượng khác đấu tranh ngăn chặn nên đến thời điểm này, thực trạng buôn bán, vận chuyển gia cầm nhập lậu đã được Công an tỉnh Lạng Sơn kiểm soát chặt chẽ. Ngày 16-4 vừa qua, tại km số 38 - Quốc lộ 1A, Công an tỉnh Lạng Sơn kiểm tra xe ôtô khách (BKS: 12B - 000.04) do Nguyễn Xuân Quý, SN 1971, trú tại thị trấn Lộc Bình, tỉnh Lạng Sơn điều khiển, phát hiện trên xe có 2 thùng xốp chứa 74kg bồ câu thịt sẵn và một số hàng hoá khác có xuất xứ từ Trung Quốc, không có giấy tờ

	chứng minh nguồn gốc và không có giấy tờ kiểm dịch theo quy định.

Nhận xét:

Dựa vào 2 đồ thị tại hình 6 và hình 7 có thể nhận thấy dễ dạng sự khác biệt giữa 2 cách đánh giá độ tương đồng giữa các câu và mức độ quan trọng của câu trong văn bản. Với TextRank, câu thứ 4 có mức độ quan trong cao nhất trong văn bản, trong khi đó, với Cosine, câu thứ 4 chỉ xếp vị trí 9.

Qua đọc nội dung thì thấy được chất lượng bản tóm tắt sử dụng TextRank có chất lượng tốt hơn Cosine. Văn bản đọc lên thấy trôi chảy về mặt nội dung, ý nghĩa. Tuy nhiên một số câu có mức độ thông tin phù hợp hơn lại không được lựa chọn do độ quan trong không cao. Do độ dài văn bản tóm tắt bị giới hạn ở mức 4 câu nên các câu có hàm lượng thông tin cao chưa được góp mặt.

Các câu dài thường được ưu tiên lựa chọn để trích xuất do khi tính toán độ tương đồng thì khả năng các câu này có độ tương đồng cao hơn so với các câu khác. Đây cũng là một nhược điểm trong thuật toán TextRank. Điều này làm giảm đi một phần chất lượng của văn bản tóm tắt.

Bên cạnh đó, do là trích rút câu, các câu không được chỉnh sửa nên trong câu sẽ tồn tại từ nối, quan hệ, từ không mang nhiều ý nghĩa trong câu. Các từ này lại chiếm số lượng không nhỏ trong văn bản. Điều này cũng làm cho độ đo tương tự giảm một phần chính xác.

Mặc dù có một vài khuyết điểm trên, thuật toán TextRank vẫn cho kết quả tóm tắt ở mức độ tốt về mặt hình thức và nội dung. Người đọc hoàn toàn hiểu được nội dung của văn bản gốc trình bày vấn đề gì khi đọc văn bản tóm tắt tự động.

Chương 3 Thực nghiệm và đánh giá kết quả

3.1.Dữ liệu thực nghiệm

Dữ liệu thực nghiệm tác giả sử dụng trong luận văn được lấy từ tập dữ liệu trong đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt", do TS. Lê Thanh Hương làm chủ nhiệm [Huo2014]. Tập dữ liệu bao gồm 205 văn bản được chia thành 06 chủ đề. Danh sách chi tiết như sau:

STT	**Chủ đề**	**Số văn bản**
1	Chính trị	31
2	Khoa học công nghệ	28
3	Khoa học – giáo dục	22
4	Kinh tế	53
5	Văn hóa	34
6	Xã hội	35

Bảng 3 Danh sách chủ đề và số lượng văn bản tương ứng

Độ dài văn bản tóm tắt được giới hạn là 03 câu. Độ dài này gần tương đương với độ dài văn bản do người tóm tắt. Dữ liệu được đánh giá bằng phương pháp ROUGE với các tham số:

- Đánh giá toàn bộ văn bản trong mỗi bộ dữ liệu
- Sử đánh giá dựa vào n-gram độ dài 1 từ
- Khoảng tin cậy 95%
- Không sử dụng từ gốc
- Bao gồm cả từ dừng trong đánh giá

- Kết quả đánh giá cuối cùng là kết quả trung bình của toàn bộ tập dữ liệu
- Điểm đánh giá được tính theo công thức (1) và các kết quả Precision, F-score được tính toán từ điểm đánh giá đó.

3.2. Thực nghiệm và đánh giá với độ đo Cosine

Các dữ liệu thử nghiệm được thực hiện với độ đo Cosine theo mô hình đã trình bày ở phần trên.

STT	Tên tập dữ liệu	Recall	Precision	F-score
1	Chính trị	0.87378	0.64363	0.72965
2	Khoa học công nghệ	0.85973	0.63783	0.71215
3	Khoa học - Giáo dục	0.73545	0.71963	0.70619
4	Kinh tế	0.64762	0.76895	0.68793
5	Văn hóa	0.75158	0.65879	0.6722
6	Xã hội	0.72597	0.72784	0.70254

Bảng 4 Kết quả đánh giá hệ thống tóm tắt tự động sử dụng độ đo Cosine

Nhận xét:

Độ đo tương đồng Cô-sin cho kết quả khả quan, các điểm đánh giá trên toàn bộ các tập dữ liệu đều trên 0,5. Tuy nhiên có một vài tập dữ liệu có kết quả thấp so với các tập còn lại như "Kinh tế", "Xã hội" và "Khoa học – Giáo dục".

Kết quả đánh giá này sẽ là một thách thức đối với thuật toán TextRank. Nó đã đạt mức trên trung bình, đối với một thuật toán được cài đặt và tính toán đơn giản. Theo như các con số thể hiện thời gian tóm tắt của từng văn bản theo độ đo Cô-sin theo bảng 5 thì tốc độ nhanh. Thời gian tóm tắt trung bình không quá 2 giây cho mỗi văn bản. Với tốc độ như này, thì phương pháp đánh giá này có thể áp dụng được trong nhiều miền ứng dụng.

STT	Tên tập dữ liệu	TG tóm tắt 1 VB	Tổng TG tóm tắt	Thời gian đánh giá
1	Chính trị	0.5982	18.5440	3.2985
2	Khoa học công nghệ	0.5048	14.1359	3.2988
3	Khoa học - Giáo dục	1.0935	24.0573	2.3830
4	Kinh tế	1.0023	53.1222	2.9300
5	Văn hóa	1.7855	60.7068	2.3212
6	Xã hội	0.8456	29.5963	2.2829

Bảng 5 Thời gian tóm tắt và đánh giá các bộ dữ liệu dùng Cosine

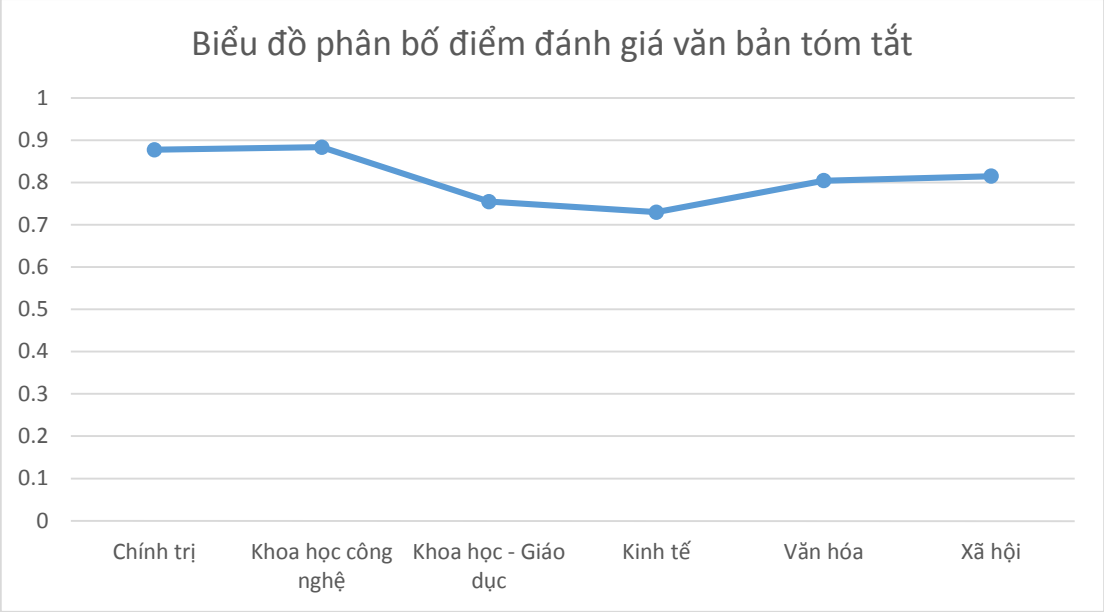
3.3.Thực nghiệm và đánh giá với độ đo TextRank

Kết quả đánh giá với các chủ đề:

STT	Tên tập dữ liệu	Recall	Precision	F-score
1	Chính trị	0.87758	0.61987	0.71288
2	Khoa học công nghệ	0.88368	0.62573	0.71409
3	Khoa học - Giáo dục	0.75484	0.72373	0.72842
4	Kinh tế	0.72985	0.74748	0.72482
5	Văn hóa	0.80426	0.66301	0.69521
6	Xã hội	0.81503	0.74516	0.76178

Bảng 6 Kết quả đánh giá hệ thống tóm tắt tự động sử dụng TextRank

Từ dữ liệu tại bảng 6, ta có biểu đồ như hình 8. Biểu đồ thể hiện điểm đánh giá (Recall) của 6 tập dữ liệu.



Hình 8 Biểu đồ phân bố điểm đánh giá văn bản tóm tắt 6 tập mẫu

Nhận xét:

Độ đo tương đồng TextRank cho kết quả rất tốt, các điểm đánh giá trên toàn bộ các tập dữ liệu đều trên 0,7. Tập dữ liệu cho kết quả tốt nhất là "Khoa học công nghệ" với điểm số đạt 0.88368. Tuy nhiên có một vài tập dữ liệu có kết quả thấp so với các tập còn lại như "Kinh tế", "Khoa học – Giáo dục" và "Văn hóa". Biểu đồ tại hình 8 cho thấy sự khác biệt rõ giữa điểm đánh giá của các tập dữ liệu. Đó cũng thể hiện mức độ chính xác, chất lượng của phương pháp TextRank đối với các tập dữ liệu với các đặc điểm khác nhau.

So sánh với độ tương đồng Cô-sin thì phương pháp TextRank cho kết quả tố hơn rất nhiều. Mức độ chênh lệch điểm số đánh giá lên đến 12,69%. Đây là mức độ chênh lệch lớn cho thấy chất lượng và hiệu quả của phương pháp TextRank. Ngoài ra, sự khác biệt còn thấy ở điểm đánh giá các bộ có điểm số thấp không giống nhau. Điều này cho thấy phương pháp TextRank có cách đánh giá và lựa chọn các câu khác so vớ Cô-sin.

Có một điểm giống nhau giữa hai phương pháp đó là tập dữ liệu "Kinh tế" có điểm đánh giá thấp nhất. Điều này chứng tỏ tồn tại các văn bản khó áp dụng được phương pháp trích rút văn bản. Đây là điểm gợi ý cho tác giả đưa ra một số

kiến nghị nhằm nâng cao kết quả tóm tắt văn bản bằng cách kết hợp nhiều phương pháp khác nhau.

STT	Tên tập dữ liệu	TG tóm tắt 1 VB	Tổng TG tóm tắt	Thời gian đánh giá
1	Chính trị	0.3759	11.6545	24.4524
2	Khoa học công nghệ	0.3278	9.1792	31.0360
3	Khoa học - Giáo dục	0.6095	13.4111	19.0962
4	Kinh tế	0.5538	29.3524	43.7601
5	Văn hóa	1.0105	34.3579	20.0343
6	Xã hội	0.5326	18.6443	19.3999

Bảng 7 Thời gian tóm tắt và đánh giá các bộ dữ liệu dùng TextRank

Nhận xét:

Từ bảng 5, bảng 7 và phân tích dữ liệu thực nghiệm, tác giả nhận thấy rằng tốc độ tóm tắt văn bản phụ thuộc vào độ dài văn bản và độ dài câu. Điều này phù hợp với thuật toán TextRank. Thuật toán TextRank tính toán đệ quy trên toàn văn bản, chính vì vậy, khi độ dài văn bản càng lớn thì thời gian càng lâu. Đây là nhược điểm của thuật toán. Từ đặc điểm này mà thuật toán sẽ khó áp dụng trong các miền ứng dụng mà độ dài dữ liệu lớn. Như vậy, phương pháp tóm tắt này phù hợp với các loại hình văn bản dạng tin tức, văn bản nội dung ngắn gọn.

Trong bảng 7, thời gian tóm tắt trung bình của 1 văn bản trên toàn bộ tập dữ liệu 205 văn bản rất thấp, ở mức xấp xỉ 0,6 giây. Đây là một con số ấn tượng. Nó cho thấy tiềm năng áp dụng phương pháp TextRank vào thực tế. Đặc biệt là trong các ứng dụng thời gian thực. Đặc biệt, sau khi cải tiến phương pháp và nâng cao chất lượng văn bản tóm tắt tự động thì phương pháp này sẽ có thể áp

dụng vào việc tóm tắt nội dung tin tức của các báo điện tử Tiếng Việt. Đây cũng là một mong muốn của tác giả khi thực hiện luận văn thạc sĩ.

Trong tổng số 205 văn bản được thử nghiệm và đánh giá, tác giả lựa chọn 194 văn bản có kết quả tốt nhất và phân loại thành 13 bộ dữ liệu có điểm đánh giá ROUGE theo các nhóm. Các bộ dữ liệu này được lựa chọn với các tiêu chí: điểm đánh giá ROUGE cùng khoảng. Các văn bản có độ dài câu trung bình tương tự nhau sẽ nằm trong một bộ. Các văn bản theo đánh giá của tác giả mà khó có khả năng tóm tắt theo trích rút sẽ được loại bỏ.

Chi tiết về 13 bộ dữ liệu được lưu tại địa chỉ: http://summarizer.dongsukien.com/site/data-sample/. Từ đó nhận thấy được sự khác nhau về đặc điểm các văn bản trong từng bộ dữ liệu. Đây chính là cơ sở để đưa ra các đề xuất cải tiến, nâng cao chất lượng văn bản tóm tắt tự động.

Kết quả của 13 bộ dữ liệu được tóm tắt sử dụng TextRank như sau:

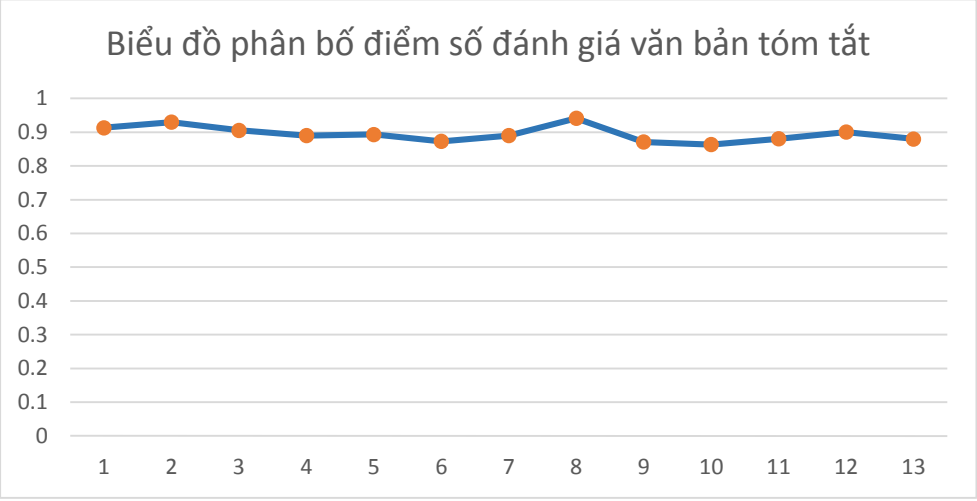
STT	Recall	Precision	F-score	TG tóm tắt	TG đánh giá	Số văn bản
1	0.91326	0.65657	0.7478	9.677	21.205	15
2	0.93004	0.54271	0.66829	11.255	18.44	15
3	0.90592	0.59224	0.6908	8.113	15.315	15
4	0.89013	0.7042	0.77305	9.86	21.624	15
5	0.89323	0.62655	0.71146	12.217	20.589	15
6	0.87302	0.65482	0.7338	8.209	18.159	15
7	0.89008	0.64024	0.72404	8.129	23.099	15
8	0.94137	0.62938	0.74311	7.097	15.501	15
9	0.87133	0.67401	0.73531	11.427	26.872	15

10	0.86354	0.73462	0.78217	7.842	21.716	15
11	0.8806	0.58109	0.66558	8.898	17.557	15
12	0.90053	0.66994	0.74627	9.401	17.872	15
13	0.88016	0.69561	0.74795	7.753	21.575	14

Bảng 8 Kết quả đánh giá 13 bộ dữ liệu sau khi đã phân tích

Nhận xét:

Sau khi đã loại bỏ các văn bản khó trích rút thì điểm đánh giá trên toàn bộ tập dữ liệu đã tăng lên đáng kể. Điểm đánh giá cao nhất thuộc về tập số 8, đạt 0.94137. Từ dữ liệu tại bảng 8, ta có biểu đồ 9. Với biểu đồ 8, khi chú ý vào điểm số của từng tập thì các kết quả không có sự khác biệt nhiều, đường biểu diễn không quá nhấp nhô. Đây là điểm chứng tỏ rằng phương pháp TextRank sẽ cho kết quả tốt nhất ở những văn bản có khả năng trích rút và cùng chung tập đặc trưng: độ dài văn bản ngắn, độ dài câu ngắn, chứa ít các từ nối, từ quan hệ.



Hình 9 Biểu đồ phân bố điểm đánh giá văn bản tóm tắt của 13 tập dữ liệu

Ngoài việc thử nghiệm và đánh giá dữ liệu trên tập dữ liệu mẫu, tác giả đã xây dựng hệ thống tóm tắt văn bản tự động dựa trên TextRank. Đây là kết quả của quá trình nghiên cứu và thực hiện luận văn này. Hệ thống được viết dựa trên nền web có sử dụng ngôn ngữ PHP làm giao diện hiển thị, Python 3 làm hệ

thống xử lý dữ liệu và tóm tắt văn bản. Hệ thống đang được chạy tại địa chỉ: http://summarizer.dongsukien.com/.

3.4. Khuyến nghị tăng cường độ chất lượng văn bản tóm tắt

3.4.1. Khuyến nghị tăng cường độ liên quan giữa các câu

Hiện tại, trong luận văn, tác giả có sử dụng 2 phương pháp để tính độ tương đồng giữa các câu trong văn bản: TextRank và Cosine. Tuy nhiên cả 2 phương pháp trên đều không dựa vào ngữ nghĩa hay văn phạm, ngữ pháp của văn bản. Đặc biệt là chưa sử dụng các đặc trưng của Tiếng Việt. Điều này làm cho thông tin đưa vào tính độ tương đồng không đầy đủ. Vì vậy tác giả khuyến nghị sử dụng các phương pháp có sử dụng các đặc trưng ngôn ngữ để tăng cường độ liên quan giữa các câu trong văn bản. Một số phương pháp được đề xuất: sử dụng kho dữ liệu Wordnet Tiếng Việt, sử dụng mạng ngữ nghĩa Wikipedia.

- Sử dụng khi dữ liệu Wordnet Tiếng Việt. Theo tác giả được biết đề tài xây dựng Wordnet Tiếng Việt đang được hoàn tất ở giai đoạn cuối cùng. Đây là điều mà được mong chờ từ rất lâu. Khi có được kho dữ liệu Wordnet này, việc xác định quan hệ ngữ nghĩa giữa các từ trong câu và các câu trong văn bản sẽ được dễ dàng, không cần phải đi qua các bước trung gian.
- Sử dụng mạng ngữ nghĩa Wikipedia. Đối với sử dụng phương pháp này thì cần phải tính độ tương đồng giữa các khái niệm trong mạng ngữ nghĩa Wikipedia. Độ tương đồng này được nhiều nghiên cứu đưa ra như Ponzetto và cộng sự trong các năm 2006, 2007 [SP06, PSM07], Torsten Zesch và cộng sự năm 2007 [ZG07, ZGM07],… Các nghiên cứu trên tập trung chủ yếu vào việc áp dụng và cải tiến các độ đo phổ biến về tính độ tương đồng từ trên tập dữ liệu Wordnet để có thể tính độ tương đồng giữa các khái niệm trên mạng ngữ nghĩa Wikipedia. Giống như, Wordnet các độ đo này được chia thành hai loại: độ đo dựa vào khoảng cách giữa các khái niệm: Path Length (PL, 1989), Leacock & Chodorow (LC, 1998), Wu and Palmer (WP, 1994) [ZG07, SP06] và độ đo dựa vào nội dung

thông tin: Resnik (Res, năm 1995), Jiang and Conrath (JC, 1997), Lin (Lin, 1998) [ZG07]. Khi áp dụng các độ đo tương đồng ngữ nghĩa trên sang đo độ tương đồng giữa các câu trong văn bản sử dụng cosine sẽ gặp một vấn đề, đó là giá trị các độ đo trên không bị ràng buộc giá trị trong khoảng giá trị [0,1]. Để khắc phục vấn đề này, Li và cộng sự năm 2006 [LLB06] đã đưa ra công thức cải tiến mà không làm ảnh hưởng đến các kết quả trước đó. Đối với việc áp dụng mạng ngữ nghĩa Wikipedia trong tóm tắt văn bản Tiếng Việt, hiện nay Wikipedia đã có khoảng 230.000 chủ đề Tiếng Việt và khoảng 1.000.000 bài viết Tiếng Việt (tính đến tháng 11/2015). Đây là lượng dữ liệu phong phú để áp dụng phương pháp này cho kết quả tốt.

3.4.2. Khuyến nghị tăng cường chất lượng văn bản tóm tắt

Sau khi tăng cường được độ tương đồng giữa các câu trong văn bản thì chất lượng văn bản tóm tắt sẽ được nâng cao. Nguyên nhân là do các câu quan trọng trong câu được xếp hạng cao hơn. Tuy nhiên, do phương pháp TextRank là trích rút câu nên khi ghép các câu lại với nhau sẽ không được tự nhiên về mặt ngôn ngữ. Vì vậy, cần phải loại bỏ được yếu tố này để đảm bảo được văn bản tóm tắt tự động giống người hơn. Tác giả đề xuất kết hợp phương pháp tóm tắt trích rút câu sử dụng TextRank với phương pháp tóm lược câu sử dụng cấu trúc cú pháp. Phương pháp tóm lược câu này sẽ giúp rút gọn câu, đưa câu trở về dạng đơn giản, ngắn gọn. Đồng thời nó cũng tạo cho các câu trong văn bản tóm tắt sử dụng trích rút câu không còn cảm giác gượng gạo do có sự xuất hiện của các từ thừa do cách sử dụng các cấu trúc cú pháp phức tạp.

Tổng kết

Những vấn đề đã giải quyết được trong luận văn

Luận văn đã nghiên cứu giải quyết vấn đề tóm tắt văn bản tiếng Việt sử dụng phương pháp TextRank. Bài toán này có tính ứng dụng thực tế cao và không cần những kiến thức chuyên sâu về ngôn ngữ học. Phương pháp trong luận văn sử dụng chủ yếu dựa vào các thông tin trong chính văn bản được tóm tắt. Dựa vào việc tìm ra các mối quan hệ của thông tin chứa trong các câu trong văn bản mà thuật toán tính toán được mức độ quan trong của từng câu trong văn bản. Từ đó, hệ thống sẽ đưa ra được một văn bản tóm tắt tự động có mức độ chính xác cao, chứa thông tin đầy đủ so với bản gốc. Bên cạnh đó, luận văn cũng có trình bày hoàn chỉnh mô hình của một hệ thống tóm tắt văn bản tự động. Kết quả trực tiếp là một hệ thống tóm tắt văn bản tự động đã được đưa vào hoạt động trên thực tế. Đây là một nỗ lực của tác giả trong việc đưa những nghiên cứu trong quá trình làm luận văn ứng dụng vào thực tế. Việc này góp phần nâng cao được chất lượng của các nghiên cứu sau này khi có sự phản hồi từ phía người sử dụng.

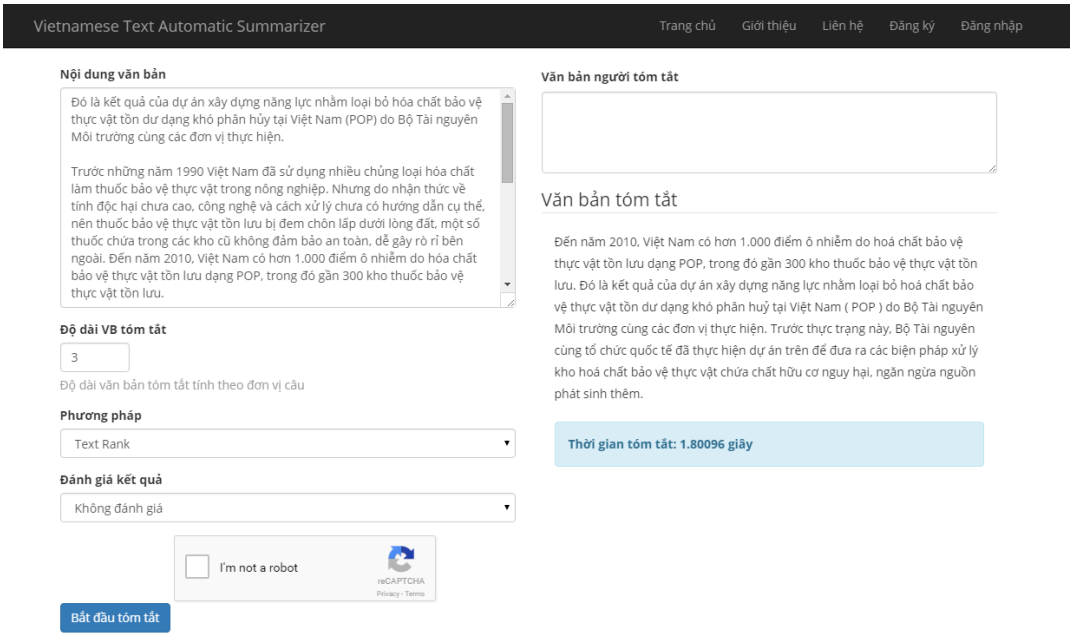
Công việc tương lai cần làm

- Phát triển hoàn thiện hơn hệ thống tóm tắt văn bản tự động đang triển khai.
- Nghiên cứu áp dụng các phương pháp khác vào việc nâng cao độ tương đồng giữa các câu, từ đó tìm ra được các câu quan trọng trong văn bản.
- Nghiên cứu áp dụng các phương pháp giúp nâng cao chất lược văn bản tóm tắt bằng việc rút gọn các câu trong văn bản tóm tắt. Giúp cho văn bản tóm tắt giống người hơn.
- Triển khai xây dựng hệ thống tóm tắt đa văn bản sử dụng phương pháp TextRank kết hợp với các phương pháp nâng cao chất lượng. Từ đó cung cấp một sản phẩm tự động tổng hợp tin tức theo chủ đề từ các bài báo trên Internet thành một bài viết ngắn gọn, súc tích, đầy đủ thông tin. Công cụ

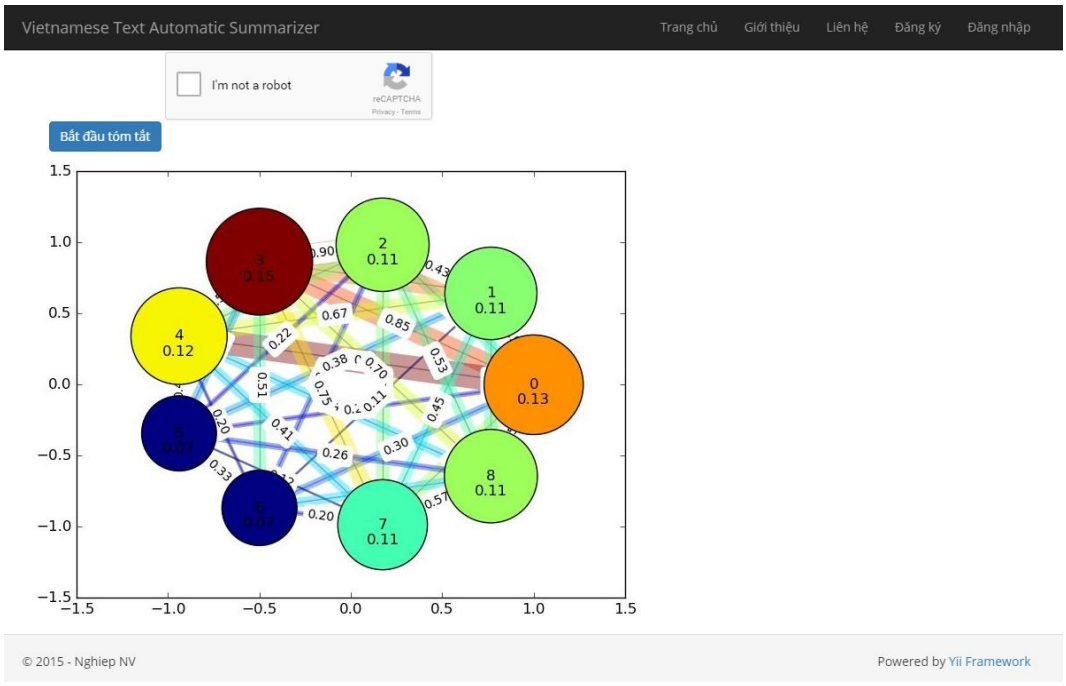
này giúp cho người đọc không cần phải vất vả tìm kiếm thông tin hữu ích trong môi trường Internet ngập lụt thông tin.

Sản phẩm phần mềm

Hệ thống tóm tắt văn bản tự động: http://summarizer.dongsukien.com/



Hình 10 Giao diện chương trình tóm tắt văn bản tự động



Hình 11 Giao diện hiển thị đồ thị quan hệ giữa các câu trong văn bản

Phụ lục

Phụ lục 1: Danh sách các văn bản được sử dụng

Chính trị	**Khoa học - CN**	KHGD3	KT12	KT44	VH22	XH19
CT01	KHCN1	KHGD4	KT13	KT45	VH23	XH20
CT02	KHCN2	KHGD5	KT14	KT46	VH24	XH21
CT03	KHCN3	KHGD6	KT15	KT47	VH25	XH22
CT04	KHCN4	KHGD7	KT16	KT48	VH26	XH23
CT05	KHCN5	KHGD8	KT17	KT49	VH27	XH24
CT06	KHCN6	KHGD9	KT18	KT50	VH28	XH25
CT07	KHCN7	KHGD10	KT19	KT51	VH29	XH26
CT08	KHCN8	KHGD11	KT20	KT52	VH30	XH27
CT09	KHCN9	KHGD12	KT21	KT53	VH31	XH28
CT10	KHCN10	KHGD13	KT22	**Văn hóa**	VH32	XH29
CT11	KHCN11	KHGD14	KT23	VH01	VH33	XH30
CT12	KHCN12	KHGD15	KT24	VH02	VH34	XH31
CT13	KHCN13	KHGD16	KT25	VH03	**Xã hội**	XH32
CT14	KHCN14	KHGD17	KT26	VH04	XH01	XH33
CT15	KHCN15	KHGD18	KT27	VH05	XH02	XH34
CT16	KHCN16	KHGD19	KT28	VH06	XH03	XH35
CT17	KHCN17	KHGD20	KT29	VH07	XH04	

CT18	KHCN18	KHGD21	KT30	VH08	XH05	
CT19	KHCN19	KHGD22	KT31	VH09	XH06	
CT20	KHCN20	**Kinh tế**	KT32	VH10	XH07	
CT21	KHCN21	KT1	KT33	VH11	XH08	
CT22	KHCN22	KT2	KT34	VH12	XH09	
CT23	KHCN23	KT3	KT35	VH13	XH10	
CT24	KHCN24	KT4	KT36	VH14	XH11	
CT25	KHCN25	KT5	KT37	VH15	XH12	
CT26	KHCN26	KT6	KT38	VH16	XH13	
CT27	KHCN27	KT7	KT39	VH17	XH14	
CT28	KHCN28	KT8	KT40	VH18	XH15	
CT29	**Khoa học – GD**	KT9	KT41	VH19	XH16	
CT30	KHGD1	KT10	KT42	VH20	XH17	
CT31	KHGD2	KT11	KT43	VH21	XH18	

Phụ lục 2: Danh sách thư viên sử dụng trong hệ thống tóm tắt văn bản Tiếng Việt tự động

STT	Tên thư viện	Ngôn ngữ
1	Yii Framework version 2	PHP
2	networkx	Python
3	matplotlib	Python
4	nltk	Python
5	collections	Python
6	symbol	Python
7	codecs	Python
8	math	Python
9	os	Python
10	re	Python
11	vietsegment	Python

Tài liệu tham khảo

Tiếng Việt

1. Diệp Quang Ban (chủ biên) , Hoàng Văn Thung (1996), Ngữ pháp tiếng Việt T1 - T2, NXB Giáo dục, Hà Nội.

2. Lê Biên (1993), Từ loại tiếng Việt hiện đại, ĐH Sư phạm I Hà Nội.

3. Nguyễn Tài Cẩn (1996), Ngữ pháp tiếng Việt,NXB ĐH Quốc gia HN.

4. Mai Ngọc Chừ, Vũ Đức Nghiệu, Hoàng Trọng Phiến (1997), Cơ sở ngôn ngữ học và tiếng Việt, NXB Giáo dục.

5. Đinh Văn Đức (1986), Ngữ pháp tiếng Việt: Từ loại, NXB Đại học và trung học chuyên nghiệp.

6. Nguyễn Thiện Giáp (chủ biên), Đoàn Thiện Thuật, Nguyễn Minh Thuyết (1996), Dẫn luận ngôn ngữ học, NXB Giáo dục.

7. Lê Thanh Hương, Hà Quang Thụy, Trần Mai Vũ, Vũ Đức Thi, Nguyễn Thị Thu Trang, Hoàng Anh Việt và Đỗ Bá Lâm (2014), Báo cáo tổng kết đề tài B2012 - 01 – 24 “Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt”, Trường Đại học Bách Khoa Hà Nội, 2014.

8. Vương Hữu Lễ, Hoàng Dũng (1994), Ngữ âm tiếng Việt, NXB Giáo dục.

9. Lương Chi Mai và Hồ Tú Bảo (2009). Báo cáo Tổng kết đề tài KC.01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt" và Về xử lý tiếng Việt trong công nghệ thông tin (2006), Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ Việt Nam, 2009.

10. Tạ Văn Thông (2003), "Hình dung các bộ phận cơ thể người qua "loại từ" tiếng Việt", Tạp chí Ngôn ngữ và đời sống số 9 (95).

11. Trần Mai Vũ (2009), Tóm tắt đa văn bản dựa vào trích xuất câu, Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2009

Tiếng Anh

[Ba07] Barry Schiffman (2007). Summarization for Q&A at Columbia University for DUC 2007, In Document Understanding Conference 2007 (DUC07), Rochester, NY, April 26-27, 2007.

[BE97] Regina Barzilay and Michael Elhadad. Using Lexical Chains for Text Summarization, In Advances in Automatic Text Summarization (Inderjeet Mani and Mark T. Maybury, editors): 111–121, The MIT Press, 1999.

[BKO07] Blake,C., Kampov, J., Orphanides, A., West,D., & Lown, C. (2007). UNCCH at DUC 2007: Query Expansion, Lexical Simplification, and Sentence Selection Strategies for Multi-Document Summarization, In DUC07.

[BL06] Blei, M. and Lafferty, J. (2006). Dynamic Topic Models, In the 23th International Conference on Machine Learning, Pittsburgh, PA.

[BME02] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown (2002). Inferring strategies for sentence ordering in multidocument news summarization, Journal of Artificial Intelligence Research: 35–55, 2002.

[BME99] Barzilay R., McKeown K., and Elhadad M. Information fusion in the context of multidocument summarization, Proceedings of the 37th annual meeting of the Association for Computational Linguistics: 550–557, New Brunswick, New Jersey, 1999.

[BMI06] D. Bollegara, Y. Matsuo, and M. Ishizuka (2006). Extracting key phrases to disambiguate personal names on the web, In CICLing 2006.

[BP98] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, In To Appear: Proceedings of the Seventh International Web Conference (WWW 98), 1998.

[CG98] Jaime Carbonell, Jade Goldstein (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In SIGIR-98, Melbourne, Australia, Aug. 1998.

[CSO01] John M Conroy, Judith D Schlesinger, Dianne P O'Leary, Mary Ellen Okurowski (2001). Using HMM and Logis-tic Regression to Generate Extract Summaries for DUC, In DUC 01, Nat'l Inst. of Standards and Technology, 2001.

[Da12] Davide Buscaldi, Ronan Tournier, Nathalie Aussenac-Gilles and Josiane Mothe, IRIT: Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method, France, 2012.

[Ed69] H. Edmundson (1969). New methods in automatic abstracting, Journal of ACM, 16 (2):264-285, 1969.

[FMN07] K. Filippova, M. Mieskes, V. Nastase, S. Paolo Ponzetto, M. Strube (2007). Cascaded Filtering for Topic-Driven Multi-Document Summarization, In EML Research gGmbH, 2007.

[Ji98] H. Jing (1998). Summary generation through intelligent cutting and pasting of the input document, Technical Report, Columbia University, 1998.

[KST02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL): 311–318, 2002.

[LH03] Chin-Yew Lin and Eduard Hovy (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics, In Human Technology Coference 2003.

[LH97] Chin-Yew Lin and Eduard Hovy (1997). Identifying topics by position, Fifth Conference on Applied Natural Language Processing: 283–290, 1997.

[LLB06] Yuhua Li, David McLean, Zuhair Bandar, James O'Shea, Keeley A. Crockett (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics, IEEE Trans. Knowl. Data Eng. 18(8): 1138-1150.

[Lu58] H. Luhn (1958). The automatic creation of literature abstracts, IBM Journal of Research and Development, 2(2):159-165, 1958.

[Ma01] Inderjeet Mani (2001). Automatic Summarization, John Benjamins Publishing Co., 2001.

[MM99] Inderjeet Mani and Mark T. Maybury (eds) (1999). Advances in Automatic Text Summarization, MIT Press, 1999, ISBN 0-262-13359-8.

[MR95] Kathleen R. McKeown and Dragomir R. Radev (1995). Generating summaries of multiple news articles, ACM Conference on Research and Development in Information Retrieval (SIGIR'95): 74–82, Seattle, Washington, July 1995.

[MT04] Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Texts, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, July 2004.

[PKC95] Jan O. Pendersen, Kupiec Julian and Francine Chen (1995). *A trainable document summarizer*, Research and Development in Information Retrieval: 68 - 73, 1995.

[SD08] P. Senellart and V. D. Blondel (2008). Automatic discovery of similar words. Survey of Text Mining II: Clustering, Classification and Retrieval (M. W. Berry and M. Castellanos, editors): 25–44, Springer-Verlag, January 2008.

[VSB06] Lucy Vanderwende, Hisami Suzuki, Chris Brockett (2006). Task-Focused Summarization with Sentence Simplification and Lexical Expansion, Microsoft Research at DUC2006, 2006.

[WC07] R. Wang and W. Cohen (2007). Language-independent set expansion of named entities using the web, In ICDM07, 2007.

[YuYa13] Yuntong Liu, Yanjun Liang, A sentence semantic similarity calculating method based on segmented semantic comparison, Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, 2013.

[ZG07] T. Zesch and I. Gurevych (2007). Analysis of the Wikipedia Category Graph for NLP Applications, In Proc. of the TextGraphs-2 Workshop, NAACL-HLT, 2007.