

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

————— * —————

ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC
NGÀNH CÔNG NGHỆ THÔNG TIN

TÓM TẮT VĂN BẢN HƯỚNG TRUY VẤN

Sinh viên thực hiện : **Hoàng Đức Thọ**

Lớp : **HTTT-K53**

Giáo viên hướng dẫn : **PGS.TS Lê Thanh Hương**

HÀ NỘI, 5-2013

PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

1. Thông tin về sinh viên

Họ và tên sinh viên: Hoàng Đức Thọ

Điện thoại liên lạc: 0988238277

Email: hoangtho2010@gmail.com

Lớp: Hệ thống thông tin K53

Hệ đào tạo: Kỹ sư

Đồ án tốt nghiệp được thực hiện tại: Trường ĐHBK Hà Nội

Thời gian làm ĐATN: Từ ngày 01/01/2013 đến 01/05/2013

2. Mục đích nội dung của ĐATN

Đề xuất và thử nghiệm phương pháp tóm tắt văn bản hướng truy vấn cho tiếng Việt, áp dụng cho đơn văn bản.

3. Các nhiệm vụ cụ thể của ĐATN

- Tìm hiểu về tóm tắt văn bản tự động
- Đề xuất và phân tích các bước thực hiện một mô hình tóm tắt văn bản hướng truy vấn cho tiếng Việt
- Cài đặt chương trình thử nghiệm và đánh giá kết quả

4. Lời cam đoan của sinh viên:

Tôi – Hoàng Đức Thọ - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của PGS.TS Lê Thanh Hương. Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

Hà Nội, ngày 19 tháng 5 năm 2013

Tác giả ĐATN

Hoàng Đức Thọ

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:

Hà Nội, ngày tháng năm

Giáo viên hướng dẫn

LỜI CẢM ƠN

Trong thời gian học tập tại trường Đại học Bách khoa Hà Nội, em đã học hỏi được rất nhiều kiến thức bổ ích từ các thầy cô giáo, đặc biệt là các thầy cô trong viện Công nghệ thông tin và Truyền thông. Thầy cô đã trang bị cho em rất nhiều kiến thức quý báu, đó cũng như là hành trang và nền tảng để em vững bước hơn khi vào môi trường làm việc đầy thử thách ngoài xã hội.

Em xin gửi lời cảm ơn chân thành tới các thầy cô trong viện, đặc biệt là cô Lê Thanh Hương, người đã tận tình hướng dẫn em trong thời gian thực hiện đồ án này.

TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP

Ngày nay, việc truy cập thông tin qua mạng internet đã trở nên rất phổ biến. Tuy nhiên với lượng thông tin khổng lồ và tăng lên nhanh chóng mỗi ngày, con người không đủ thời gian và sức lực đọc hết các tài liệu để tìm thông tin cần thiết cho mình. Tóm tắt văn bản hướng truy vấn là giải pháp cho vấn đề đó, đây là một dạng đặc biệt của bài toán tóm tắt văn bản tự động mà văn bản sẽ được tóm tắt theo mong muốn của người sử dụng.

Trong đồ án này em đề xuất và thử nghiệm một phương pháp tóm tắt văn bản hướng truy vấn dành cho các văn bản tiếng Việt dựa trên tần số từ và độ tương đồng câu. Nội dung của đồ án gồm 3 phần chính sau:

- Phần 1. Đặt vấn đề và định hướng giải quyết: Mô tả bài toán, tìm hiểu tóm tắt văn bản, tóm tắt hướng truy vấn, đề xuất hướng giải quyết cho tiếng Việt.
- Phần 2: Giải quyết vấn đề: Phân tích chi tiết các bước thực hiện mô hình, xây dựng các công cụ và kiểm thử trên tập mẫu.
- Phần 3: Kết luận và đề xuất: Đánh giá các phần đã làm được, các tồn tại, đề xuất hướng phát triển.

Kết quả kiểm thử cho thấy mô hình này cho kết quả tương đối chính xác, tốc độ xử lý nhanh, có thể cài đặt sử dụng trong thực tế.

Mục lục

DANH MỤC CÁC BẢNG.....	8
DANH MỤC CÁC HÌNH VẼ.....	9
DANH MỤC TỪ VIẾT TẮT.....	10
MỞ ĐẦU.....	11
1. Lý do chọn đề tài.....	11
2. Phạm vi nghiên cứu.....	11
3. Tóm tắt báo cáo.....	11
PHẦN 1. ĐẶT VẤN ĐỀ VÀ ĐỊNH HƯỚNG GIẢI QUYẾT.....	13
I. ĐẶT VẤN ĐỀ.....	13
II. TỔNG QUAN VỀ TÓM TẮT VĂN BẢN TỰ ĐỘNG.....	13
2.1. Định nghĩa.....	13
2.2. Các tiêu chí đánh giá.....	13
2.3. Ứng dụng của tóm tắt văn bản.....	14
2.4. Phân loại tóm tắt văn bản.....	14
2.4.1. Theo đầu vào hệ thống.....	14
2.4.2. Theo đầu ra hệ thống.....	14
2.4.3. Theo mục đích tóm tắt.....	14
2.5. Mô hình biểu diễn văn bản.....	15
2.5.1. Mô hình boolean.....	15
2.5.2. Mô hình không gian vector.....	15
2.5.3. Mô hình tập thô dung sai.....	17
2.6. Mô hình tóm tắt văn bản.....	17
2.7. Các phương pháp áp dụng trong các pha.....	18
2.7.1. Pha Phân tích.....	18
2.7.1.1. Phương pháp thống kê.....	18
2.7.1.2. Phương pháp cấu trúc.....	19
2.7.2. Pha Biến đổi.....	20
2.7.2.1. Giảm lược về cấu trúc câu.....	20
2.7.2.2. Giảm lược về mặt ngữ nghĩa.....	20
2.7.3. Pha Hiện thị.....	21
2.7.3.1. Phương pháp hiện thị phân đoạn.....	21
2.7.3.2. Phương pháp hiện thị liên kết.....	21

2.8.	Đánh giá kết quả tóm tắt.....	21
2.8.1.	Sử dụng so khớp n-gram	22
2.8.2.	Sử dụng các độ đo ROUGE	22
2.9.	Một số hệ thống tóm tắt văn bản tiêu biểu	22
III.	BÀI TOÁN TÓM TẮT VĂN BẢN HƯỚNG TRUY VẤN.....	24
3.1.	Định nghĩa	24
3.2.	Ứng dụng của bài toán.....	24
3.3.	Một số hướng tiếp cận phổ biến	24
3.3.1.	Dựa trên đồ thị.....	24
3.3.2.	Dựa trên cấu trúc diễn ngôn	25
3.3.3.	Dựa trên tần số từ và độ tương đồng câu	25
3.4.	Đề xuất hướng giải quyết cho tiếng Việt.....	25
PHẦN 2.	GIẢI QUYẾT VẤN ĐỀ	28
I.	PHÂN TÍCH MÔ HÌNH THỰC HIỆN BÀI TOÁN	28
1.1.	Giai đoạn phân tích.....	29
1.1.1.	Chuẩn hóa.....	29
1.1.2.	Tách câu, tách từ.....	29
1.1.3.	Loại bỏ từ dừng	30
1.1.4.	Xử lý từ đồng nghĩa.....	31
1.1.5.	Mô hình hóa văn bản	32
1.1.6.	Chọn câu phù hợp tạo tóm tắt.....	32
1.2.	Giai đoạn hiển thị	34
II.	CÀI ĐẶT THỬ NGHIỆM	35
2.1.	Chương trình thử nghiệm	35
2.1.1.	Các công cụ đã xây dựng.....	35
2.1.1.1.	Chương trình tóm tắt.....	35
2.1.1.2.	Công cụ tạo tập mẫu	35
2.1.1.3.	Công cụ kiểm thử.....	36
2.2.	Thử nghiệm một văn bản.....	37
2.2.1.	Đầu vào.....	37
2.2.2.	Kết quả tóm tắt	38
2.2.3.	Nhận xét.....	38
2.3.	Thử nghiệm trên tập mẫu.....	38
2.3.1.	Dữ liệu thử nghiệm.....	38

2.3.2.	Độ đo BLEUS	39
2.3.3.	Kết quả thử nghiệm	40
2.3.4.	Nhận xét, đánh giá mô hình.....	41
PHẦN 3. KẾT LUẬN VÀ ĐỀ XUẤT.....		42
1.	Các công việc đã thực hiện được	42
2.	Đề xuất hướng phát triển.....	42
TÀI LIỆU THAM KHẢO.....		43

DANH MỤC CÁC BẢNG

Bảng 1: Ví dụ một số từ dừng	31
Bảng 2: Một số mục từ đồng nghĩa.....	32
Bảng 3: Thông tin tập mẫu sử dụng để đánh giá	39
Bảng 4: Ví dụ về n-gram.....	39
Bảng 5: Kết quả kiểm thử độ đo BLEUS của tập mẫu	40

DANH MỤC CÁC HÌNH VẼ

Hình 1: Mô hình chung của tóm tắt văn bản.....	17
Hình 2: Mô hình tóm tắt văn bản trích rút	18
Hình 3: Mô hình tóm tắt văn bản hướng truy vấn.....	28
Hình 4: Minh họa quá trình chọn câu quan trọng	33
Hình 5: Giao diện chương trình demo	35
Hình 6: Chương trình quản lý tập mẫu	36
Hình 7: Giao diện chương trình kiểm thử	36

DANH MỤC TỪ VIẾT TẮT

Viết tắt	Ý nghĩa
VSM	Vector Space Model
TF.IDF	Term Frequency. Inverse Document Frequency
TF.ISF	Term Frequency. Inverse Sentence Frequency
DUC	Document Understanding Conferences
TAC	Text Analysis Conference

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay, sự phát triển nhanh chóng của công nghệ thông tin cùng các thiết bị sử dụng, việc chia sẻ, truy cập thông tin qua internet đã trở nên rất phổ biến. Mỗi ngày, vô số thông tin về tình hình kinh tế, xã hội, kinh nghiệm sống, học tập, làm việc... được chia sẻ trên các báo mạng, diễn đàn... Tuy nhiên do lượng thông tin rất lớn hơn nữa còn trùng lặp, dư thừa nhiều, nên con người không đủ thời gian và công sức duyệt hết các văn bản để tìm thông tin hữu ích cho mình. Do đó, cần các hệ thống tổng hợp thông tin một cách ngắn gọn, chính xác, liên quan đến vấn đề mà người dùng quan tâm. Giải pháp cho vấn đề này là bài toán ***Tóm tắt văn bản hướng truy vấn***, một dạng của bài toán tóm tắt văn bản tự động.

Bài toán tóm tắt văn bản tự động vô cùng phức tạp nhưng rất hữu dụng, do đó đã có nhiều công ty lớn, các nhà khoa học, nhóm nghiên cứu đầu tư thời gian và tiền bạc để tìm ra các hướng giải quyết hiệu quả nhất. Các hội nghị nổi tiếng liên quan đến tóm tắt văn bản như: DUC(2001-2007), TAC(2008), ALC(2001-2007)... đã đưa ra rất nhiều kết quả phân tích và các giải pháp hữu ích. Một số hệ thống tóm tắt văn bản đã được ứng dụng vào thực tế và cho thấy lợi ích của nó như MEAD, LexRank, AutoSummarize trong Microsoft Office Word... Đối với tóm tắt hướng truy vấn, cũng có rất nhiều công trình nghiên cứu, ứng dụng, chủ yếu là sử dụng trong các máy tìm kiếm hoặc hệ thống hỏi đáp tự động.

Tuy nhiên các công trình đó phần lớn dành cho tiếng Anh, với tiếng Việt thì chưa có nhiều nghiên cứu, vì thế trong đề án này, em xin chọn đề tài “***Tóm tắt văn bản hướng truy vấn***”. Với mục đích tìm hiểu quy trình tóm tắt văn bản và các vấn đề liên quan, tổng hợp một số kỹ thuật sử dụng thường sử dụng trong tóm tắt văn bản, dựa vào đó đề xuất, cài đặt thử nghiệm một hướng tiếp cận phù hợp với bài toán tóm tắt đơn văn bản hướng truy vấn cho tiếng Việt.

2. Phạm vi nghiên cứu

- ✓ Các vấn đề xoay quanh tóm tắt văn bản
- ✓ Một số hướng tiếp cận tóm tắt văn bản hướng truy vấn
- ✓ Thực hiện một phương pháp tóm tắt trích rút, đơn văn bản, hướng truy vấn phù hợp với tiếng Việt

3. Tóm tắt báo cáo

Nội dung của báo cáo bao gồm các phần cụ thể như sau:

Phần 1. Đặt vấn đề và định hướng giải quyết

I. Đặt vấn đề: Nêu vấn đề cần giải quyết trong đề án

II. Tổng quan về tóm tắt văn bản tự động: Trình bày các định nghĩa, phân loại, cách biểu diễn văn bản, quy trình thực hiện bài toán tóm tắt văn bản, các kỹ thuật thường dùng, các tiêu chí và một số phương pháp đánh giá hệ thống tóm tắt.

III. Bài toán tóm tắt văn bản hướng truy vấn: Trình bày một số hướng tiếp cận cho bài toán tóm tắt hướng truy vấn và đề xuất một hướng tiếp cận phù hợp cho văn bản tiếng Việt.

Phần 2. Giải quyết vấn đề

I. Phân tích mô hình thực hiện bài toán: Đưa ra mô hình cụ thể, và phân tích chi tiết các bước thực hiện dựa trên hướng tiếp cận đã đề xuất.

II. Cài đặt thử nghiệm: Xây dựng các công cụ và dữ liệu mẫu để thực hiện kiểm thử, đánh giá mô hình. Từ đó nhận xét ưu nhược điểm và khả năng ứng dụng.

Phần 3. Kết luận và đề xuất: Trình bày các vấn đề đã giải quyết được trong đồ án, các vấn đề tồn tại và đề xuất hướng phát triển.

PHẦN 1. ĐẶT VẤN ĐỀ VÀ ĐỊNH HƯỚNG GIẢI QUYẾT

I. ĐẶT VẤN ĐỀ

Như đã nêu ở trên, mục tiêu cụ thể của đề án là đề xuất và thử nghiệm một hướng tiếp cận cho bài toán tóm tắt hướng truy vấn đơn văn bản áp dụng được cho tiếng Việt. Cụ thể bài toán cần giải quyết được phát biểu như sau:

Đầu vào: Văn bản, truy vấn, độ rút gọn

Đầu ra: Bản tóm tắt của văn bản đầu vào xoay quanh vấn đề nêu trong truy vấn

Để giải quyết được bài toán này, việc trước hết là tìm hiểu cơ sở lý thuyết về tóm tắt văn bản, tóm tắt hướng truy vấn, từ đó xác định hướng giải quyết và thực hiện cài đặt thử nghiệm.

II. TỔNG QUAN VỀ TÓM TẮT VĂN BẢN TỰ ĐỘNG

2.1. Định nghĩa

Tóm tắt văn bản là quá trình làm giảm độ dài, độ phức tạp của văn bản mà vẫn giữ lại được nội dung quan trọng của văn bản đó. Công việc tóm tắt văn bản đã xuất hiện từ rất lâu đời, và nó được làm thủ công, do con người đọc, rút ra các ý chính rồi trình bày lại một cách ngắn gọn, dễ hiểu. Mục đích là giúp người sử dụng có cái nhìn tổng quan về nội dung trình bày trong văn bản, để quyết định sử dụng văn bản đó hợp lý. Tuy nhiên với lượng văn bản nhiều và dài thì việc làm thủ công vô cùng tốn thời gian, công sức.

Ngày nay, thời đại công nghệ thông tin phát triển mạnh, tóm tắt văn bản tự động (gọi tắt là tóm tắt văn bản) được nghiên cứu phát triển nhằm mục đích làm thay con người công việc nặng nhọc đó. Đã có rất nhiều định nghĩa được đưa ra, tuy nhiên có thể sử dụng định nghĩa ngắn gọn sau:

“Tóm tắt văn bản là quá trình rút ra những thông tin quan trọng nhất từ một hay nhiều nguồn văn bản để tạo ra một văn bản gọn hơn phục vụ cho một số nhiệm vụ hay người dùng cụ thể”

2.2. Các tiêu chí đánh giá

➤ **Độ mạch lạc** (Coherence): đánh giá mức độ rõ ràng của văn bản tóm tắt, tính súc tích, khả năng có thể đọc và hiểu được của bài viết...

➤ **Độ hàm chứa thông tin** (Informationness): tỉ lệ thông tin của văn bản gốc trong văn bản tóm tắt.

➤ **Độ liên quan** (Relevance): xác định mức độ phù hợp của văn bản tóm tắt với chủ đề cho trước (chủ đề có thể là một câu truy vấn).

➤ **Độ dễ đọc hiểu** (Reading Comprehence): một người được giao việc đọc văn bản kết quả, sau đó trả lời các câu hỏi, hệ thống sẽ phải cho điểm và từ đó đưa ra phần trăm những câu trả lời đúng.

2.3. Ứng dụng của tóm tắt văn bản

Tóm tắt văn bản có nhiều ứng dụng trong thực tế, một số ứng dụng nổi bật như:

- ✓ Tóm tắt tự động các tin tức trên báo điện tử.
- ✓ Trợ giúp thông minh việc đọc và khai thác thông tin.
- ✓ Tóm lược danh sách tìm kiếm từ các Search Engine.
- ✓ Giảm lược nội dung trình bày cho các thiết bị cầm tay.
- ✓ Sinh tự động chủ đề, tiêu đề, dẫn đường văn bản.
- ✓ Hỗ trợ tóm lược nội dung cuộc họp, website, chương trình phát thanh và truyền hình, sổ tay công việc.

2.4. Phân loại tóm tắt văn bản

Có nhiều cách phân loại tóm tắt, phụ thuộc vào tiêu chí sử dụng để phân loại, sau đây là một số cách phân loại cần quan tâm:

2.4.1. Theo đầu vào hệ thống

Tóm tắt đơn văn bản là từ một văn bản nguồn cho ra bản ngắn gọn của văn bản đó. Ngược lại, **tóm tắt đa văn bản** là từ nhiều văn bản nguồn cũng chỉ cho ra một đoạn tóm tắt, chứ không có nghĩa là thực hiện nhiều việc tóm tắt một văn bản đồng thời cho nhiều văn bản khác nhau. Rõ ràng, tóm tắt đa văn bản thì khó hơn, vì ngoài những công việc của tóm tắt đơn văn bản, tóm tắt đa văn bản còn phải thực hiện các công việc như tiền xử lý trích rút, tích hợp thống nhất khuôn dạng và hiển thị kết quả theo cách riêng. Ngoài ra, tóm tắt đa văn bản còn phải đối mặt với các vấn đề như dư thừa trùng lặp dữ liệu giữa các văn bản nguồn, nội dung các văn bản nguồn phân tán, độ rút gọn yêu cầu cao, thời gian xử lý cần phải nhanh trong khi sự phức tạp trong xử lý lớn.

2.4.2. Theo đầu ra hệ thống

Tóm tắt trích rút là quá trình thu gọn văn bản mà trong kết quả ra chứa các đơn vị ngữ liệu văn bản nguồn. **Tóm tắt tóm lược** là quá trình thu gọn văn bản mà trong kết quả ra có một số các đơn vị ngữ liệu mới được sinh ra từ các đơn vị ngữ liệu văn bản nguồn.

2.4.3. Theo mục đích tóm tắt

Tóm tắt chung là tóm tắt theo quan điểm ban đầu của tác giả văn bản gốc. **Tóm tắt hướng truy vấn** là tóm tắt theo quan điểm mong muốn của người dùng ứng dụng

thông qua các tham số truyền vào câu truy vấn. Tóm tắt hướng truy vấn được cài đặt và áp dụng nhiều hơn nhưng trong lĩnh vực hẹp hơn, đi sâu vào các chuyên ngành cụ thể.

2.5. Mô hình biểu diễn văn bản

Văn bản thông thường là dạng dữ liệu phi cấu trúc, do vậy muốn xử lý chúng trước hết phải biểu diễn thành dạng có cấu trúc. Các cấu trúc này phải có khả năng thao tác bằng các phép toán cơ bản như cộng, nhân, đại số quan hệ... Có ba mô hình thỏa mãn yêu cầu đó thường được sử dụng là:

2.5.1. Mô hình boolean

Trong mô hình boolean, văn bản, vốn là tập hợp của các term (thuật ngữ), được biểu diễn bởi chỉ số từng term và trọng số của chúng. Trọng số của từng term - dùng để đánh giá độ quan trọng của chúng - trong mô hình này chỉ mang hai giá trị 0 và 1, tùy theo sự xuất hiện của term đó trong văn bản.

$$w_i = \begin{cases} 1 & t_i \in D \\ 0 & t_i \notin D \end{cases}$$

Trong đó w_i là trọng số của term t_i trong văn bản D .

Đối với vấn đề truy vấn, trong mô hình này câu truy vấn bao gồm các văn bản tìm kiếm liên hệ với nhau thông qua các phép đại số quan hệ cơ bản như NOT (phủ định), AND (và) hay OR (hoặc). Câu truy vấn có thể biểu diễn thành dạng vector với các thành phần liên kết và các phép toán quan hệ cơ bản. Từ đây, độ liên quan giữa một văn bản và truy vấn được xác định thông qua các thành phần liên kết. Độ liên quan này chỉ có thể mang hai giá trị : 0 – văn bản không phù hợp với truy vấn và 1 – văn bản phù hợp.

Do vậy có thể thấy rằng hạn chế lớn nhất của mô hình này đó là việc đánh giá độ liên quan chỉ trả về hai kết quả, hoặc phù hợp hoặc không, như vậy yêu cầu của hệ thống khi cần sắp xếp và chọn lựa các văn bản theo mức độ liên quan đến truy vấn sẽ không đạt. Độ liên quan của mô hình này không thể phân chia thành các mức khác nhau, do vậy không phản ánh được thực tế là việc liên quan giữa văn bản và truy vấn có thể là mờ, không chắc chắn. Hạn chế này được loại bỏ khi ta sử dụng một mô hình tổng quát hơn – Mô hình không gian vector (VSM).

2.5.2. Mô hình không gian vector

Như trên đã đề cập, mô hình không gian vector là mô hình tổng quát hơn mô hình Boolean. Các văn bản được biểu diễn thành các vector nhiều chiều, với trọng số không chỉ mang hai giá trị là 0 hay 1 mà có thể mang các giá trị khác tùy theo cách đánh giá, tính toán. Một khác biệt nữa so với mô hình boolean là các phép toán cơ bản của mô hình không gian vector. Các phép toán đại số quan hệ dĩ nhiên không phù hợp nữa,

thay vào đó là các phép toán vector như cộng hai vector, nhân hai vector, tích vô hướng...

Khi biểu diễn văn bản thành các vector, vấn đề về truy vấn và xác định độ liên quan hoàn toàn được giải quyết. Truy vấn là kết quả của các phép toán vector giữa các vector biểu diễn cho những văn bản cấu thành nên truy vấn, như vậy, truy vấn trong trường hợp này cũng là một văn bản đặc biệt. Việc xác định độ liên quan giữa truy vấn và văn bản được quy thành độ liên quan giữa văn bản và văn bản. Hai văn bản là hai vector, vậy khoảng cách hay góc giữa chúng đều có thể đại diện cho sự liên quan giữa hai văn bản này. Tất nhiên, để áp dụng được các phép toán vector cơ bản, hai vector cần chuẩn hóa về số chiều (độ dài).

Các chỉ số sử dụng trong phương pháp này:

➤ **Tần suất thuật ngữ** của một từ w trong một văn bản d , ký hiệu $TF(w,d)$, có thể sử dụng các công thức sau, với f_{ij} là số lần xuất hiện của từ w_i trong văn bản d_j :

$$\begin{aligned}TF(w_i,d_j) &= f_{ij} \\TF(w_i,d_j) &= 1 + \log(f_{ij}) \\TF(w_i,d_j) &= \text{sqrt}(f_{ij})\end{aligned}$$

➤ **Tần suất văn bản** của một từ w , ký hiệu $DF(w)$ là số lượng văn bản mà từ w có xuất hiện. Nghịch đảo của tần suất văn bản của một từ w , ký hiệu $IDF(w)$ được cho bởi công thức:

$$IDF(w) = \log\left(\frac{m}{h}\right)$$

Trong đó: m là tổng số văn bản, h là số văn bản chứa từ w

➤ **Tần suất TF-IDF** là kết hợp của hai loại tần suất nói trên:

$$TF\text{-}IDF(w,d) = TF(w,d) * IDF(w)$$

Theo mô hình này, mỗi văn bản sẽ được biểu diễn dưới dạng $D(t_1, t_2, \dots, t_n)$ với n là tổng số thuật ngữ xuất hiện, mỗi thuật ngữ sẽ được đánh index, t_i là trọng số của thuật ngữ thứ i (trong danh sách thuật ngữ) trong văn bản D . Khi đó độ liên quan giữa hai văn bản biểu diễn bởi 2 vector $X(x_1, x_2, \dots, x_n)$ và $Y(y_1, y_2, \dots, y_n)$ được tính bằng công thức Cosin:

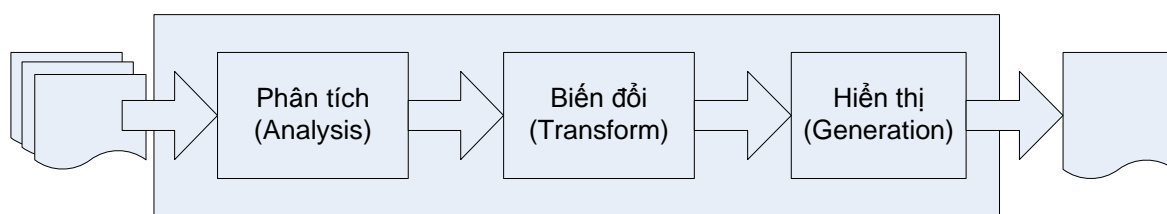
$$\cos(X,Y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum (x_i)^2} \cdot \sqrt{\sum (y_i)^2}}$$

2.5.3. Mô hình tập thô dung sai

Mô hình tập thô dung sai (Tolerance Rough Set Model) là một mô hình mới, tiên tiến dựa trên lý thuyết về logic mờ và tập mờ (Fuzzy Set). Điều cốt lõi của lý thuyết này là việc xác định chính xác một giả thiết nào đó (ví dụ như hai văn bản này có phù hợp, có giống nhau không...) là một điều rất khó. Tuy nhiên chúng ta có thể chỉ ra một cặp xấp xỉ trên và xấp xỉ dưới để khẳng định được giả thiết đó là đúng. Sử dụng các suy diễn hợp lý để xác định và “làm đẹp” các ngưỡng này. Các phép toán cơ bản trong mô hình tập thô dựa trên các quan hệ tương đương các tính chất như đối xứng, phản xạ, bắc cầu... Lý thuyết logic mờ đã và đang được ứng dụng rất mạnh mẽ trong lĩnh vực Trí tuệ nhân tạo.

Mô hình tập thô gần đây được sử dụng nhiều cho các bài toán tìm kiếm cũng như phân nhóm văn bản... Tuy nhiên khi áp dụng mô hình tập thô cho quá trình xử lý văn bản thì tính chất bắc cầu không còn phù hợp. Nhóm tác giả Hồ Tú Bảo, Saori Kawasaki, Nguyễn Ngọc Bình đã đề xuất ra mô hình tập thô dung sai trong đó bỏ đi tính chất bắc cầu trong quá trình xử lý văn bản. Lý thuyết tập thô được các nhà nghiên cứu Trí tuệ nhân tạo phát triển và ngày càng thể hiện được tính ưu việt không chỉ trong việc biểu diễn và thao tác văn bản mà còn trong các vấn đề khác của lĩnh vực này.

2.6. Mô hình tóm tắt văn bản

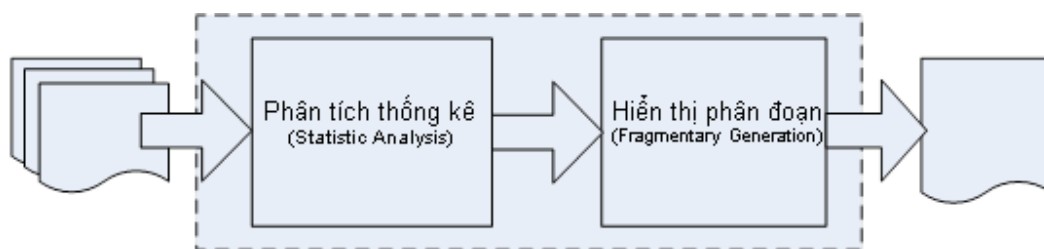


Hình 1: Mô hình chung của tóm tắt văn bản

Một mô hình tóm tắt văn bản tổng quát gồm các pha sau:

- **Phân tích** (Analysis): Phân tích văn bản đầu vào để đưa ra những mô tả bao gồm các thông tin dùng để tìm kiếm, đánh giá các đơn vị ngữ liệu quan trọng cũng như các tham số đầu vào cho việc tóm tắt.
- **Biến đổi** (Transformation): Lựa chọn các thông tin trích chọn được, biến đổi để giản lược và thống nhất, kết quả là các đơn vị ngữ liệu đã được tóm tắt.
- **Hiển thị** (Generation): Từ các đơn vị ngữ liệu đã tóm tắt, liên kết chúng lại thành đoạn theo một thứ tự nào đó hoặc theo cấu kết ngữ pháp rồi hiển thị phù hợp với yêu cầu người dùng.

Một hệ Tóm lược (Abstraction) bao gồm tất cả các pha trên, tuy nhiên một hệ Trích rút (Extraction) chỉ gồm pha Phân tích và Pha Hiển thị, không có pha biến đổi. Thậm chí trong các pha phân tích và hiển thị, chỉ có một số công đoạn được sử dụng.



Hình 2: Mô hình tóm tắt văn bản trích rút

Như vậy một hệ Trích rút tiến hành ít bước hơn, các phương pháp thường dùng là thống kê, học trên ngữ liệu. Còn hệ Tóm lược thì phức tạp, do kết hợp các phương pháp của xử lý ngôn ngữ tự nhiên. Vì vậy, kết quả của các hệ Tóm lược thường thuyết phục hơn (về mặt dễ đọc, dễ hiểu, liên kết ngôn ngữ tốt, gần gũi với con người).

Trong mỗi pha có thể áp dụng nhiều kỹ thuật xử lý khác nhau, chi tiết sẽ được trình bày ở phần tiếp theo.

2.7. Các phương pháp áp dụng trong các pha

2.7.1. Pha Phân tích

Ở pha này văn bản nguồn sẽ được tách thành các đoạn, câu, từ, kết hợp với các thông số đầu vào và áp dụng một số thuật toán cụ thể để chọn ra các đoạn hoặc câu phù hợp làm đầu vào cho pha tiếp theo.

Các phương pháp áp dụng trong pha Phân tích được chia thành hai loại: Phương pháp thống kê và Phương pháp cấu trúc.

2.7.1.1. Phương pháp thống kê

Phương pháp này sử dụng các số liệu thống kê về độ quan trọng của từ, câu hay đoạn, nhận được từ các nghiên cứu về ngôn ngữ học hay thông qua các phương pháp học máy dựa trên tập mẫu để trích rút ra các đơn vị ngữ liệu quan trọng

➤ Phương pháp vị trí

Phương pháp vị trí bao gồm các phương pháp xác định độ quan trọng dựa trên thống kê về vị trí của từ, ngữ hay câu trong văn bản. Các thống kê này tất nhiên phụ thuộc vào thể loại văn bản...

- ✓ **Chủ đề - Tiêu đề** (Title-based): Chủ đề các đoạn văn bản hay tiêu đề các bảng thường chứa các từ và ngữ quan trọng, nên trích rút thông tin từ đây.
- ✓ **Đầu - cuối đoạn** (First - Last Sentence): Xác suất câu đầu đoạn hay câu cuối đoạn chứa ý chính của cả đoạn là rất lớn, đặc biệt là câu đầu đoạn. Ngoài ra, các đoạn đầu và cuối trong văn bản cũng quan trọng hơn các đoạn giữa.
- ✓ **Minh họa - Chú thích** (Comments): Trong các câu chú thích, câu minh họa cho ảnh hay đồ thị thường chứa các thông tin quan trọng. Tuy nhiên, các câu

này thường chỉ được dùng để đánh giá độ quan trọng của các câu khác liên quan, chứ không được chọn làm đầu vào cho pha tiếp.

➤ **Phương pháp ngữ cố định**

Các ngữ cố định có đặc điểm thống kê rất tốt. Sau các ngữ này thường là các câu hay từ có độ quan trọng là xác định. Người ta chia thành hai loại ngữ cố định, một loại mang lại độ quan trọng cho thành phần đi sau, được gọi là ngữ nhấn mạnh, một loại giúp ta loại bỏ, không xét đến những thành phần đi sau vì nó không có nhiều giá trị trong việc trích rút, được gọi là ngữ dư thừa:

- ✓ **Ngữ nhấn mạnh** (Bonus phrase - Emphasizer): Ngữ nhấn mạnh gồm các ngữ như “nói chung là...”, “đặc biệt là...”, “cuối cùng thì...”, “trong bài viết này tôi muốn chỉ ra...”, “bài viết nói về...”, “nội dung gồm...”, ..v.v...
- ✓ **Ngữ dư thừa** (Stigma phrases): Một số ngữ dư thừa: “hiếm khi mà...”, “bài này không nói đến...”, “Không thể nào...”, ..v.v...

➤ **Phương pháp thống kê tần suất từ**

Độ quan trọng của từ phụ thuộc vào số lần xuất hiện của từ đó trong các văn bản liên quan. Các kỹ thuật như TF.IPF hay Tập thuật ngữ thường xuyên (Frequent Item Set) dùng cho công việc xác định tần suất của từ.

2.7.1.2. Phương pháp cấu trúc

Là các phương pháp sử dụng các mối liên hệ cấu trúc - ngữ pháp - ngữ nghĩa để xác định các đơn vị ngữ liệu quan trọng. Tư tưởng chính của các phương pháp này là những đơn vị ngữ liệu nào có chứa các thành phần liên kết nhiều với các thành phần khác sẽ có độ quan trọng lớn. Việc đánh giá các mối quan hệ sẽ dựa trên các mạng ngữ nghĩa, các quan hệ cú pháp hoặc thông qua các phương pháp xác định độ liên quan truyền thống.

➤ **Phương pháp quan hệ lẫn nhau:** Phương pháp này xác định mối quan hệ giữa các đoạn trong văn bản hay các câu trong đoạn với nhau thông qua các kỹ thuật thu thập thông tin ở mức văn bản. Các đoạn (câu) trong văn bản nguồn được tính toán độ liên quan lẫn nhau sử dụng các kỹ thuật như Cosine, TF.IPF hay N-gram Overlap. Sau đó chọn ra đoạn (câu) có độ liên quan lớn nhất.

➤ **Phương pháp liên kết từ vựng:** Phương pháp liên kết từ vựng sử dụng các từ điển quan hệ từ vựng để xây dựng các chuỗi từ liên kết với nhau về mặt ngữ nghĩa. Ví dụ “cây” là một loại “thực vật”, có bộ phận là “lá”, chất liệu là “gỗ”. Các từ “cây”, “thực vật”, “lá”, “gỗ” có quan hệ ngữ nghĩa nào đó với nhau. Sau khi xây dựng được các chuỗi từ này, đánh giá độ mạnh của chúng và có những trích chọn phù hợp.

➤ **Phương pháp Liên kết tham chiếu:** Phương pháp liên kết tham chiếu còn được gọi là phương pháp trích chọn trùng lặp (Anaphora-based Method). Theo

phương pháp này, các cụm trùng lặp được chọn ra, phân rõ xem đâu là từ tham chiếu và từ được tham chiếu. Sau khi phân tách các cụm trùng lặp, chúng ta tạo chuỗi các từ tham chiếu đến cùng một từ được tham chiếu. Chuỗi dài nhất sẽ được coi là trọng tâm của đoạn, các câu chứa các từ trong chuỗi này có một độ ưu tiên nào đó khi xét trích chọn.

➤ **Phương pháp quan hệ câu:** Dựa trên các từ thể hiện mối quan hệ giữa các câu chúng ta cấu trúc hóa đoạn văn bản từ các đơn vị thành phần như ngữ, mệnh đề, câu... Sau đó đơn vị được coi như trung tâm sẽ được trích chọn.

2.7.2. Pha Biến đổi

Ở pha này, các câu sẽ được biến đổi, làm gọn lại hoặc kết hợp nhiều câu tạo thành câu mới ngắn gọn hơn. Các phương pháp trong pha này không làm tăng thêm độ chính xác mà chỉ giúp cho văn bản kết quả ngắn gọn hơn mà vẫn sát nghĩa và thuật toán thường rất phức tạp. Có thể chia làm 2 loại:

2.7.2.1. Giảm lược về cấu trúc câu

Giảm lược về cấu trúc câu là việc lược bỏ trong câu các phần thừa, ít mang giá trị, làm cho cấu trúc câu thu gọn lại. Công việc này thường dựa trên phân tích cú pháp các thành phần trong câu.

2.7.2.2. Giảm lược về mặt ngữ nghĩa

➤ **Phương pháp trừu tượng hóa khái niệm**

Tư tưởng của phương pháp này là từ các khái niệm cụ thể thay thế bằng khái niệm chung.

Ví dụ: “Tôi ăn dâu, táo và đào” => “Tôi ăn trái cây”

➤ **Phương pháp thay thế bộ phận**

Tư tưởng của phương pháp này là từ các khái niệm bộ phận thay thế bằng khái niệm toàn bộ.

Ví dụ: “Xích, lốp, ghi đông, bàn đạp ...” => “Cái xe đạp...”.

➤ **Phương pháp thay thế ngữ tương đương**

Tư tưởng của phương pháp này là các ngữ đóng vai trò như nhau trong câu được thay bằng một ngữ chung.

Ví dụ: “Anh ấy bước vào, ngồi xuống ghế, xem thực đơn, gọi món, ăn, trả tiền và ra về” => “Anh ấy đi ăn tiệm”.

➤ **Phương pháp thay thế từ, ngữ đồng nghĩa ngắn hơn**

Một phương pháp khác khá dễ hiểu đây là việc thay thế một từ, ngữ bằng một từ, ngữ khác đồng nghĩa hoặc gần nghĩa nhưng có độ dài ngắn hơn. Điều này thường thông qua một từ điển các từ đồng nghĩa (Thesaurus).

➤ ***Phương pháp thay thế bởi đại diện***

Tư tưởng của phương pháp này là thay thế một ngữ bằng một ngữ khác có ý nghĩa đại diện cho ngữ ban đầu.

Ví dụ: “Người phát ngôn viên của chính phủ Hoa Kỳ thông báo...” => “Washington thông báo...”.

2.7.3. Pha Hiện thị

2.7.3.1. Phương pháp hiện thị phân đoạn

Đây là phương pháp đơn giản nhất. Các đơn vị ngữ liệu được trích rút hay giản lược từ các pha trước được liên kết lại thành đoạn theo thứ tự tiền định của chúng, không thêm bớt từ nối và cũng không sắp xếp lại các đơn vị ngữ liệu. Văn bản kết quả của phương pháp này có độ dễ đọc dễ hiểu kém, thậm chí lủng củng về nghĩa vì các đơn vị ngữ liệu được trích rút mắc phải một số lỗi như mập mờ tham chiếu, không có từ nối hoặc là thừa từ và ngữ.

2.7.3.2. Phương pháp hiện thị liên kết

Việc hiện thị liên kết là tiếp nhận các đơn vị ngữ liệu đã được trích rút và giản lược từ các pha trước đó, phân tích mối quan hệ về nghĩa của các câu rồi thêm bớt các từ nối, từ dẫn và sắp xếp theo một thứ tự mới dựa vào những gì đã thu thập sao cho thỏa mãn yêu cầu về hiện thị và yêu cầu về độ dễ đọc, dễ hiểu của người dùng.

2.8. Đánh giá kết quả tóm tắt

Đánh giá một bản tóm tắt là một công việc khó bởi không tồn tại một bản tóm tắt lý tưởng cho một (hoặc một tập) văn bản đưa ra. Hơn nữa, việc đánh giá nội dung tóm tắt cũng rất khó khăn. Trường hợp kết quả là một câu trả lời cho một câu hỏi, ta có thể xác định được câu trả lời đó đúng hay sai, nhưng trong các trường hợp khác, thật khó trả lời liệu đầu ra là phải một kết quả đúng hay không? Thực tế luôn có khả năng một hệ thống sinh ra một bản tóm tắt tốt nhưng lại sai khác với bản tóm tắt do người thực hiện. Bên cạnh đó, khi việc đánh giá được thực hiện bởi con người thì chi phí đánh giá sẽ rất cao. Mặt khác, tóm tắt văn bản còn liên quan đến tỉ lệ nén văn bản, do đó, việc đánh giá bản tóm tắt cần phải quan tâm đến vấn đề này, khi đó độ phức tạp và chi phí đánh giá sẽ tăng cao.

Dưới đây là hai phương pháp đánh giá tự động thường sử dụng:

2.8.1. Sử dụng so khớp n-gram

Phương pháp này được Lin và Hovy đưa ra năm 2002 dựa trên mô hình n-gram của độ đo BLEU (Bilingual Evaluation Understudy [1], độ đo đánh giá kết quả dịch máy). Ý tưởng của phương pháp này là so khớp n-gram liên tiếp của bản tóm tắt thủ công và tóm tắt tự động, theo công thức sau:

$$\text{Score} = \alpha_1 * \text{Score}_1 + \alpha_2 * \text{Score}_2 + \alpha_3 * \text{Score}_3 + \alpha_4 * \text{Score}_4$$

Trong đó:

$\text{Score}_i = \text{Số } i\text{-gram trùng nhau} / \text{Tổng số } i\text{-gram của bản tóm tắt thủ công}$

α_i là hệ số đánh giá độ quan trọng của các Score_i

2.8.2. Sử dụng các độ đo ROUGE

ROUGE (Recall-Oriented Understudy of Gisting Evaluation [2]) cũng được đưa ra bởi Lin, vào năm 2009, đây là tập hợp các độ đo dựa trên mô hình n-gram của BLEU với nhiều cách tính khác nhau. Thường sử dụng nhất là độ đo ROUGE-N, với n là giá trị của mô hình n-gram, $n = \{1, 2, 3, 4\}$.

Công thức của độ đo ROUGE-N như sau: Cho $R = (r_1, r_2, \dots, r_n)$ là tập các tóm tắt mẫu, s là tóm tắt tự động, $\Omega_n(d)$ là vector biểu diễn mô hình n-gram của văn bản d.

$$\text{ROUGE} - N(s) = \frac{\sum_{r \in R} \langle \Omega_n(r), \Omega_n(s) \rangle}{\sum_{r \in R} \langle \Omega_n(r), \Omega_n(r) \rangle}$$

Độ đo ROUGE được sử dụng làm độ đo chính thức của các hội nghị DUC 2004-2007 và TAC 2008-2012.

2.9. Một số hệ thống tóm tắt văn bản tiêu biểu

Hiện tại, trên thế giới đã có rất nhiều nghiên cứu và dự án xây dựng các ứng dụng tóm tắt văn bản. Các ứng dụng này có thể đáp ứng rất nhiều các mục đích khác nhau. Có thể kể ra một số ứng dụng Tóm tắt văn bản tiêu biểu như sau:

- **SUMMARIST**: Một hệ thống Trích rút văn bản năm thứ tiếng (tiếng Anh, tiếng Nhật, tiếng Tây Ban Nha, tiếng Ả-rập và tiếng Hàn Quốc). Hiện tại SUMMARIST đang nghiên cứu để cải tiến trở thành một hệ thống Tóm lược văn bản và hỗ trợ nhiều ngôn ngữ hơn như tiếng Pháp và Indonesia.
- **SweSUM**: Ứng dụng Tóm tắt văn bản đa ngôn ngữ của Học viện công nghệ hoàng gia Thụy Điển. SweSUM có thể tóm tắt các văn bản có ngôn ngữ vùng Scandinavi như Thụy Điển, Đan Mạch, Na Uy và các ngôn ngữ khác như tiếng Anh, Pháp, Đức, Tây Ban Nha và cả tiếng Iran.

- **SumUM:** Hệ thống Tóm lược văn bản kỹ thuật của nhóm nghiên cứu xử lý ngôn ngữ tự nhiên trường Đại học Montréal, Canada. SumUM có thể thực hiện cả chức năng tóm tắt chỉ định và tóm tắt thông tin rất tốt..
- **FJCL:** Hệ thống Rút trích văn bản tiếng Nhật được phát triển trong phòng nghiên cứu Ikeda của trường đại học Gifu. Đây là một hệ thống sử dụng các phương pháp áp dụng cho hệ ngôn ngữ đơn âm tiết (monosyllabic language system) như tiếng Nhật, Hàn Quốc, Trung Quốc và Việt Nam.
- **Pertinence Summarizer:** Hệ thống tóm tắt tin tức đa ngôn ngữ trực tuyến nổi tiếng. Hiện tại để thử nghiệm khả năng của mình, Pertinence đã được tích hợp với Google và tóm tắt tự động danh sách tìm kiếm trả về từ Google thông qua câu truy vấn đưa vào. Chúng ta có thể thử nghiệm hệ thống này trên trang web: www.pertinence.net.
- **MEAD:** Nền tảng cho các hệ thống Tóm tắt nhiều văn bản và đa ngôn ngữ. Đây là một bộ công cụ xây dựng trên nền Linux và Solaris, sử dụng ngôn ngữ Perl - Một ngôn ngữ có khả năng xử lý văn bản rất linh hoạt và mạnh mẽ. MEAD biểu diễn, lưu trữ dữ liệu ở dạng XML, cung cấp cho chúng ta khung ứng dụng để cài đặt các ứng dụng Tóm tắt văn bản cho ngôn ngữ mà ta muốn. Ngoài ra MEAD cũng cung cấp các công cụ để xây dựng các ứng dụng đánh giá hệ thống tóm tắt theo các tiêu chí và các tập mẫu nổi tiếng. MEAD được xây dựng bởi các chuyên gia nổi tiếng về Xử lý ngôn ngữ ở khắp nơi trên thế giới dưới sự tài trợ của Chương trình Nghiên cứu Công nghệ thông tin của Tổ chức Khoa học quốc gia Mỹ. MEAD được cung cấp ở dạng mã nguồn mở để nghiên cứu và kế thừa. Hiện tại phiên bản mới nhất của MEAD là MEAD v3.07.
- **Microsoft Word AutoSummary:** Microsoft cũng cài đặt chức năng Trích rút và sinh tiêu đề trong Microsoft Word từ phiên bản Word '97. Chúng ta có thể thử bằng cách chọn Tools - AutoSummarize trên thanh công cụ (có thể khác tùy vào phiên bản). Công cụ này cho phép chúng ta chọn thông số về độ rút gọn, trích rút hay sinh tiêu đề...

Ngoài ra còn các hệ thống Tóm tắt văn bản nổi tiếng khác như ANES hay SUMMONS. Tuy nhiên tại Việt Nam hiện nay chưa có một nghiên cứu và ứng dụng Tóm tắt văn bản chính thức nào.

III. BÀI TOÁN TÓM TẮT VĂN BẢN HƯỚNG TRUY VẤN

3.1. Định nghĩa

Theo định nghĩa ở trên, tóm tắt văn bản hướng truy vấn là một dạng tóm tắt văn bản (khi phân chia theo mục đích tóm tắt), điểm đặc trưng là ở giai đoạn tiền xử lý, việc tính toán sẽ phụ thuộc một phần vào truy vấn người dùng.

3.2. Ứng dụng của bài toán

Tóm tắt hướng truy vấn thường sử dụng trong việc tóm tắt kết quả trả về của máy tìm kiếm thông tin, hoặc trong các hệ thống hỏi đáp tự động.

Hiện nay, đối với máy tìm kiếm, hệ thống sẽ tóm tắt văn bản theo tóm tắt đơn văn bản thông thường, lưu vào cơ sở dữ liệu, và thực hiện tìm kiếm trên bản tóm tắt đó để giảm thời gian tìm kiếm. Sau khi xác định được văn bản phù hợp, văn bản đó sẽ được tóm tắt lại theo truy vấn người dùng để đưa ra hiển thị kèm với kết quả. Đối với hệ thống hỏi đáp tự động, hệ thống sẽ tiến hành phân loại câu hỏi và thực hiện so khớp hoặc tính tương đồng với câu hỏi trong cơ sở dữ liệu để xác định câu trả lời phù hợp nhất, sau đó tóm tắt văn bản chứa câu trả lời, sử dụng câu trả lời như truy vấn, và hiển thị kèm với câu trả lời, có đánh dấu câu trả lời.

Tóm lại, tóm tắt hướng truy vấn thường được tích hợp ở giai đoạn xử lý kết quả của hệ thống tìm kiếm thông tin và hỏi đáp tự động, mục đích là thêm thông tin để kết quả rõ ràng và dễ hiểu hơn với người dùng

3.3. Một số hướng tiếp cận phổ biến

3.3.1. Dựa trên đồ thị

Phương pháp này được đưa ra bởi [3] Jagadeesh và đồng sự, áp dụng cho tóm tắt trích rút đa văn bản. Đồ thị của văn bản sẽ được xây dựng dựa trên việc phân tích các câu trong đó để tìm ra các cụm danh từ (noun phrases), sau đó phân tích các cụm danh từ này để tìm ra mối quan hệ giữa các danh từ sử dụng các hàm heuristic. Đồ thị thu được sẽ bao gồm 2 dạng nút, nút thành phần (là các danh từ trích rút từ văn bản) và nút liên kết, có 2 loại nút liên kết là *isa* (là một) và *related_to* (liên quan với).

Sau khi xây dựng đồ thị cho mỗi câu, chúng sẽ được kết hợp để tạo đồ thị cho toàn văn bản. Một thuật toán tìm kiếm sẽ được sử dụng để tìm các câu quan trọng đưa vào tóm tắt. Có 3 giải thuật có thể áp dụng:

- *Dựa trên tâm các đồ thị*: một đồ thị trung tâm cho tất cả văn bản được xây dựng, tích hợp thêm đồ thị của truy vấn. Sau đó các câu có đồ thị tương đồng với tâm lớn nhất sẽ được chọn
- *Dựa trên đồ thị truy vấn*: các câu có đồ thị tương đồng với đồ thị truy vấn lớn nhất sẽ được chọn

- *Dựa trên việc kết hợp câu đã chọn*: giống bước trên nhưng sau khi chọn được mỗi câu thì kết hợp câu đó vào tâm tạo thành tâm mới

Phương pháp này cho kết quả tương đối chính xác nhưng phụ thuộc chủ yếu vào giải đoạn phân tích cú pháp để tìm các cụm danh từ, do đó cần bộ phân tích cú pháp chính xác.

3.3.2. Dựa trên cấu trúc diễn ngôn

Phương pháp này được trình bày bởi W. Bosma [4], mục đích là tạo ra bản tóm tắt ngắn gọn chứa câu trả lời để đưa ra kết quả trong hệ thống hỏi đáp tự động. Trong đó mỗi văn bản được biểu diễn bởi đồ thị có trọng số dựa trên lý thuyết diễn ngôn, mỗi đỉnh đại diện cho một câu, trọng số trên mỗi cạnh là khoảng cách giữa hai câu. Một thuật toán tìm kiếm đồ thị sẽ được sử dụng để chọn ra các câu có tổng trọng số trên đường đi tới câu trả lời(vai trò như truy vấn) nhỏ nhất.

3.3.3. Dựa trên tần số từ và độ tương đồng câu

Phương pháp này trình bày bởi Siva kumar và đồng sự [5] áp dụng cho tóm tắt trích rút đa văn bản. Trước tiên các văn bản sẽ được biểu diễn trong mô hình không gian vector, mỗi câu được tính khoảng cách với câu truy vấn, sau đó sử dụng thuật toán phân cụm, chia các câu vào các cụm. Mỗi câu được tính điểm số vị trí và điểm số độ quan trọng trong cụm, sau đó từ các cụm có điểm số cao nhất, trích rút ra các câu có điểm số cao nhất tạo thành tóm tắt.

3.4. Đề xuất hướng giải quyết cho tiếng Việt

Qua tìm hiểu về các vấn đề liên quan trong tóm tắt và đặc trưng của tiếng Việt, dễ nhận thấy rằng việc tiếp cận ở mức cú pháp và ngữ nghĩa là khá khó khăn, một phần là vì công cụ và dữ liệu hỗ trợ, tuy đã có một số công cụ gán nhãn từ vựng và phân tích cú pháp cho độ chính xác cao nhưng thường chỉ áp dụng cho lĩnh vực hẹp, và còn ở mức nghiên cứu, chưa được công bố chính thức. Mặt khác, do đặc trưng về ngữ pháp nên các hướng tiếp cận đó thường không chính xác với tiếng Việt.

Do đó em xin đề xuất mô hình trích rút các câu quan trọng cho bài toán tóm tắt hướng truy vấn *dựa trên tần số từ và độ tương đồng câu*, áp dụng cho tóm tắt đơn văn bản. Mô tả sơ lược như sau: Đầu tiên sử dụng câu truy vấn làm tâm tóm tắt, sau đó tìm câu có độ tương đồng với tâm lớn nhất, mỗi câu được chọn sẽ kết hợp với tâm tạo nên tâm mới. Sau khi kết thúc sẽ loại bỏ câu truy vấn khỏi kết quả. Phương pháp này dựa theo ý tưởng ở giải thuật thứ 2 trong hướng tiếp cận dựa trên đồ thị đã nêu ở trên, nhưng các câu ở đây biểu diễn theo mô hình không gian vector và độ tương đồng sử dụng độ đo cosin.

Phạm vi ứng dụng hướng tới của mô hình là tích hợp vào modul trả kết quả của bộ máy tìm kiếm văn bản(search engine), thực hiện tóm tắt văn bản kết quả theo tập

từ khóa đã tìm kiếm(chính là truy vấn người dùng). Do đó có một số ràng buộc với dữ liệu đầu vào.

Vì văn bản đã được máy tìm kiếm lựa chọn nên nội dung của văn bản và truy vấn sẽ liên quan với nhau. Do đó các câu chứa nhiều từ khóa trong truy vấn, hay trong trường hợp này là độ tương đồng lớn, sẽ mang các thông tin quan trọng liên quan đến truy vấn mà người dùng quan tâm. Tuy nhiên trong vấn đề tìm kiếm, phần lớn người dùng thường không nắm rõ được nội dung mình muốn biết nên mới sử dụng tìm kiếm, mà chỉ biết các từ khóa liên quan tới vấn đề đó. Ví dụ như tìm kiếm thông tin về giá vàng, người ta không biết giá vàng tăng hay giảm, có biến động gì gần đây. Hoặc tìm cách sửa một lỗi máy tính thì người dùng sẽ đưa ra các thông tin về lỗi đó, sau khi xem bản tóm tắt của các kết quả từ máy tìm kiếm, sẽ biết được kết quả nào phù hợp để quyết định đọc hay không.

Trong giải thuật chọn câu, các câu được chọn sẽ được thêm vào truy vấn, với mục đích làm thêm từ khóa liên quan đến truy vấn. Nhưng không phải từ nào trong các câu đó cũng đều quan trọng nên các từ xuất hiện trong truy vấn gốc được nhân lên một trọng số α . Do đó kết quả tóm tắt sẽ ưu tiên các từ khóa trong truy vấn, và các từ khóa xuất hiện nhiều trong các câu được chọn. Theo đó thì bản tóm tắt sẽ dễ hiểu hơn vì bao gồm các thông tin liên quan tới truy vấn.

Tổng quan về modul đó như sau:

➤ **Đầu vào**

- *Văn bản*: văn bản đầu vào sử dụng bộ mã Unicode utf-8, chỉ chứa text, chính xác về chính tả, dấu câu, không quá ngắn(5 câu trở lên), nội dung phải liên quan tới truy vấn.
- *Truy vấn*: sử dụng bộ mã như văn bản, là một đoạn văn bản chứa các từ khóa cần tìm kiếm, nếu cần chính xác thì dùng dấu phẩy để ngăn cách các từ khóa
- *Độ rút gọn*: có thể là số lượng từ (100-150 từ) hoặc phần trăm văn bản nguồn (10-20%).

➤ **Thực hiện tóm tắt**

Bước này áp dụng mô hình tóm tắt đã đề xuất để tạo kết quả

- *Chuẩn hóa*: bước này sẽ thực hiện xử lý tiêu đề, các đoạn văn trong ngoặc đơn
- *Tách câu, tách từ*: thực hiện tách câu, tách từ sử dụng công cụ VNTokenizer
- *Loại bỏ từ dừng*: tìm kiếm và loại bỏ các từ dừng dựa trên danh sách có sẵn
- *Xử lý từ đồng nghĩa*: đồng bộ các từ đồng nghĩa về cùng 1 dạng
- *Mô hình hóa văn bản*: tính TF.IDF và chuyển các câu về dạng vector

- *Trích rút câu, tạo tóm tắt*: đây là giải thuật đã đề xuất, thực hiện tính toán độ tương đồng sử dụng độ đo cosin và một số phép toán trên vector để tìm kiếm các câu phù hợp đưa vào kết quả tóm tắt, và được ghép lại theo phương pháp hiển thị phân đoạn.

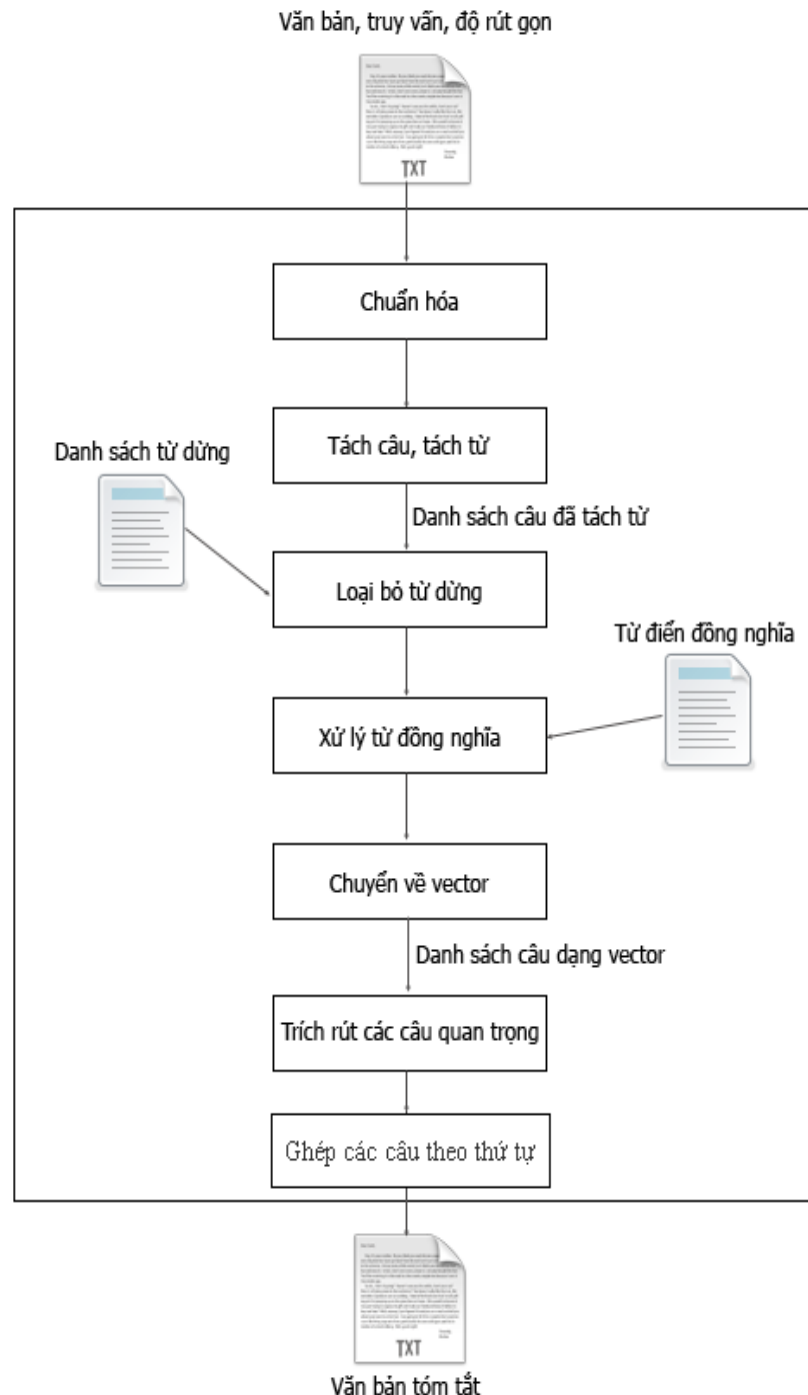
➤ ***Đầu ra***: văn bản tóm tắt

Chi tiết các kỹ thuật sử dụng trong các bước sẽ trình bày ở phần sau.

PHẦN 2. GIẢI QUYẾT VẤN ĐỀ

I. PHÂN TÍCH MÔ HÌNH THỰC HIỆN BÀI TOÁN

Dựa vào các kiến thức về tóm tắt văn bản đã trình bày ở trên, trong phần này em sẽ trình bày chi tiết các kỹ thuật áp dụng trong từng bước của mô hình xử lý đã đề xuất.



Hình 3: Mô hình tóm tắt văn bản hướng truy vấn

1.1. Giai đoạn phân tích

1.1.1. Chuẩn hóa

➤ Xử lý câu tiêu đề

Câu tiêu đề của một văn bản (nếu có) thường mang nội dung chính trình bày trong văn bản, do đó các từ khóa trong đó cũng được dùng để phát hiện tóm tắt (một số giải thuật còn tăng trọng số cho những từ xuất hiện trong tiêu đề), nhưng không đưa câu tiêu đề vào kết quả tóm tắt, nên cần phát hiện để loại bỏ khỏi kết quả. Việc phát hiện câu tiêu đề có thể dựa vào dấu hiệu “câu tiêu đề là câu duy nhất của đoạn đầu tiên”. Trong giải thuật này chỉ sử dụng câu tiêu đề như câu thông thường, sau đó loại khỏi kết quả (nếu nó được chọn vào kết quả).

➤ Xử lý các cụm từ trong ngoặc

Các cụm từ trong ngoặc có thể là chú thích hoặc viết tắt của cụm từ nào đó, nếu là chú thích thì có thể bỏ qua còn từ viết tắt thì khá quan trọng, nhất là đối với tóm tắt hướng truy vấn.

Ví dụ: Sinh viên tình nguyện(SVTN) đi đến các vùng sâu để giúp đỡ đồng bào

Các câu sau câu này sẽ sử dụng cụm từ SVTN, nếu truy vấn có từ khóa “sinh viên tình nguyện” thì các câu sử dụng từ viết tắt sẽ không được quan tâm.

Việc xử lý từ viết tắt không đơn giản là phát hiện các từ trong ngoặc, tùy từng loại văn bản của chuyên ngành nào đó, các từ viết tắt vẫn được sử dụng mà không gây hiểu lầm cho người đọc, vì trong các lĩnh vực ấy nó chỉ có thể thay thế cho cụm từ cố định nào đó, hoặc do thói quen, sử dụng nhiều thì mọi người đều biết.

Ví dụ: UBND thường được dùng thay thế cho “Ủy ban nhân dân”

Trong giải thuật này chỉ xử lý các cụm từ viết tắt chữ đầu trong ngoặc đơn, còn các trường hợp khác do chưa xây dựng được bộ dữ liệu cụ thể nên không xét đến. Các cụm từ trong ngoặc đơn khác sẽ bị xóa đi.

1.1.2. Tách câu, tách từ

Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau, có khoảng 70% các từ gồm 2 âm tiết, và 14% các từ gồm 3 âm tiết, còn lại là 1 âm tiết. Hơn nữa, việc kết hợp các âm tiết có nhiều cách, mỗi cách cho một nghĩa khác nhau. Vì thế, để xử lý tiếng Việt, bài toán tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất. Ngoài tiếng Việt, có khá nhiều các ngôn ngữ châu Á khác cũng cần bước tách từ, ví dụ như: tiếng Nhật, tiếng Trung, tiếng Hàn,... do đó vấn đề này nhận được sự quan tâm rộng rãi và có nhiều hướng tiếp cận khác nhau.

Một số phương pháp có thể áp dụng:

- So khớp từ dài nhất (*Longest Matching*)
- So khớp cực đại (*Maximum Matching*)
- Mô hình Markov ẩn (*Hidden Markov Models- HMM*)
- Học dựa trên sự cải biến (*Transformation-based Learning – TBL*)
- Chuyển đổi trạng thái trọng số hữu hạn (*Weighted Finite State Transducer*)
- Độ hỗn loạn cực đại (*Maximum Entropy – ME*)
- Máy học sử dụng vector hỗ trợ (*Support Vector Machines*)
- Trường xác suất có điều kiện (*CRFs*)

Bài toán tách từ khá phức tạp, do đó việc tách từ trong bước này sẽ sử dụng công cụ VNTokenizer, được phát triển bởi nhóm tác giả Lê Hồng Phương.

Đây là công cụ tách từ tự động cho tiếng Việt, mã nguồn mở, được viết bằng ngôn ngữ Java. Phiên bản cũ nhất là phiên bản vnTokenizer 2.0 được xây dựng vào năm 2005 khi đó nó mới là một ứng dụng đơn với giao diện đơn giản. Để sử dụng trong chương trình lần này, phiên bản mới nhất 4.1.1c, mã nguồn của công cụ được tải tại website của dự án VLSP [6].

Công cụ này được xây dựng sử dụng kết hợp từ điển (từ điển tiếng Việt được lấy từ đề tài VLSP) và ngram, trong đó mô hình ngram được huấn luyện sử dụng treebank tiếng Việt (70,000 câu đã được tách từ), treebank là kho ngữ liệu câu được chú giải ngữ pháp.

Với độ chính xác xấp xỉ 97% (theo thống kê của tác giả trên website) là kết quả rất cao so với công cụ tách từ hiện nay.

Ngoài ra việc tách câu khá đơn giản nhưng cần xử lý các trường hợp nhập nhầm dấu chấm câu và dấu chấm trong từ (trong email, số thập phân, địa chỉ web). Do đó để tiết kiệm thời gian, việc tách câu trong phần này sử dụng luôn modul tách câu trong công cụ VNTokenizer.

1.1.3. Loại bỏ từ dừng

Từ dừng (StopWord) là những từ thường xuất hiện nhiều trong các tài liệu nhưng thường chỉ mang ý nhấn mạnh, bổ nghĩa... nó có ý nghĩa lớn trong một số phương pháp dựa trên dấu hiệu đặc biệt, nhưng trong phương pháp dựa trên tần số từ đang xét thì các từ này làm giảm độ chính xác. Trong giải thuật này chủ yếu dựa trên trọng số từ nên việc loại bỏ từ dừng là rất cần thiết.

Từ dừng sẽ được loại bỏ nhờ một danh sách từ dừng xây dựng sẵn, tham khảo tại [7], sau khi tách từ, các từ xuất hiện trong từ điển từ dừng sẽ bị xóa. Dưới đây là một số từ dừng trích trong file sẽ sử dụng.

thậm chí	vì vậy	tuy nhiên
thật ra	với lại	thế là
trước kia	đáng lẽ	sau cùng
tuy vậy	ắt hẳn	quả thật

Bảng 1: Ví dụ một số từ dùng

Ngoài ra ở bước này, các dấu câu, dấu phẩy cũng bị xóa vì nó cũng giống từ dùng.

1.1.4. Xử lý từ đồng nghĩa

Có 3 loại từ đồng nghĩa cần xét đến:

➤ ***Từ có nghĩa giống nhau hoặc gần giống nhau.***

Ví dụ: siêng năng, chăm chỉ, cần cù, ...

➤ ***Từ đồng nghĩa hoàn toàn***

Ví dụ: hổ, cọp, hùm, ...

➤ ***Từ đồng nghĩa không hoàn toàn***

Ví dụ:

Ăn, xoi, chén, ... (biểu thị thái độ, tình cảm khác nhau đối với người đối thoại hoặc điều được nói đến).

Mang, khiêng, vác, ... (biểu thị những cách thức hành động khác nhau).

Với loại 1 và loại 2 thì các từ đồng nghĩa có thể thay thế cho nhau. Còn loại 3 thì phải xét đến ngữ nghĩa của từ trong ngữ cảnh của văn bản, đây có thể coi là bài toán phức tạp nhất trong xử lý ngôn ngữ, hiện nay chưa có nhiều nghiên cứu.

Việc xử lý từ đồng nghĩa là rất quan trọng, nhất là trong bài toán tóm tắt hướng truy vấn. Trong mô hình lần này, do chỉ xử lý ở mức nông, nên không xét đến các vấn đề ở mức cú pháp và ngữ nghĩa, nhưng để tăng độ chính xác, bài toán sẽ sử dụng việc đồng nhất các từ đồng nghĩa (xử lý chung cho cả 3 loại trên) dựa trên từ điển đồng nghĩa thô xây dựng sẵn, bộ từ điển này gồm gần 2800 mục, xây dựng bằng cách dùng công cụ tải các trang của từ điển Việt – Việt tại trang tratu.soha.vn, sau đó tách thẻ có chứa các từ đồng nghĩa rồi ghép lại. Mỗi mục gồm các từ gần nghĩa hoặc đồng nghĩa với nhau về mặt nào đó, và mỗi từ chỉ xuất hiện trong một mục, trên thực tế có những từ có thể ở nhiều mục, nhưng số lượng các từ đó không nhiều nên trong bộ từ điển này sẽ sử dụng nghĩa phổ biến nhất của các từ đó. Tuy chưa được đầy đủ và xử lý đơn giản nhưng cũng góp phần tăng độ chính xác cho việc tóm tắt. Dưới đây là một số mục từ trong bộ từ đồng nghĩa sẽ sử dụng.

lãnh thổ, bờ cõi, biên thủy, biên giới, biên cương
rối rã, rối, rảnh rỗi, rảnh rang, rảnh
thương nhân, nhà buôn, thương gia, doanh nhân, doanh gia
quả cảm, gan góc, dũng cảm, gan dạ, dũng mãnh, can đảm, anh dũng
tả, mô tả, miêu tả, diễn tả, diễn đạt, biểu đạt

Bảng 2: Một số mục từ đồng nghĩa

Sau bước tách từ và loại bỏ từ dừng, các câu sẽ được xử lý theo theo cách duyệt tất cả các từ, với mỗi từ, tìm từ đó trong từ điển đồng nghĩa, nếu có thì thực hiện thay thế từ đó bằng từ đầu tiên trong mục từ chứa nó.

1.1.5. Mô hình hóa văn bản

Việc cuối cùng trong giai đoạn tiền xử lý là mô hình hóa văn bản, sử dụng mô hình không gian vector. Tương tự các công thức dùng để mô hình hóa văn bản ở trên, để mô hình hóa câu, ta sử dụng công thức sau TF.ISF, công thức này tương tự như TF.IDF nhưng các thông số ở trong phạm vi câu và văn bản. Cụ thể mỗi từ tần số của mỗi từ w_i trong câu s_j được tính như sau:

$$w_{ij} = \begin{cases} \alpha * [1 + \log(f_{ij})] * \log \frac{m}{h_i} & \text{nếu } h_{ij} > 0 \\ 0 & \text{nếu ngược lại} \end{cases}$$

Trong đó:

f_{ij} là số lần xuất hiện của từ t_i trong câu s_j ,

m là tổng số câu trong văn bản

h_i là tổng số câu mà từ t_i xuất hiện.

α là hệ số đánh giá độ quan trọng của từ, nếu từ xuất hiện trong truy vấn thì $\alpha > 1$, còn lại thì $\alpha = 1$

Với hệ số α cho từ xuất hiện trong truy vấn, trong quá trình kiểm thử trên tập mẫu thì $\alpha = 4$ cho kết quả tốt nhất.

1.1.6. Chọn câu phù hợp tạo tóm tắt

Bước này sẽ áp dụng các giải thuật đánh giá câu quan trọng để đưa vào kết quả tóm tắt. Để hạn chế hiện tượng trùng lặp thông tin trong kết quả tóm tắt, trước khi đưa vào lựa chọn, các câu sẽ được so sánh với nhau để tìm các câu gần tương tự nhau, và loại bỏ câu có vị trí xa tiêu đề hơn. Độ đo sử dụng để loại bỏ câu trùng lặp và chọn câu phù hợp tạo tóm tắt là độ đo cosin đã trình bày ở trên, nhưng hai vector được tính toán bây giờ là biểu diễn cho hai câu.

Giải thuật loại bỏ câu trùng lặp như sau:

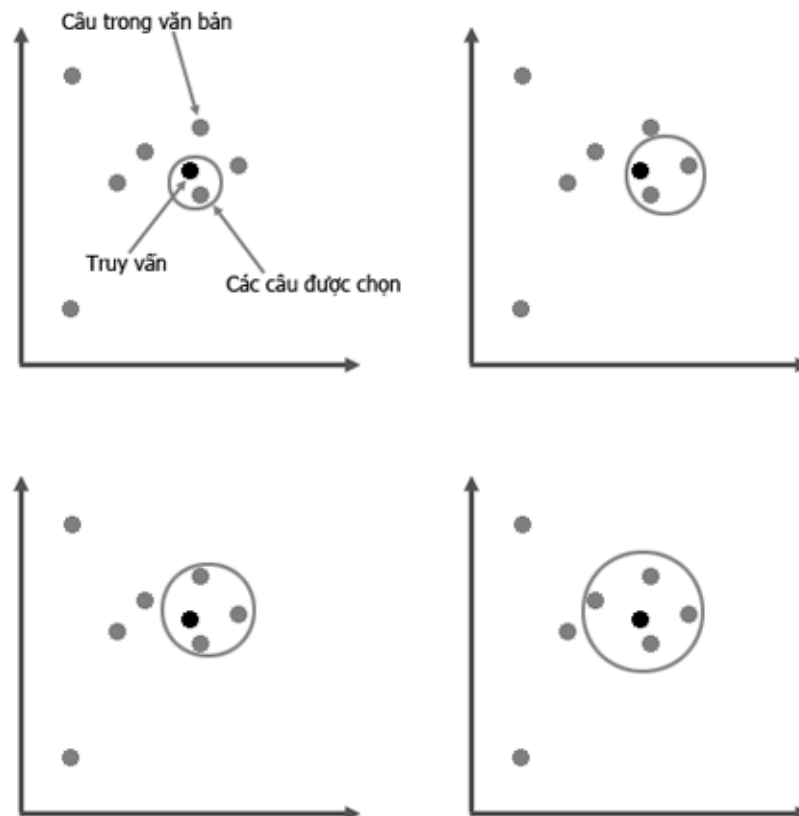
Bước 1: xét câu s_i , tính độ tương đồng với các câu sau nó s_j

Bước 2: với mỗi câu s_j , nếu độ tương đồng $\Omega_{ij} > \alpha$ thì loại bỏ câu s_j

Bước 3: nếu hết văn bản thì dừng lại, không thì tăng i lên 1 và quay lại **bước 1**

Qua thực nghiệm trên một số văn bản, cho thấy ngưỡng $\alpha=0.8$ cho kết quả tương đối chính xác. Do đó trong bước này sẽ thực hiện loại bỏ một câu nếu có độ tương tự lớn hơn 0.8 với câu nào đó đứng trước nó, theo thứ tự vị trí trong văn bản.

Quá trình chọn câu quan trọng sẽ thực hiện như hình dưới đây



Hình 4: Minh họa quá trình chọn câu quan trọng

Sau khi chuyển biểu diễn các câu về mô hình không gian vector, mỗi câu sẽ là một vector, văn bản là danh sách các vector, độ tương đồng giữa các câu sẽ được tính toán sử dụng độ đo cosin.

Giải thuật chọn câu theo các bước sau:

Bước 1: khởi tạo tâm là truy vấn

Bước 2: tính độ tương đồng Ω của các câu trong văn bản với tâm

Bước 3: chọn câu có Ω lớn nhất, kết hợp vào tâm, xóa câu đó khỏi văn bản

Bước 4: kiểm tra độ dài, nếu chưa đủ, tính toán lại tâm và quay lại **bước 2**

Tâm của tóm tắt sẽ được tính toán lại dựa trên công thức tính vector trọng tâm của nhóm, và độ tương tự của 1 câu với tâm sẽ là độ tương tự với vector đó.

*) Véc tơ trọng tâm của nhóm

Giả sử có một tập câu = $\{s_1, s_2, \dots, s_m\}$ có lần lượt các véc tơ biểu diễn là v_1, v_2, \dots, v_m . Khi đó, véc tơ trọng tâm của tập câu được tính theo công thức:

$$\overrightarrow{V_{cen}} = \frac{\sum_{i=1}^m \vec{v}_i}{m}$$

1.2. Giai đoạn hiển thị

Ở bước này, văn bản tóm tắt sẽ được tạo ra bằng cách ghép các câu được chọn theo thứ tự trong văn bản, đó chính là phương pháp hiển thị phân đoạn.

II. CÀI ĐẶT THỬ NGHIỆM

2.1. Chương trình thử nghiệm

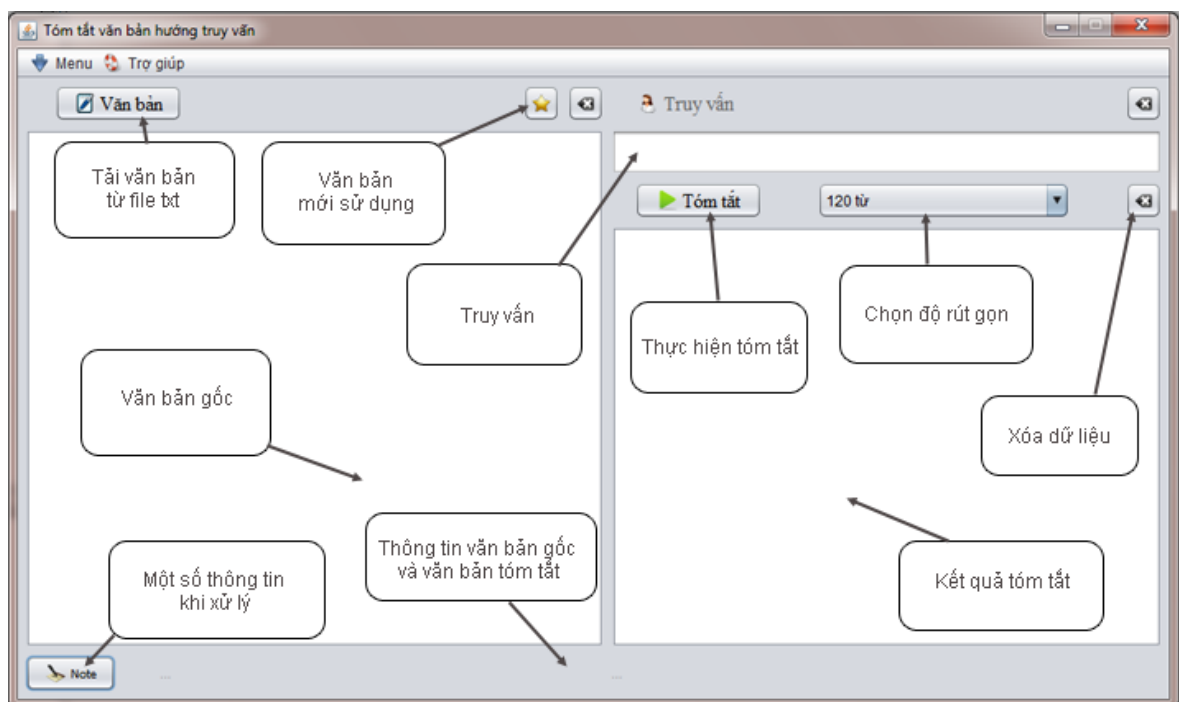
Để thực hiện thử nghiệm em đã xây dựng một số công cụ phục vụ tóm tắt 1 văn bản, công cụ tạo mẫu và công cụ kiểm thử trên mẫu:

- Môi trường cài đặt: Java JDK 7u17, Windows 7 32bit.
- Công cụ lập trình Netbeans 7.3.

2.1.1. Các công cụ đã xây dựng

2.1.1.1. Chương trình tóm tắt

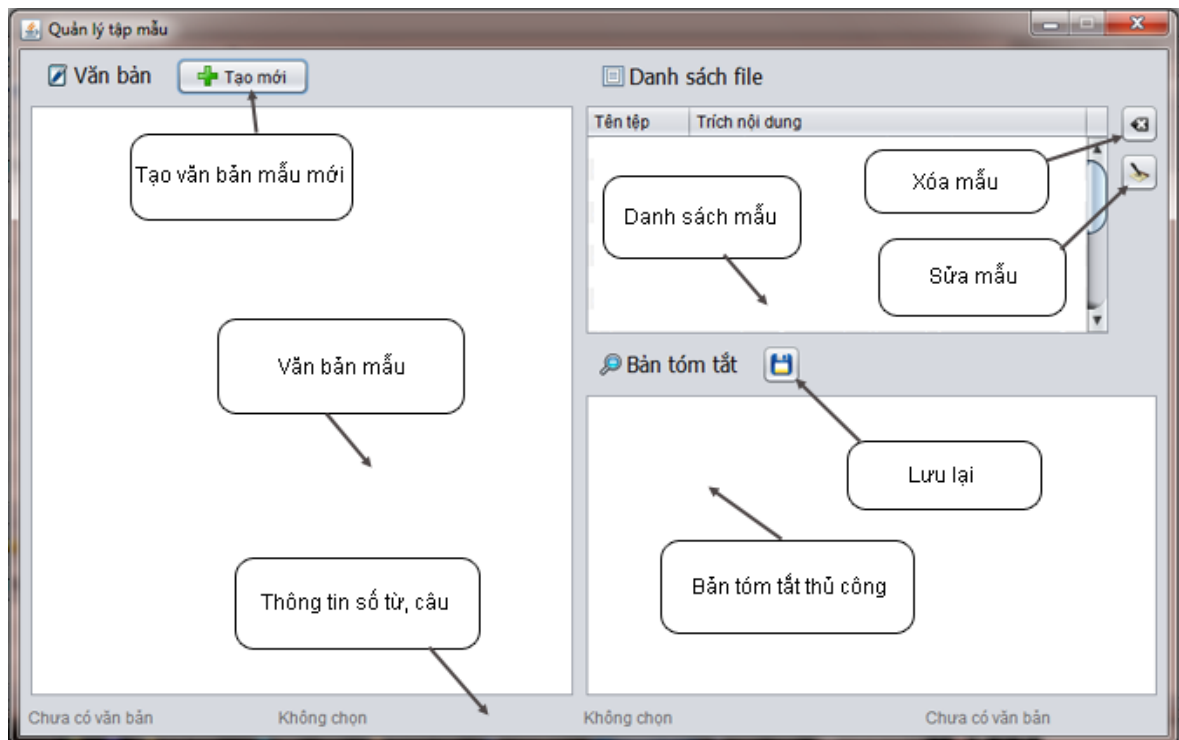
Đây là chương trình thực hiện tóm tắt một văn bản dựa trên giải thuật đã phân tích ở trên. Chi tiết các chức năng đã ghi chú đầy đủ trên ảnh giao diện chương trình. Đầu vào của chương trình là văn bản gốc, truy vấn, và độ rút gọn, đầu ra sẽ là văn bản tóm tắt, có thể xem chi tiết một số bước xử lý ở chức năng Note góc dưới trái giao diện.



Hình 5: Giao diện chương trình demo

2.1.1.2. Công cụ tạo tập mẫu

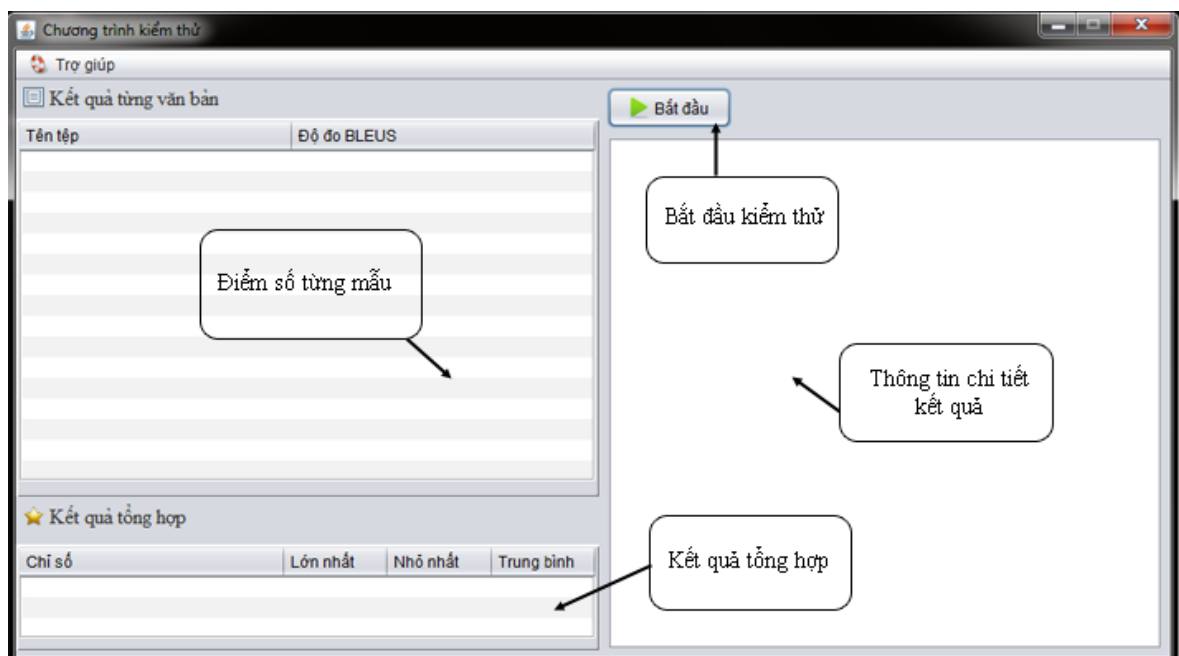
Công cụ này hỗ trợ, tạo, chỉnh sửa các bản tóm tắt thủ công. Chức năng chính là quản lý các văn bản mẫu bao gồm văn bản gốc và bản tóm tắt thủ công, được tích hợp chức năng tách từ, tách câu của VNTokenizer nên việc tạo văn bản mẫu sẽ chính xác và hiệu quả hơn. Ngoài ra còn có chức năng phát hiện ra các văn bản lỗi font, các văn bản này không thể sử dụng trong các công cụ đi kèm nên cần loại bỏ.



Hình 6: Chương trình quản lý tập mẫu

2.1.1.3. Công cụ kiểm thử

Công cụ này được xây dựng dựa trên việc tích hợp giải thuật đã đề xuất ở trên và tích hợp thêm hai giải thuật để so sánh, việc so sánh dựa trên độ đo BLEUS, chi tiết về cách thực hiện sẽ trình bày ở phần sau.



Hình 7: Giao diện chương trình kiểm thử

2.2. Thử nghiệm một văn bản

Phần này em sử dụng công cụ tóm tắt đã xây dựng để thử nghiệm một văn bản. Kết quả thực hiện thu được như sau:

2.2.1. Đầu vào

➤ Văn bản:

Bảo vệ vững chắc độc lập chủ quyền lãnh thổ bằng biện pháp hòa bình{1}

Chiều ngày 26-4, Chủ tịch nước Trương Tấn Sang và Tổ Đại biểu Quốc hội (ĐBQH) số 1, Đoàn ĐBQH TP Hồ Chí Minh tiếp tục có buổi tiếp xúc với gần 400 cử tri của quận 1{2}. Ghi nhận các ý kiến của cử tri, Chủ tịch nước đánh giá cao tinh thần đóng góp ý kiến của mọi người, nhất là vấn đề sửa đổi Hiến Pháp và các đạo luật{3}. Trả lời câu hỏi được đông đảo cử tri quan tâm về chủ trương bảo vệ chủ quyền lãnh thổ, củng cố quốc phòng - an ninh, Chủ tịch nước Trương Tấn Sang khẳng định chủ trương của Đảng, Nhà nước trước sau như một là kiên quyết bảo vệ vững chắc độc lập chủ quyền lãnh thổ bằng biện pháp hòa bình, theo hệ thống luật pháp quốc tế{4}. Tuy nhiên, Chủ tịch nước cũng khẳng định “không bao giờ bảo vệ chủ quyền bằng nói miện”; chủ trương hòa hiếu không có nghĩa là không làm gì{5}. Nước ta cũng mua sắm trang bị vũ khí, nhưng không phải để gây chiến tranh, chạy đua vũ trang mà là tăng cường phòng thủ, bảo vệ chủ quyền lãnh thổ{6}. Chủ tịch nước cho biết, chủ trương hòa hiếu luôn được các nước bạn bè trên thế giới ủng hộ{7}.

Đề cập đến tình hình biển, đảo, Chủ tịch nước bày tỏ thông cảm với những lo lắng, bức xúc của cử tri, mong cử tri phải bình tĩnh, không nghe những lời kích động của kẻ xấu{8}. Những mâu thuẫn trên Biển Đông là có, nhưng biện pháp hòa hiếu của nước ta đã có kết quả tốt, Nhà nước luôn hỗ trợ ngư dân ra khơi, số lượng tàu cá đánh bắt xa bờ ngày càng tăng{9}. Nước ta phấn đấu đến năm 2020 sẽ phát triển kinh tế biển đạt 52%-53% GDP, trong đó, dầu khí, vận tải biển, đánh bắt hải sản là thế mạnh lớn{10}. Mục tiêu cuối cùng của nước ta là chủ quyền lãnh thổ vững chắc, quốc phòng - an ninh ổn định, kinh tế phát triển{11}.

Liên quan đến các vấn đề kinh tế - xã hội, Chủ tịch nước Trương Tấn Sang cho biết kinh tế nước nhà có những phát triển đáng kể, nông nghiệp đạt nhiều thắng lợi, các ngành thuộc về dầu khí tăng trưởng khá{12}. Tuy nhiên, Chủ tịch nước mong cử tri hiểu kinh tế Việt Nam dùng chủ yếu là tiền mặt, nên sẽ có những hệ quả về giá cả thị trường, thu nhập người dân liên quan đến việc tăng lương, tăng giảm giá vàng, giá xăng, dầu{13}.

Tổng số 259 từ / 13 câu

➤ Câu truy vấn: bảo vệ chủ quyền lãnh thổ

➤ **Độ rút gọn:** 100 từ

2.2.2. Kết quả tóm tắt

Kết quả được chọn theo thứ tự 4, 11, 6, 5, 7, 9 tổng số 111 từ / 6 câu

{4} Trả lời câu hỏi được đồng đảo cử tri quan tâm về chủ trương bảo vệ chủ quyền lãnh thổ, củng cố quốc phòng - an ninh, Chủ tịch nước Trương Tấn Sang khẳng định chủ trương của Đảng, Nhà nước trước sau như một là kiên quyết bảo vệ vững chắc độc lập chủ quyền lãnh thổ bằng biện pháp hòa bình, theo hệ thống luật pháp quốc tế.

{5} Tuy nhiên, Chủ tịch nước cũng khẳng định “không bao giờ bảo vệ chủ quyền bằng nói miệng”; chủ trương hòa hiếu không có nghĩa là không làm gì.

{6} Nước ta cũng mua sắm trang bị vũ khí, nhưng không phải để gây chiến tranh, chạy đua vũ trang mà là tăng cường phòng thủ, bảo vệ chủ quyền lãnh thổ.

{7} Chủ tịch nước cho biết, chủ trương hòa hiếu luôn được các nước bạn bè trên thế giới ủng hộ.

{9} Những mâu thuẫn trên Biển Đông là có, nhưng biện pháp hòa hiếu của nước ta đã có kết quả tốt, Nhà nước luôn hỗ trợ ngư dân ra khơi, số lượng tàu cá đánh bắt xa bờ ngày càng tăng.

{11} Mục tiêu cuối cùng của nước ta là chủ quyền lãnh thổ vững chắc, quốc phòng - an ninh ổn định, kinh tế phát triển.

2.2.3. Nhận xét

Sau khi chạy thử nghiệm trên một số văn bản, em nhận thấy kết quả tóm tắt khá chính xác, đã nêu lên được các vấn đề liên quan tới truy vấn mà trong văn bản trình bày. Để đánh giá chất lượng thực sự của mô hình, trong phần sau sẽ thực hiện kiểm thử trên lượng dữ liệu đủ lớn.

2.3. Thử nghiệm trên tập mẫu

2.3.1. Dữ liệu thử nghiệm

Các mẫu trong thử nghiệm lần này được tạo ra bằng cách sử dụng công cụ hỗ trợ đã nêu ở trên với văn bản mẫu là các bài báo trên các báo điện tử:

- <http://vnexpress.net/>
- <http://laodong.com.vn/>
- <http://dantri.com.vn/>

Các bài báo sử dụng được lấy từ các chuyên mục: Văn hóa, Xã hội, Chính trị, Pháp luật, Kinh tế.

Độ dài mỗi bản tóm tắt thủ công là xấp xỉ 120 từ. Bảng mã Unicode utf-8, định dạng .txt, số lượng 50 mẫu.

Thông tin độ dài của tập mẫu được trình bày ở bảng dưới đây:

	Lớn nhất	Nhỏ nhất	Trung bình
Độ dài văn bản theo từ (đã loại bỏ từ dừng)	823	180	371
Độ dài văn bản theo câu	35	9	18

Bảng 3: Thông tin tập mẫu sử dụng để đánh giá

2.3.2. Độ đo BLEUS

Độ đo sẽ sử dụng trong phần đánh giá này là BLEUS, cải tiến của độ đo BLEU, sử dụng cho n-gram của từ.

❖ N-gram

N-gram của từ là chuỗi gồm n từ, tập các n-gram của một văn bản được tạo nên bằng cách ghép n từ liên tiếp cho tới khi hết văn bản.

N-gram	Hôm_nay trời mưa to
Unigram(1-gram)	Hôm_nay, trời, mưa, to
Bigram(2-gram)	Hôm_nay trời, trời mưa, mưa to
Trigram(3-gram)	Hôm_nay trời mưa, trời mưa to
Fourgram(4-gram)	Hôm_nay trời mưa to

Bảng 4: Ví dụ về n-gram

❖ Độ đo BLEUS

BLEU là độ đo dựa trên sự đồng hiện của các n-gram, bao gồm 1-gram, 2-gram, 3-gram, 4-gram.

Công thức của độ đo này như sau:

$$BLEU(D1, D2) = \beta * e^{\frac{1}{4} \sum_{k=1}^4 \ln \left(\frac{x_k}{y_k} \right)}$$

Trong đó:

D1 là bản tóm tắt tự động(do chương trình tạo ra)

D2 là bản tóm tắt thủ công

X_k là số k-gram trùng nhau ở hai văn bản

Y_k là số k-gram trong văn bản D1

β là điểm phạt, được tính như sau:

$$\beta = \begin{cases} e^{1-\frac{a}{b}} & \text{nếu } a < b \\ 1 & \text{nếu ngược lại} \end{cases}$$

với a là số 1-gram trong D_2 , b là số 1-gram trong D_1

Nhược điểm của độ đo BLEU là sẽ trả về 0 nếu như hai văn bản không có 4-gram nào trùng nhau, do đó ta sẽ sử dụng một dạng khác của BLEU đó là BLEUS [11] (Smooth BLEU). Độ đo này khắc phục nhược điểm của độ đo BLEU bằng cách thay $X_k = 0$ thành 2^{-n} , cụ thể như sau:

- Nếu $k=2$ thì $X_2=1/2$, $X_3=1/4$, $X_4=1/8$
- Nếu $k=3$ thì $X_3=1/2$, $X_4=1/4$
- Nếu $k=4$ thì $X_4=1/2$

2.3.3. Kết quả thử nghiệm

Để tính toán kết quả, mỗi mẫu sẽ được thực hiện để tạo bản tóm tắt và thực hiện tính điểm BLEUS, chi tiết như sau:

Bước 1: Tải văn bản gốc và văn bản tóm tắt tương ứng, tách tiêu đề của văn bản gốc làm truy vấn.

Bước 2: Gọi modul tóm tắt với đầu vào là văn bản gốc và độ rút gọn 120 từ.

Bước 3: Thực hiện tách từ (sử dụng công cụ vntokenizer) cho văn bản tóm tắt thủ công và văn bản tóm tắt tự động.

Bước 4: Tính điểm số dựa vào độ đo BLEUS cho từng kết quả tóm tắt.

Bước 5: Hiện thị kết quả lên giao diện.

Sau khi thực hiện giải thuật đánh giá, kết quả thống kê thu được như sau:

	Lớn nhất	Nhỏ nhất	Trung bình
Điểm số	0.806	0.254	0.518

Bảng 5: Kết quả kiểm thử độ đo BLEUS của tập mẫu

Một số nhận xét về kết quả kiểm thử:

- Tốc độ thực hiện nhanh (trung bình khoảng 60ms)
- Theo Papineni và đồng sự [9], thì độ đo BLEU từ 0.3 là chấp nhận được, từ 0.5 là tương đối tốt, như vậy thì điểm số trung bình của mẫu là tương đối tốt, một số mẫu có điểm số nhỏ nhưng số lượng không đáng kể

2.3.4. Nhận xét, đánh giá mô hình

- Ưu điểm: tốc độ nhanh, kết quả tóm tắt tương đối tốt, không cần sử dụng dữ liệu học.
- Nhược điểm: vẫn mang nhược điểm của phương pháp tóm tắt trích rút là đứt mạch, nhập nhằng tham chiếu
- Khả năng ứng dụng: với tốc độ thực hiện nhanh, cài đặt đơn giản, tuy vẫn có nhược điểm của tóm tắt trích rút nhưng mô hình này hoàn toàn có thể cài đặt sử dụng trong thực tế, vì với máy tìm kiếm thì đưa ra thông tin quan trọng hơn sự liên mạch của văn bản tóm tắt.

PHẦN 3. KẾT LUẬN VÀ ĐỀ XUẤT

1. Các công việc đã thực hiện được

Về cơ bản, đồ án đã thực hiện được các mục tiêu đề ra ban đầu:

- Tìm hiểu về tóm tắt văn bản tự động
- Đề xuất và phân tích các bước thực hiện một mô hình tóm tắt văn bản hướng truy vấn cho tiếng Việt
- Cài đặt chương trình thử nghiệm và đánh giá kết quả

Tuy nhiên, do tài liệu tham khảo chưa nhiều và thời gian có hạn nên vẫn còn một số việc chưa thực hiện hoặc chưa tốt:

- Cơ sở lý thuyết trình bày còn sơ sài
- Độ chính xác của dữ liệu thử nghiệm chưa cao và số lượng còn ít dẫn đến kết quả đánh giá chưa thật chính xác

2. Đề xuất hướng phát triển

Nhằm tăng chất lượng của mô hình để đưa vào ứng dụng thực tế, em có một số đề xuất như sau:

- Xử lý chi tiết các từ viết tắt
- Tích hợp phân giải đồng tham chiếu
- Có một giải thuật xác định từ dùng thay vì dùng danh sách có sẵn
- Xây dựng từ điển đồng nghĩa đầy đủ và chính xác hơn
- Có thêm modul xử lý các bảng mã khác nhau

TÀI LIỆU THAM KHẢO

Danh sách tài liệu

- [3] Ahmed A. Mohamed; Sanguthevar Rajasekaran, "Query-Based Summarization Based on Document Graphs," Author, Mansfield, Connecticut, US, 2006.
- [4] Wauter Bosma, "Query-Based Summarization using Rhetorical Structure Theory," in Human Media Interaction, Enschede The Netherlands, 2003.
- [5] A. P. Siva kumar ; Dr. P. Premchand; Dr. A. Govardhan, Query-Based Summarizer Based on Similarity of Sentences and Word Frequency, JNTUACE Anantapur, India: International Journal of Data Mining & Knowledge Management Process, 2011.
- [8] Chin-Yew Lin and Franz Josef Och, ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.
- [9] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in 02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 2002.
- [12] Lê Quý Tài, "Nghiên cứu các phương pháp xử lý tiếng Việt ứng dụng cho tóm tắt văn bản," Luận văn thạc sĩ, Hà Nội, 2011.

Danh sách website

- [1] "BLEU - Wikipedia," Wikipedia, [Online]. Available: <http://en.wikipedia.org/wiki/BLEU>. [Accessed 10 05 2013].
- [2] Wikipedia, "ROUGE Metric," Wikipedia, [Online]. Available: [http://en.wikipedia.org/wiki/ROUGE_\(metric\)](http://en.wikipedia.org/wiki/ROUGE_(metric)). [Accessed 30 05 2013].
- [6] "Xử lý văn bản," [Online]. Available: <http://vlsp.vietlp.org:8080/demo/?page=resources>. [Accessed 09 05 2013].
- [7] "KLTN10-wiki," [Online]. Available: <https://code.google.com/p/kltn10-wiki/source/browse/>. [Accessed 18 5 2013].
- [10] "N-Gram Wikipedia," [Online]. Available: <http://en.wikipedia.org/wiki/N-gram>. [Accessed 09 05 2013].
- [11] "KantanMt.com," [Online]. Available: <http://www.kantanmt.com/whatisbleuscore.php>. [Accessed 10 05 2013].