

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NÔNG TIẾN CÔNG

TÓM TẮT VĂN BẢN DỰA VÀO TRÍCH XUẤT CÂU

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

LẠNG SƠN, 2018

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NÔNG TIẾN CÔNG

TÓM TẮT VĂN BẢN DỰA VÀO TRÍCH XUẤT CÂU

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS-TS Đoàn Văn Ban

LẠNG SƠN, 2018

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là sản phẩm nghiên cứu, tìm hiểu của cá nhân tôi. Những điều được trình bày trong luận văn hoặc là của cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Lạng Sơn, tháng 4 năm 2018

TÁC GIẢ LUẬN VĂN

Nông Tiến Công

MỤC LỤC

MỤC LỤC.....	i
DANH MỤC CÁC HÌNH.....	iii
DANH MỤC CÁC BẢNG.....	iv
MỞ ĐẦU.....	1
Chương 1 TÓM TẮT VĂN BẢN TIẾNG VIỆT.....	2
1.1. Bài toán tóm tắt văn bản	2
1.1.1. Phân loại tóm tắt	2
1.1.3. Mô hình tóm tắt văn bản và một số phương pháp tiếp cận	5
1.2. Các phương pháp đánh giá	9
1.2.1. Các phương pháp đánh giá trong.....	10
1.2.2. Các phương pháp đánh giá ngoài	11
1.3. Tóm tắt văn bản tiếng Việt dựa vào trích xuất câu và một số vấn đề liên quan.....	12
1.3.1. Đặc điểm ngôn ngữ trong văn bản tiếng Việt.....	12
1.3.2. Một số hướng tiếp cận bài toán tóm tắt văn bản tiếng Việt ...	15
1.3.3. Mô hình tóm tắt văn bản tiếng Việt dựa vào trích xuất câu ...	17
1.4. Tổng kết chương	18
Chương 2 PHƯƠNG PHÁP TÓM TẮT VĂN BẢN DỰA TRÊN ĐỘ TƯƠNG ĐỒNG CÂU.....	19
2.1. Một số khái niệm và phương pháp tính độ tương đồng câu	19
2.1.1. Độ tương đồng	19

2.1.2. Độ tương đồng ngữ nghĩa và phương pháp trích xuất câu dựa trên độ tương đồng ngữ nghĩa câu.	20
2.1.3. Tính độ tương đồng theo độ đo Cosine	21
2.1.4. Phương pháp tính độ tương đồng câu dựa vào chủ đề ẩn	22
2.1.5. Phương pháp tính độ tương đồng câu dựa vào mạng Wikipedia	25
2.2. Mô hình tóm tắt văn bản tiếng Việt dựa trên trích xuất câu quan trọng theo phương pháp tính độ tương đồng câu.....	28
2.2.1. Giai đoạn tiền xử lý	29
2.2.2. Giai tạo danh sách câu khả dụng	32
2.2.3. Giai đoạn sinh văn bản tóm tắt	34
2.3. Tổng kết chương	34
Chương 3 THỰC NGHIỆM MÔ HÌNH TÓM TẮT VĂN BẢN TIẾNG VIỆT	35
3.1. Môi trường thực nghiệm	35
3.2. Chương trình tóm tắt văn bản	35
3.3. Tiến hành thực nghiệm	37
3.3.1. Cơ sở dữ liệu tổng thể.....	37
3.3.2. Mô hình suy luận chủ đề ẩn.....	37
3.3.3. Dữ liệu thực nghiệm	38
3.3.4. Đánh giá độ chính xác của mô hình tóm tắt văn bản	38
3.4. Tổng kết chương	46
KẾT LUẬN	47
TÀI LIỆU THAM KHẢO.....	48

DANH MỤC CÁC HÌNH

Hình 1.1. Mô hình hệ thống tóm tắt văn bản [13]	5
Hình 1.2. Mô hình chung cho tóm tắt văn bản tiếng Việt dựa vào trích xuất câu.....	17
Hình 2.1. Mô hình tính độ tương đồng câu với chủ đề ản	24
Hình 2.2. Mối quan hệ giữa đồ thị bài viết và đồ thị chủ đề Wikipedia	26
Hình 2.3. Mô hình tóm tắt văn bản tiếng Việt	28
Hình 2.4. Các câu sau khi tách trong cửa sổ nhỏ góc dưới bên trái	29
Hình 2.5. Văn bản sau khi chuẩn hóa	30
Hình 2.6. Xác định từ dừng và ký tự vô ích.....	30
Hình 3.1. Giao diện chương trình	36
Hình 3.2. Các từ đặc trưng của lĩnh vực giáo dục có xác suất xuất hiện cao ở chủ đề 83, 116, 136 trong mô hình suy luận chủ đề ản	38
Hình 3.3. Kết quả tóm tắt văn bản theo phương pháp tổ hợp với tỷ lệ nén 30%.....	40
Hình 3.4. Độ chính xác của các phương pháp tóm tắt theo tỷ lệ nén	44
Hình 3.5. Độ chính xác của các phương pháp tóm tắt ở tỷ lệ nén 30% đối với một số lĩnh vực.....	45

DANH MỤC CÁC BẢNG

Bảng 3.1. Kết quả tóm tắt 6 nhóm văn bản theo tỷ lệ nén 10%.....	41
Bảng 3.2. Kết quả tóm tắt 6 nhóm văn bản theo tỷ lệ nén 20%.....	42
Bảng 3.3. Kết quả tóm tắt 6 nhóm văn bản theo tỷ lệ nén 30%.....	43

MỞ ĐẦU

Với sự phát triển của công nghệ và Internet hiện nay, thông tin thời sự được cập nhật trên các Website với tốc độ vũ bão. Điều đó đã mang lại cho con người rất nhiều lợi ích thiết thực nhưng nó cũng khiến họ gặp phải không ít khó khăn khi sàng lọc lấy thông tin hữu ích từ nguồn dữ liệu khổng lồ ấy.

Theo đánh giá của công ty Oracle¹, hiện có đến 80% dữ liệu trên thế giới là dữ liệu văn bản. Vì vậy, việc tổ chức quản lý và khai thác hiệu quả nguồn dữ liệu này là những bài toán lớn cần được quan tâm nghiên cứu và giải quyết. Tóm tắt văn bản tự động nhằm nhanh chóng thu được những thông tin quan trọng, tăng hiệu quả xử lý thông tin là một trong các hướng tiếp cận khai thác dữ liệu văn bản nhận được sự quan tâm nghiên cứu của nhiều nhà khoa học, nhóm nghiên cứu cũng như các công ty lớn trên thế giới.

Tóm tắt văn bản tự động có nhiều ứng dụng trong thực tế như: tóm tắt tin tức, tóm tắt kết quả tìm kiếm trong các máy tìm kiếm, tóm tắt hình ảnh, tóm tắt video,...²

Do những đặc thù của ngôn ngữ nên việc giải quyết bài toán tóm tắt văn bản tiếng Việt đặt ra cho các nhà nghiên cứu những thách thức, khó khăn riêng. Các kết quả khả quan từ những nghiên cứu về tóm tắt văn bản tiếng Việt được công bố hiện nay là cơ sở cho các dự án xây dựng hệ thống tóm tắt văn bản tiếng Việt tự động hiệu quả trong tương lai [4], [5], [6], [7], [8].

Với việc chọn đề tài **“Tóm tắt văn bản dựa vào trích xuất câu”**, luận văn trung vào việc nghiên cứu, đánh giá và lựa chọn phương pháp xây dựng một mô hình tóm tắt văn bản tiếng Việt hiệu quả.

¹ <http://www.oracle.com/technetwork/testcontent/9ir2text-bwp-f-129974.pdf>

² https://en.wikipedia.org/wiki/Automatic_summarization

Chương 1

TÓM TẮT VĂN BẢN TIẾNG VIỆT

1.1. Bài toán tóm tắt văn bản

Theo Inderjeet Mani thì mục đích của tóm tắt văn bản tự động là: “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” [13].

Theo Radev: “Văn bản tóm tắt là văn bản được tạo từ một hoặc nhiều văn bản khác mà truyền tải được những thông tin quan trọng trong văn bản gốc nhưng có độ dài không quá một nửa văn bản gốc (thường ngắn hơn đáng kể)” [11].

Như vậy, tóm tắt văn bản là việc tìm các ý chính của văn bản. Bản tóm tắt là có ba đặc điểm sau [10], [11], [12], [13]:

- Bảo toàn nội dung chính so với văn bản gốc: Các nội dung quan trọng hay nổi bật của bản gốc phải được giữ lại trong bản tóm tắt.
- Ngắn gọn: bản tóm tắt thường ngắn hơn bản gốc nhiều.
- Dễ đọc: người sử dụng có thể đọc và hiểu được dễ dàng.

Việc đưa ra được một bản tóm tắt có chất lượng và không bị giới hạn bởi miền ứng dụng được xác định là rất khó khăn nên các phương pháp giải quyết bài toán tóm tắt văn bản thường chỉ hướng đến một kiểu văn bản cụ thể hoặc một kiểu tóm tắt cụ thể.

1.1.1. Phân loại tóm tắt

Có nhiều cách phân loại tóm tắt văn bản khác nhau, sau đây là một số cách phân loại tiêu biểu [13]:

1.1.1.1. Theo định dạng đầu ra

- Tóm tắt trích xuất (Extract): là một bản tóm tắt gồm các đoạn văn bản được rút trích từ văn bản gốc.

- Tóm tắt tóm lược (Abstract): là một bản tóm tắt được tạo ra dựa trên các thông tin quan trọng trong văn bản gốc.

1.1.1.2. Theo mức độ xử lý

- Tiếp cận mức ngoài (surface-level): thông tin được miêu tả dưới dạng khái niệm về các đặc trưng nông (shallow feature). Các đặc trưng nông bao gồm các thuật ngữ (term) quan trọng qua thống kê (dựa vào tần số của các thuật ngữ trong văn bản), các thuật ngữ quan trọng dựa vào vị trí, các thuật ngữ trong các cụm từ dấu hiệu hay các thuật ngữ trong câu truy vấn của người dùng. Kết quả là một bản tóm tắt dạng trích xuất (extract).

- Tiếp cận mức sâu (deeper-level): ở mức này, bản tóm tắt có thể là dạng trích xuất hoặc dạng tóm tắt (abstract) và cần phải sử dụng đến sinh tổng hợp ngôn ngữ tự nhiên. Với dạng tiếp cận này, phải cần đến những phân tích về mặt ngữ nghĩa, chẳng hạn sử dụng hướng tiếp cận thực thể để xây dựng dạng biểu diễn của các thực thể văn bản (đơn vị văn bản) và mối quan hệ giữa các thực thể rồi từ đó tìm ra phần quan trọng. Mối quan hệ giữa các thực thể gồm quan hệ ngữ nghĩa như: đồng nghĩa, trái nghĩa, nghĩa hẹp, nghĩa rộng..., quan hệ cú pháp: dựa trên cây phân tích cú pháp và các mối quan hệ khác.

1.1.1.3. Theo mục đích của bản tóm tắt

- Tóm tắt chỉ thị (Indicative): Đưa ra những thông tin ngắn gọn về chủ đề chính của văn bản. Dạng tóm tắt này thường được sử dụng trong các hệ thống tìm kiếm thông tin. Thông thường, độ dài của văn bản tóm tắt loại này chỉ từ 5 đến 10% độ dài của toàn bộ văn bản.

- Tóm tắt thông tin (Information): tóm tắt bao gồm tất cả các thông tin nổi bật có trong văn bản nguồn tại nhiều mức độ chi tiết khác nhau.

- Tóm tắt đánh giá (Evaluation): tóm tắt nhằm mục đích đánh giá vấn đề chính của văn bản nguồn, thể hiện quan điểm của tác giả đối với công việc của họ.