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

DANH MỤC CÁC HÌNH

DANH MỤC CÁC BẢNG

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 Oracle[1], 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 ngọn: 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.