ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Việt Cường

XÂY DỰNG MỤC LỤC CHO VĂN BẢN

Ngành: **Công nghệ thông tin**
Chuyên ngành: **Hệ thống thông tin**
Mã số: **60 48 05**

LUẬN VĂN THẠC SĨ

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS. TS. HÀ QUANG THUỴ

HÀ NỘI – 2007

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới PGS.TS. Hà Quang Thuỵ, người thầy đã dìu dắt tôi suốt bao năm qua trên bước đường nghiên cứu khoa học.

Tôi xin chân thành cảm ơn sự giúp đỡ và góp ý rất nhiệt tình của TS. Nguyễn Lê Minh và TS. Phan Xuân Hiếu trong suốt quá trình nghiên cứu và hoàn thành luận văn này.

Tôi xin chân thành cảm ơn sự giúp đỡ, tạo điều kiện và khuyến khích tôi trong quá trình làm việc và nghiên cứu của tập thể các thầy cô và anh chị em trong Bộ môn Các hệ thống thông tin và Phòng thí nghiệm Công nghệ tri thức và Tương tác người máy.

Và cuối cùng, tôi xin gửi lời cảm ơn tới gia đình, người thân và bạn bè – những người luôn ở bên tôi những lúc khó khăn nhất, luôn động viên tôi, khuyến khích tôi trong cuộc sống và trong công việc.

Tôi xin chân thành cảm ơn!

Hà Nội, tháng 10 năm 2007

Tác giả

Nguyễn Việt Cường

LỜI CAM ĐOAN

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

Luận văn này là mới và không sao chép nguyên bản từ bất kì một nguồn tài liệu nào khác.

MỤC LỤC

DANH MỤC CÁC KÍ HIỆU, CÁC CHỮ VIẾT TẮT

STT	Kí hiệu/Viết tắt	Diễn giải
1	TF	Term Frequency – Tần suất của khái niệm
2	TF * IDF	Term Frequency * Inverse Document Frequency
3		

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Trong vài thập kỉ qua, lượng thông tin được số hoá ngày càng nhiều. Ban đầu là các thư viện với các cuốn sách được lưu trữ số hoá, tiếp đến là các nội dung thông tin được đưa lên Internet dưới nhiều hình thức khác nhau. Hơn thế nữa, với sự ra đời của World Wide Web thì thông tin đã thực sự bùng nổ, con người ngày càng muốn có nhiều thông tin hơn và muốn tìm cách để có thể nắm bắt được thông tin nhanh, chính xác và cô đọng.

Rất nhiều bài toán trong xử lí ngôn ngữ tự nhiên đã được đặt ra và giải quyết nhằm giúp máy tính có thể hiểu được phần nào các văn bản số hoá rồi từ đó trình bày lại theo một hình thức nào đó để giúp con người tìm kiếm và thu thập thông tin nhanh hơn. Các bài toán có thể kể đến như: thu nhận thông tin, phân cụm văn bản, phân lớp văn bản, rút trích thông tin, hệ thống hỏi đáp, tóm tắt văn bản,… Những bài toán này đã phần nào được giải quyết và đã thể hiện phần nào ý nghĩa đối với người sử dụng. Ví dụ như các hệ thống máy tìm kiếm Yahoo!, Google,… đã có thể giúp người dùng thu thập thông tin theo truy vấn, trả lại trang thông tin và tóm tắt nội dung của trang thông tin để giúp con người có thể nhanh chóng tìm ra được thông tin mình cần.

Bài toán tóm tắt văn bản ra đời với vai trò giúp người truy cập thông tin có thể dễ dàng nắm bắt được những nội dung chính của văn bản ở một dạng cô đọng hơn. Một ví dụ điển hình là tủ chứa các thẻ trình bày tóm tắt thông tin về cuốn sách ở các thư viện, nó giúp người đọc có thể tìm kiếm nhanh tới cuốn sách mình cần. Hay trong thời đại thông tin được số hoá hiện nay, ở đầu mỗi bài báo hay một bài trình bày hoặc một bài viết dài về một vấn đề nào đó, người ta thường đưa thêm vào một đoạn tóm tắt ngắn của toàn bộ nội dung. Tuy nhiên, không phải lúc nào thông tin tóm tắt đó cũng có sẵn, một phần vì các tóm tắt đó được thực hiện theo phương pháp thủ công và đôi khi không phải do chính tác giả viết ra. Từ đó đặt ra vấn đề là làm sao để có thể tự động hoá quá trình tóm tắt văn bản dựa trên nội dung sẵn có.

Trên thế giới đã có rất nhiều công trình nghiên cứu về vấn đề này và cũng nghiên cứu cách thức tóm tắt theo nhiều hướng khác nhau, từ rút trích một đoạn văn, rút trích một vài câu quan trọng cho tới rút trích các cụm từ có ý nghĩa; rồi từ tóm tắt trên một văn bản tới tóm tắt trên phạm vi nhiều văn bản;… Tuy nhiên hầu hết các phương pháp hiện tại đều áp dụng cho các văn bản tương đối ngắn như tin tức, bài hướng dẫn, bài trình bày,… và không có tính chất định vị thông tin. Đối với các văn bản cỡ lớn hơn như tài liệu nghiên cứu, sách,… thì có rất ít

các công trình nghiên cứu. Trong số đó có một bài toán được quan tâm đặc biệt trong thời gian gần đây, đó là bài toán xây dựng mục lục cho văn bản. Cơ sở của bài toán này là bản thân mục lục của một tài liệu dài không những chứa một lượng lớn thông tin về nội dung của văn bản mà còn có khả năng định vị thông tin bên trong văn bản. Ngoài ra các tiêu đề nằm ở mục lục còn manh tính súc tích cao.

Với thực tế như đã trình bày ở trên, luận văn tiến hành nghiên cứu và đề xuất phương pháp xây dựng mục lục cho văn bản thông qua đề tài "***Xây dựng mục lục cho văn bản***". Mục tiêu của luận văn là nghiên cứu, giải quyết và đề xuất phương pháp giải quyết bài toán xây dựng mục lục cho văn bản cỡ trung bình và lớn thông qua các công trình nghiên cứu hiện tại trên thế giới. Cơ sở của đề tài là các kết quả nghiên cứu đã được công bố trên thế giới về bài toán phân đoạn văn bản và bài toán sinh tiêu đề cho văn bản. Luận văn cũng tiến hành thử nghiệm trên một vài văn bản với sự đánh giá của các chuyên gia là các nhà ngôn ngữ học để đánh giá về tính chính xác của kết quả đạt được. Các kết quả bước đầu đạt được cho thấy hướng nghiên cứu của luận văn là có triển vọng và có khả năng phát triển tiếp thành một bài toán tổng thể cỡ lớn hơn.

Ngoài phần mở đầu và kết luận, kết cấu của luận văn bao gồm 4 chương:

- Chương 1 "***Giới thiệu bài toán***" tóm tắt một số bài toán trong lĩnh vực tóm tắt văn bản, phát biểu bài toán xây dựng mục cho văn bản, đồng thời phần tích các công trình có liên quan và đưa ra phương hướng giải quyết.
- Chương 2 "***Các phương pháp giải quyết bài toán***" trình bày các phương pháp dùng trong quá trình xây dựng mục lục, phân tích điểm mạnh và yếu của mỗi phương pháp.
- Chương 3 "***Xây dựng mục lục cho văn bản***" sẽ đi sâu vào việc tích hợp các thuật toán để giải quyết bài toán chính của luận văn, đồng thời đề xuất một số hướng cải tiến và cơ sở lí luận của các cải tiến đó.
- Chương 4 "***Thử nghiệm và đánh giá***" sẽ trình bày quá trình thử nghiệm của luận văn và các kết quả đạt được trong quá trình thử nghiệm. Đồng thời cũng đưa ra các phân tích và đánh giá về kết quả đạt được.

Chương 1

GIỚI THIỆU BÀI TOÁN

1.1. Bài toán tóm tắt văn bản

Lượng thông tin trên Internet, trong các tài liệu và trong các cơ sở dữ liệu đang không từng tăng lên dẫn đến nhu cầu tìm kiếm và biểu diễn thông tin hiệu quả. Các hệ thống thu nhận thông tin (Information Retrieval) đã cho phép tìm kiếm và sắp xếp thông tin nhận được theo mức độ liên quan đến câu hỏi truy vấn của người dùng []. Gần đây, các hệ thu nhận thông tin còn đưa ra các đoạn tóm tắt của thông tin trả về để giúp người dùng dễ dàng chọn lựa có xem thông tin đó hay không, các đoạn tóm tắt này thường đưa ra các ý chính trong văn bản tương ứng và một đoạn tóm tắt lí tưởng là đoạn tóm tắt đưa ra được tất cả các ý chính của văn bản, đặc biệt là đưa ra được những ý mà người dùng mong muốn. Điều này thực sự có ý nghĩa khi số lượng tài liệu có liên quan đến câu truy vấn là rất lớn trong khi ta chỉ có đủ thời gian để xem những tài liệu liên quan nhiều đến vấn đề cần tìm hiểu.

Bài toán tóm tắt văn bản đã có lịch sử từ lâu đời, ví dụ như công việc của một người thư kí, có trách nhiệm tóm tắt lại những ý chính của tài liệu (tóm tắt đơn văn bản) hoặc tổng hợp thông tin trên nhiều tài liệu (tóm tắt đa văn bản). Hay trong các thư viện, người thủ thư phải đọc qua tài liệu để tóm tắt ý chính hoặc đưa ra các từ khoá trên các thẻ bài để người đọc có thể tìm thấy tài liệu dễ dàng. Trong thời kì thông tin được số hoá, bài toán tóm tắt văn bản số (sau đây gọi chung là văn bản) được giải quyết lần đầu tiên trong bài báo của Luhn năm 1958. Trong bài báo này, Luhn giải quyết bài toán tạo ra một đoạn tóm tắt (abstract) cho các tài liệu kĩ thuật. Những năm sau đó, bài toán được tiếp tục phát triển với nhiều cải tiến mới [Paice 1990, Tait 1983]. Và khi Internet thực sự đi vào cuộc sống con người (từ những năm 90) thì bài toán được quan tâm nhiều hơn. Một vài hướng tiếp cận đã được triển khai: tiếp cận theo hướng ngôn ngữ học [], và tiếp cận theo hướng thống kê [] hoặc kết hợp cả hai [].

Tóm tắt văn bản tự động để đạt được mức như con người là một bài toán khó vì việc hiểu ngôn ngữ tự nhiên là một bài toán khó. Việc xây dựng một công cụ tóm tắt tổng quát là rất khó khăn do các yếu tố ảnh hưởng đến việc tóm tắt rất đa dạng, như phong cách viết, thể loại văn bản, từ vựng, cấu trúc câu,… Do vậy, các công cụ tóm tắt văn bản thường chỉ tập trung theo một mục tiêu nào đó như theo thể loại văn bản, theo mục đích sử dụng,… Có thể kể ra một vài bài toán tóm tắt văn bản theo các hướng khác nhau như sau [Gol]:

- *Các thức xây dựng:* Một đoạn tóm tắt kiểu ngôn ngữ tự nhiên được tạo ra bằng việc sử dụng các biểu diễn ngữ nghĩa để phản anh cấu trúc và các ý chính của văn bản, trong khi tóm tắt kiểu trích dẫn chứa một vài đoạn văn bản trong văn bản gốc.
- *Kiểu:* Tóm tắt tổng quát sẽ đưa ra những ý chung của văn bản, trong khi tóm tắt hướng truy vấn sẽ đưa ra những nội dung có liên quan truy vấn của người dùng.
- *Mục đích:* Tóm tắt chỉ dẫn sẽ đưa ra thông tin tổng quan về văn bản hoặc một tập hợp văn bản, trong khi tóm tắt thông tin sẽ đưa ra nhiều thông tin hơn giúp người dùng có thể lấy ra các thông tin cốt lõi. Mục đích của tóm tắt thông tin có thể coi như một sự thay thế cho văn bản gốc.
- *Số lượng văn bản:* Tóm tắt đơn văn bản đưa ra tóm tắt của một văn bản, trong khi tóm tắt đa văn bản sẽ đưa ra thông tin tóm tắt dựa trên một tập hợp nhiều văn bản.
- *Độ dài văn bản:* Độ dài của một văn bản thường chỉ ra mức độ dư thừa thông tin của văn bản. Ví dụ, một bản tin thường chỉ liên quan đến một chủ đề, do đó sẽ chỉ chứa ít thông tin dư thừa. Tuy nhiên, một tài liệu khoa học thường trình bày về một vấn đề nào đó, diễn giải vấn đề và sau đó lặp lại ở phần kết luận.
- *Thể loại:* Thông tin về thể loại văn bản sẽ rất hữu ích cho quá trình tóm tắt văn bản. Các thể loại khác nhau bao gồm: tin tức, ý kiến/quan điểm, thư hoặc bản ghi nhớ, email, tài liệu khoa học, sách, trang web hay đoạn hội thoại.
- *Mục đích của người sử dụng:* Việc xác định mục đích của người sử dụng cũng sẽ ảnh hưởng đến việc chọn cách thức tóm tắt. Người sử dụng đang xem lướt các thông tin hay đang tìm kiếm một thông tin cụ thể?

Luận văn này sẽ tập trung vào tóm tắt văn bản theo kiểu chỉ dẫn. Hay nói chính xác hơn, mục tiêu của luận văn là đưa ra tiêu đề cho các phần khác nhau trong một văn bản cỡ trung bình và dài giúp người sử dụng có thể vừa xem được các ý chính trong văn bản, đồng thời định vị được vị trí của thông tin đó. Văn bản được coi là cỡ trung bình nếu có độ dài khoảng 3-50 trang và được coi là dài nếu có độ dài trên 50 trang [Gol]. Ví dụ các mẩu tin tức và các trang web được

coi là văn bản ngắn, còn các tài liệu khoa học (ví dụ như một báo cáo khoa học chuyên ngành) được coi là văn bản cỡ trung bình và dài.

1.2. Bài toán xây dựng mục lục cho văn bản

Các nghiên cứu giải quyết bài toán tóm tắt văn bản hầu hết chỉ tập trung vào việc xử lí các văn bản ngắn, đặc biệt là các mẩu tin tức hoặc bài viết nhỏ []. Hơn thế nữa, các phương pháp được đề ra cũng thường chỉ tập trung cho các văn bản thuộc một lĩnh vực cụ thể nào đó []. Điều này đã làm bỏ ngỏ một lĩnh vực nghiên cứu tóm tắt văn bản cho các văn bản cỡ trung bình và dài như tài liệu kĩ thuật hoặc các cuốn sách. Hiện tại cũng đã có một vài công trình được công bố nhằm giải quyết bài toán này nhưng hầu như cũng vẫn chỉ dùng các cách thức cũ để áp dụng cho bài toán lớn hơn [].

Luận văn này sẽ tiến hành nghiên cứu một bài toán khá mới mẻ, đó là bài toán xây dựng mục lục cho văn bản []. Đây là một kiểu tóm tắt chỉ dẫn rất thích hợp cho việc truy cập thông tin trong những văn bản dài. Mục lục là nơi liệt kê ra danh sách các chủ đề trong tài liệu và vị trí tương ứng của từng chủ đề. Danh sách các chủ đề trong một văn bản, xét theo một khía cạnh nào đó cũng là một dạng tóm tắt giàu thông tin vì nó thường có độ dài vừa phải và chứa được tất cả những ý cốt lõi nhất trong văn bản. Ngoài ra một mục lục có cấu trúc phân cấp sẽ càng cho được nhiều ý nghĩa hơn về các chủ đề lớn, nhỏ trong văn bản đó. Ví dụ, chúng ta thường tìm thấy mục lục trong các cuốn sách, là nơi để tìm kiếm và định vị nhanh các thông tin chúng ta quan tâm, hơn thế nữa nó còn trình bày danh sách tất cả các chủ đề được trình bày trong cuốn sách, hay nói cách khác, đó là một bản tóm tắt cho cuốn sách.

1.3. Phương hướng giải quyết bài toán

Có thể phát biểu một cách ngắn gọn bài toán xây dựng mục lục cho văn bản như sau: *Cho trước một văn bản, cần phải sinh ra một cây, trong đó mỗi nút là một đoạn văn bản và tiêu đề của đoạn văn bản tương ứng*. Quá trình này liên quan đến hai bài toán khác:

- ***Phân đoạn văn bản (Text Segmentation):*** phân văn bản thành các đoạn độc lập và liên tục với nội dung các phần có sự tách biệt về mặt ngữ nghĩa.
- ***Sinh tiêu đề (Title Generation):*** sinh ra các tiêu đề ngắn gọn, giàu thông tin cho đoạn văn bản tương ứng.

Đối với bài toán thứ nhất, phân đoạn văn bản, ta có thể giải quyết bằng cách sử dụng cấu trúc sẵn có của văn bản (chương, mục, mục con,…) hoặc sử dụng một phương pháp phân đoạn văn bản tự động []. Trong luận văn này, tôi sử dụng hướng tiếp cận thứ hai vì thực tế cho thấy, nếu một văn bản đã được chia thành các chương, mục (hướng tiếp cận thứ nhất) thì bản thân tác giả của văn bản cũng đã xác định tiêu đề cho các phần và do đó việc sinh mục lục sẽ vô cùng đơn giản. Ngoài ra, bài toán phân đoạn văn bản cũng được chia làm hai loại là phân đoạn văn bản một cấp và phân đoạn văn bản đa cấp. Luận văn này sẽ trình bày cả hai hướng tiếp cận cho vấn đề này.

Đối với nhiệm vụ thứ hai, sinh tiêu đề cho một đoạn văn bản, ta có thể sử dụng rất nhiều phương pháp có sẵn để giải quyết []. Các phương pháp có thể kể đến như lựa chọn câu quan trọng [], lựa chọn mệnh đề quan trọng [],… Tuy nhiên việc sinh tiêu đề một cách độc lập đối với từng đoạn văn bản sẽ gây ra tính không thống nhất trong nội dung của mục lục, có thể xảy ra trường hợp các tiêu đều bị lặp lại []. Ngay cả khi các tiêu đề không bị lặp lại thì các tiêu đề này cũng không đảm bảo được việc tách bạch thông tin một cách hiệu quả giữa các đoạn văn bản. Do đó, cần phải điều khiển quá trình sinh mục lục theo một phương thức phối hợp để đảm bảo tính thống nhất cục bộ cũng như toàn cục.

Như vậy, bài toán xây dựng mục lục cho văn bản sẽ được giải quyết thông qua hai bước là phân đoạn văn bản và sinh tiêu đề. Đồng thời luận văn cũng sẽ trình bày một phương pháp để đảm bảo tính thống nhất giữa tên của các tiêu đề trong mục lục.

1.4. Các công trình liên quan

Về khía cạnh độ dài và thể loại văn bản, trong khi hầu hết các nghiên cứu hiện tại tập trung vào các văn bản ngắn thì đã có một số hướng tiếp cận được triển khai để tóm tắt những văn bản dài hơn. Hầu hết các cách tiếp cận này tập trung vào một miền ngữ nghĩa cụ thể như văn bản y tế hoặc tài liệu khoa học. Với việc đưa ra các giả thiết mạnh về cấu trúc văn bản đầu vào và định dạng đầu ra, các cách tiếp cận này đã thu được những kết quả tương đối khả quan. Ví dụ, Teufel và Moens (2002) tóm tắt các văn bản khoa học bằng cách lựa chọn những yếu tố tu từ (rhetorical elements) thường được trình bày trong các đoạn tóm tắt của tài liệu khoa học. Elhadad và McKeown (2001) trình bày cách tiếp cận sinh tóm tắt của các tài liệu y tế bằng việc sử dụng một số cấu trúc mẫu trong lựa chọn nội dung. Tuy nhiên, trong luận văn này, tôi sử dụng cách tiếp cận độc lập thể loại, tức là tóm tắt văn bản mà không sử dụng các yếu tố đặc trưng liên quan để thể loại văn bản.

Về bài toán phân đoạn văn bản, đã có khá nhiều công trình nghiên cứu liên quan đến vấn đề này []. Hầu hết các công trình đều chỉ tập trung nghiên cứu bài toán phân đoạn văn bản một cấp, hay nói cách khác là phân đoạn văn bản tuyến tính []. Cũng đã có một vài công trình đề cập và giải quyết bài toán phân đoạn văn bản đa cấp (2 cấp) bằng cách kết hợp LSI và Space-scale filtering. Trong đó đáng kể nhất phải nói tới công trình của Hearst năm 1994, công trình này là cơ sở cho rất nhiều công trình khác liên quan đến bài toán này, chi tiết sẽ được trình bày ở phần sau. Trong luận văn này, tôi chủ yếu phân tích và sử dụng công trình của Hearst.

Về bài toán sinh tiêu đề, các công trình liên quan đến vấn đề này đã có rất nhiều, tiêu biểu như []. Tuy nhiên, các công trình này chủ yếu được sử dụng để sinh tiêu đề cho một tài liệu đơn lẻ trong khi bài toán của chúng ta là sinh tiêu đề cho nhiều đoạn văn bản có sự liên kết và ràng buộc với nhau về mặt nội dung. Trong luận văn này, tôi sẽ trình bày một mô hình sinh tiêu đề có khả năng đảm bảo được tính thống nhất giữa các tiêu đề được sinh ra. Mô hình này sử dụng kết hợp cả phương pháp lựa chọn lẫn phương phương pháp sử dụng các đặc trưng về ngữ pháp.

Trong chương tiếp theo, luận văn sẽ trình bày ngắn gọn về các phương pháp được sử dụng để giải quyết bài toán phân đoạn văn bản và sinh tiêu đề cho văn bản.

Chương 2

PHÂN ĐOẠN VĂN BẢN VÀ SINH TIÊU ĐỀ

2.1. Phân đoạn văn bản

Bài toán phân đoạn văn bản có thể được hiểu là bài toán với một văn bản cho trước, hãy xác định những vị trí mà ở đó chủ đề thay đổi. Đối với các văn bản ngắn như bài báo hay bản tin thì chỉ có một chủ đề xuyên suốt toàn văn bản, sự phân lập về mặt chủ đề giữa các đoạn hầu như không có. Tuy nhiên trong các văn bản dài hơn, như tài liệu khảo cứu khoa học thì có rất nhiều phần khác nhau, mỗi phần nói về một vấn đề riêng biệt tuy cùng chung một mục đích là giải quyết mục tiêu của văn bản. Bài toán này đã được giải quyết theo một vài hướng khác nhau []. Trong phần tiếp theo, luận văn sẽ trình bày một vài trong số các phương pháp này để làm tiền đề cho các thử nghiệm của luận văn.

Bài toán phân đoạn văn bản không chỉ có ý nghĩa với các văn bản thông thường, nó còn có ý nghĩa lớn với các bài toán liên quan đến văn bản dạng nói hay hình ảnh []. Ví dụ, trong bài phát biểu chào mừng năm học mới của hiệu trưởng, có rất nhiều phần khác nhau như: chào mừng học sinh mới, sơ kết năm học cũ và phương hướng năm học mới. Các phần này không được phân tách một cách rõ ràng như trong văn bản viết, do đó sẽ khó khăn hơn trong vấn đề phân đoạn. Phương pháp được đưa ra là sử dụng các đặc trưng về khoảng lặng giữa các phần, âm điệu thay đổi khi chuyển phần,… hay có thể sử dụng một module nhận dạng giọng nói sau đó sử dụng các phương pháp phân đoạn văn bản thông thường. Hay trong một đoạn băng video dài, người ta có thể quay nhiều lần khác nhau với những chủ đề khác nhau để tiết kiệm đoạn băng. Việc tự động tìm ra các vị trí thay đổi chủ đề một cách tự động thực sự có ý nghĩa khi biên tập.

Khi tiến hành phân đoạn văn bản, chúng ta sẽ gặp phải những vị trí mà sự thay đổi chủ đề là "lớn" (ví dụ, chuyển hẳn sang chủ đề) hoặc sự thay đổi chủ đề là "nhỏ" (ví dụ, vẫn nói về chủ đề lớn nhưng theo một cách tiếp cận khác). Điều đó làm nảy sinh bài toán phân đoạn văn bản đa cấp. Theo cách này, một văn bản ban đầu được chia thành các đoạn lớn mang các chủ đề riêng, sau đó mỗi đoạn văn bản này lại được phân đoạn tiếp để thu được các chủ đề nhỏ hơn, giúp người đọc dễ theo dõi hơn. Việc giải quyết bài toán này có thể đi theo một trong hai hướng tiếp cận: phân đoạn kiểu đệ quy tức là phân đoạn theo từng mức một hoặc phân đoạn một lần tức là chỉ một lần áp dụng thuật toán mà ta thu được văn bản đã được phân thành vài cấp.

Trong phần tiếp theo, luận văn sẽ trình bày một số phương pháp trong phân đoạn văn bản.

2.2. Các phương pháp phân đoạn văn bản

2.2.1. Sử dụng mối liên kết từ vựng

Một số cách tiếp cận giải quyết bài toán phân đoạn văn bản đã được công bố dựa trên độ đo về sự khác nhau trong việc sử dụng từ của hai phân đoạn ở hai phía của đường biên phân đoạn: nếu có sự khác biệt lớn trong việc sử dụng từ ở hai phía của một vị trí phân tác thì đó được coi là đường biên.

Điển hình cho phương pháp này là hệ thống TextTiling của Hearst (1994). Hearst đã sử dụng ý tưởng về mỗi quan hệ liên kết từ vựng trong văn bản trong [Halliday] để tìm ra những vị trí mà ở đó diễn ra sự thay đổi đồng thời của rất nhiều yếu tố như không gian, thời gian, cấu trúc, sự kiện,… và sự thay đổi này là đạt cực đại tại điểm đó. Trong TextTiling, Heart chia văn bản thành các "*tile*". Các *tile* mang ý nghĩa tương đương với các đoạn bị phân lập do quá trình phân đoạn văn bản. Sau đây, luận văn sẽ trình bày tóm tắt thuật toán của Hearst dùng để tìm ra cấu trúc chủ đề nhỏ cho văn bản (mỗi đoạn được coi là 1 chủ đề).

Các nghiên cứu trước đây của [Halliday, Tannen, Walker] cho thấy rằng sự lặp lại của các khái niệm chỉ ra mối liên kết chặt chẽ về mặt ngữ nghĩa. Điều đó đã chỉ ra rằng sự lặp lại của các khái niệm sẽ rất có ích trong việc xác định cấu trúc phân đoạn của văn bản và chúng ta sẽ sử dụng yếu tố lặp lại của các khái niệm với vai trò là yếu tố chỉ ra mối liên kết từ vựng.

Thuật toán này một mở rộng của thuật toán được trình bày trong [Morris] với khả năng ghi lại chuỗi các khái niệm lặp lại. Thuật toán này xác định đường biên bằng cách xem xét các vị trí mà ở đó có sự kết thúc của một chuỗi khái niệm và bắt đầu một chuỗi khái niệm mới.

Thuật toán bao gồm ba phần chính:

- Tokenization.
- Xác định độ tương tự.
- Nhận diện biên.

Tokenization là quá trình chia văn bản đầu vào thành các đơn vị từ vựng độc lập. Trong quá trình này, văn bản được chia thành các "câu giả" với độ dài cố định *w* cho trước (đây là một tham số của thuật toán) chứ không phải là dùng các câu được xác định mang tính cú pháp hoàn chỉnh mặc dù điều này sẽ gây ra

vấn đề chuẩn hoá. Quá trình này sẽ tạo ra các nhóm token và được gọi là *chuỗi token*. Theo các kết quả thực nghiệm, độ dài *w* là 20 sẽ phù hợp với hầu hết các loại văn bản khác nhau. Các token được phân tính hình thái và được lưu trong một bảng và tương ứng với mỗi token là số thứ tự của chuỗi token và tần suất xuất hiện của token tương ứng trong chuỗi token. Đồng thời với nó là vị trí các điểm ngắt đoạn (paragraph break) trong văn bản cũng được lưu trữ. Những từ dừng và từ quá phổ biến cũng được loại ra trong quá trình phân tích hình thái.

Bước tiếp theo sau quá trình tokenization là tiến hành so sánh độ tương tự từ vựng của các cặp khối (block) liền kề của các chuỗi từ vựng. Một tham số quan trọng khác của thuật toán được đưa vào là *kích thước khối* (blocksize) được định nghĩa là số các chuỗi token được nhóm lại cùng nhau để so sánh với một nhóm chuỗi token liền kề khác. Giá trị này được kí hiệu là *k* thay đổi tuỳ theo các văn bản khác nhau, tuy nhiên người ta thường lấy nó là độ dài trung bình tính theo chuỗi token của các đoạn văn bản (paragraph). Trong thực tế, giá trị *k* là 6 sẽ phù hợp với hầu hết các loại văn bản khác nhau. Các đoạn thực sự trong văn bản không được sử dụng do độ dài của nó không đều nhau và gây ra việc so sánh không cân bằng.

Giá trị tương tự sẽ được tính cho tất cả các vị trí ở giữa các chuỗi token. Nghĩa là tại mỗi vị trí *i* ở giữa các chuỗi token, độ tương tự sẽ được tính trên hai khối, khối thứ nhất là các chuỗi token từ $i-k$ tới *i* và khối thứ hai là từ $i+1$ tới $i+k+1$. Các tiếp cận theo kiểu cửa sổ trượt này sẽ làm mỗi chuỗi token được tính $2k$ lần.

Độ tương tự *sim* sẽ được tính theo độ đo cosin cho hai khối b_1 và b_2 với độ dài *k* chuỗi token cho mỗi khối:

$$\mathrm{sim}\left(b_1,b_2\right)=\frac{\sum_t w_{t,b_1}w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2\sum_{t=1}^{n} w_{t,b_2}^2}}$$

Trong đó khái niệm *t* được tính trên tập tất cả các token thu được trong quá trình tokenization và w_{t,b_i} là trọng số được gán cho khái niệm *t* trong khối b_i. Ở đây, trong số được tính đơn giản bằng tần suất của khái niệm tương ứng trong khối (*TF*). Ngoài ra trọng số còn có thể được tính theo công thức $TF*IDF$, tuy nhiên trong các thử nghiệm cho thấy việc chỉ dùng độ đo *TF* thường cho kết quả tốt hơn. Theo công thức này thì nếu độ tương tự giữa hai khối là cao thì chứng tỏ hai khối có nhiều khái niệm chung. Giá trị của độ đo này nằm trong đoạn $\left[0;1\right]$. Ví dụ ta có 2 khối với nội dung như sau:

Khối 1: I like apples.

Khối 2: Apples are good for you

Khi biểu diễn dưới dạng vectơ, hai khối này có nội dung như sau:

Bảng 1. Ví dụ về độ tương tự giữa 2 khối văn bản

Từ	Apples	Are	For	Good	I	Like	You
Khối 1	1	0	0	0	1	1	0
Khối 2	1	1	1	1	0	0	1

Khi đó độ đo tương tự giữa 2 khối này có giá trị:

$$\mathrm{sim}(K_1, K_2) = \frac{1*1+0*1+0*1+0*1+1*0+1*0+0*1}{\sqrt{\left(1^2+0^2+0^2+0^2+1^2+1^2+0^2\right)\left(1^2+1^2+1^2+1^2+0^2+0^2+1^2\right)}} = 0.26$$

Độ đo tương tự này có thể được đồ thị hoá để có cái nhìn trực quan hơn về sự biến đổi trong đó trục x là số thứ tự của token và trục y là giá trị độ đo tương tự. Tuy nhiên, do độ đo tương tự được tính giữa hai khối b_1 và b_2, trong đó b_1 bao gồm các chuỗi token từ $i-k$ đến i và b_2 bao gồm các chuỗi token từ $i+1$ đến $i+k+1$ nên độ đo sẽ rơi vào vị trí giữa chuỗi token i và $i+1$. Và trong thuật toán này, chúng ta sẽ sử dụng đồ thị khác đi với trục x là số thứ tự của điểm giữa của các chuỗi token. Đồ thị được làm trơn bằng kĩ thuật làm trơn trung bình. Trong thực nghiệm cho thấy, việc sử dụng kĩ thuật làm trơn trung bình với kích thước cửa sổ là 3 thích hợp với hầu hết các văn bản và chỉ cần sử dụng một vòng làm trơn.

Các vị trí biên được xác định thông qua sự thay đổi trong chuỗi các độ đo tương tự thu được ở bước trước. Số thứ tự của các điểm giữa của các chuỗi token không được sắp xếp theo giá trị độ đo tương tự ở đó mà lại được sắp xếp phụ thuộc vào mức độ dốc của đồ thị tại điểm đó so với các điểm xung quanh. Với một điểm giữa của chuỗi token *i*, thuật toán sẽ xem xét độ đo tương tự tại điểm giữa của chuỗi token bên trái của *i* miễn là giá trị của nó đang tăng. Khi giá trị so với bên trái đạt cực đại, sự sai khác về độ đo tương tự giữa độ đo tại điểm đạt cực đại và độ đo tại *i* được ghi lại. Công việc này được áp dụng tiếp tục với các điểm giữa của các chuỗi token phía bên phải của *i*, độ tương tự của các điểm đó sẽ được kiểm tra, miễn là chúng vẫn tiếp tục tăng. Độ cao tương đối của điểm cực đại so với bên phải của *i* được cộng với độ cao tương đối của điểm cực đại so với điểm bên trái (Một điểm giữa xuất hiện tại điểm cực đại sẽ có độ đo bằng 0 vì cả hai điểm bên cạnh đều không cao hơn nó). Độ đo mới này được gọi là độ

sâu, tương ứng với mức độ thay đổi xuất hiện ở hai phía của một điểm giữa của chuỗi token. Đường biên của các phân đoạn sẽ được ấn định cho các điểm giữa của các chuỗi token có độ đo tương ứng lớn nhất và sẽ được điều chỉnh để lấy được điểm ngăn cách thực sự giữa các đoạn. Một thủ tục kiểm tra sẽ được thực hiện để đảm bảo các phân đoạn không quá gần nhau. Theo thực nghiệm, nên có ít nhất 3 chuỗi token giữa 2 đường biên. Điều này sẽ giúp ngăn những văn bản có thông tin tiêu đề giả và các đoạn chỉ có một câu. Một ví dụ cho trường hợp này chính là trong văn bản có sẵn câu tiêu đề cho mỗi đoạn và thông thường câu đó được ngăn với đoạn văn bản tương ứng cũng bằng một dấu ngắt đoạn.

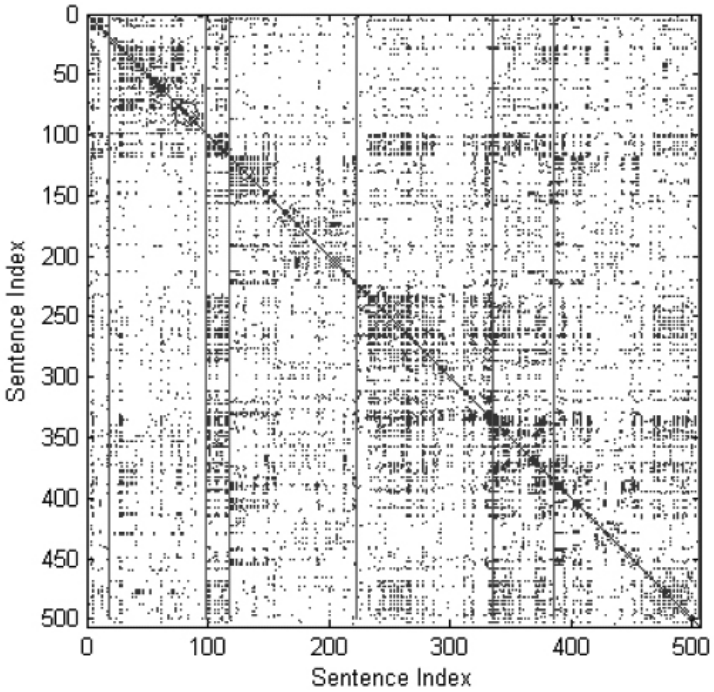
Thuật toán phải xác định có bao nhiêu phân đoạn (segment) sẽ được ấn định cho một văn bản vì mỗi đoạn (paragraph) cũng có thể là một đường biên tiềm năng. Không thể có một ngưỡng cố định cho trường hợp này vì nó phụ thuộc theo kiểu văn bản và độ dài văn bản.

Hearst đã đưa ra một phương pháp tham ăn cho phép xác định số lượng đường biên được ấn định phụ thuộc theo chiều dài văn bản và phụ thuộc theo các độ đo tương tự trong văn bản đó: giá trị ngưỡng là một hàm của giá trị trung bình và độ lệch chuẩn của độ sâu của văn bản sau khi được phân tích. Theo đó, một đường biên được ấn định chỉ khi độ sâu của nó vượt qua $\overline{s} - \sigma/2$, trong đó $\overline{s}$ là giá trị trung bình còn σ là độ lệch chuẩn.

Thuật toán của Hearst đã được triển khai thành công cụ TextTiling. Trong bài báo của mình, Hearst đã trình bày độ đo đánh giá giải thuật thông qua độ chính xác và độ hồi tưởng mà sau này, trong [Pevzner, Hearst], bà đã trình bày một độ đo khác *WindowDiff* là mở rộng của độ đo P_k do Beeferman đưa ra vào năm 1997 [Beeferman, Berger, Lafferty]. Trong phần cuối của chương này sẽ trình bày các độ đo phổ dụng cho bài toán phân đoạn văn bản.

Một thuật toán tương tự thuật toán của Hearst là thuật toán của Reynar [Reynar, 1994]. Thuật toán này cũng thực hiện bước phân tích hình thái để loại bỏ các từ dừng và từ phổ biến. Tuy nhiên ở bước sau, thuật toán này sử dụng các câu thay vì các câu giả với độ dài cố định như trong thuật toán của Hearst và sau đó tiến hành tính toán độ tương tự trên tất cả các cặp câu trong văn bản. Do đó thuật toán này còn được gọi là thuật toán tính độ tương tự toàn cục so với thuật toán của Hearst tính toán độ tương tự cục bộ. Tiếp đó sẽ dựng một đồ thị theo kĩ thuật *dotplotting* được trình bày trong [Church, 1993]. Hình 1 là ví dụ của đồ thị dotplotting, như ta thấy trên đồ thị, các vùng văn bản có độ tương tự cao sẽ đậm hơn (mật độ cao) và tập trung quanh đường chéo chính. Các đường kẻ dọc là vị trí phân đoạn thực tế trong văn bản. Và như quan sát thấy trên đồ thị, các điểm

phân các giữa các vùng có mật độ cao trên đường chéo chính hầu như trùng với điểm phân đoạn thực tế của văn bản.



Hình 1. Đồ thị dotplotting cho một văn bản

Như vậy có thể nói thuật toán này tương tự như thuật toán của Hearst, chỉ khác là thuật toán này sử dụng câu cú pháp thay vì câu giả và sử dụng kĩ thuật *dotplotting* để xác định điểm biên thay vì dùng các phương pháp giải tích như của Hearst. Theo [Regina, 2006], việc sử dụng chuỗi token có độ dài cố định hay thay đổi có tác dụng gần như nhau, sự khác biệt không đáng kể.

Ngoài ra, trong [Choi 2000], các tác giả đã trình bày một phương pháp tổng hợp dựa trên kĩ thuật của Hearst và kĩ thuật dotplotting cải tiến với việc áp dụng các phép toán xử lí ảnh (nhân chập với một ma trận vuông kích thước 3×3) để làm rõ nét hơn vị trí của các đường biên và qua đó tìm được chính xác hơn vị trí phân tách. Thuật toán này đã được triển khai thành công cụ C99 được sử dụng khá phổ biến. Trong phần thực nghiệm, luận văn sẽ sử dụng công cụ C99 là một trong hai công cụ để phân đoạn văn bản.

2.2.2. Sử dụng mô hình nhát cắt cực tiểu

Ngoài việc sử dụng các mối liên kết từ vựng, chúng ta còn có thể ứng dụng lí thuyết đồ thị để giải quyết bài toán phân đoạn văn bản. Tiêu biểu cho phương pháp này là mô hình nhát cắt cực tiểu được trình bày trong [Malioutov, Regina 2006]. Mô hình này sử dụng phép phân hoạch đồ thị thoả mãn điều kiện nhát cắt chuẩn hoá (normalized-cut criterion) [Shi, Malik 2000].

Trong khi các các tiếp cận trước đây sử dụng độ đo tương tự để phân đoạn thì trong mô hình này, các tác giả mô hình hoá đối tượng của bài toán thông qua các nhát cắt trên đồ thị. Mô hình này sẽ tìm cách cực đại độ tương tự trong mỗi phân đoạn và cực tiểu độ tương tự giữa các phân đoạn khác nhau.

Mô hình nhát cắt cực tiểu

Cho đồ thị $G = \{V, E\}$ là một đồ thị vô hướng có trọng số trong đó V là tập hợp các đỉnh tương ứng với các câu trong văn bản và E là tập hợp các cạnh có trọng số. Trọng số $w(u,v)$ định nghĩa độ tương tự giữa hai đỉnh u và v, trong đó trọng số cao hơn chỉ ra rằng độ tương tự cao hơn. Chi tiết về cách thức xây dựng đồ thị sẽ được trình bày ở phần xây dựng đồ thị.

Trước hết ta sẽ xem xét bài toán phân hoạch đồ thị thành hai tập hợp đỉnh A và B. Chúng ta sẽ phải làm cực tiểu giá trị của nhát cắt mà giá trị này được định nghĩa là tổng trọng số của các cạnh nối giữa hai tập hợp đỉnh. Hay nói cách khác, ta muốn chia các câu thành hai tập hợp có độ phân biệt đạt cực đại bằng cách chọn A và B để cực tiểu hoá giá trị nhát cắt:

$$cut(A,B) = \sum_{u \in A, v \in B} w(u,v)$$

Tuy nhiên, cần đảm bảo rằng không chỉ làm cực đại hoá sự khác nhau giữa hai tập hợp mà tự bản thân mỗi tập hợp phải cực đại hoá độ đo tương tự. Điều này được thực hiện thông qua nhát cắt chuẩn hoá, trong đó giá trị nhát cắt được chuẩn hoá thông qua giá trị của mỗi tập hợp tương ứng. Giá trị của mỗi tập hợp được tính bằng tổng trọng số của các cạnh nối từ bên trong tập hợp ra toàn đồ thị:

$$vol(A) = \sum_{u \in A, v \in V} w(u,v)$$

Điều kiện nhát cắt chuẩn hoá ($Ncut$) được định nghĩa như sau:

$$Ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

Thông qua việc cực tiểu hoá giá trị này, chúng ta sẽ vừa cực tiểu hoá được độ tương tự giữa các tập hợp lại vừa cực đại hoá độ tương tự bên trong mỗi tập hợp. Công thức này cũng cho phép chúng ta phân chia giá trị mục tiêu thành tổng của các số hạng riêng lẻ, và cho phép một giải pháp quy hoạch động cho bài toán nhát cắt nhiều đường (nhiều nhắt cắt tại một thời điểm).

Điều kiện này có thể dễ dàng được mở rộng cho trường hợp nhát cắt chuẩn hoá k-đường:

$$Ncut_k(V) = \frac{cut(A_1, V - A_1)}{vol(A_1)} + \cdots + \frac{cut(A_k, V - A_k)}{vol(A_k)}$$

trong đó $A_1, \ldots, A_k$ là một phân hoạch của đồ thị.

Trong [Shi, Malik] đã chỉ ra rằng bài toán cực tiểu nhát cắt chuẩn hoá trên đồ thị là bài toán NP đầy đủ. Tuy nhiên, trong bài toán này, nhát cắt đa đường bị ràng buộc duy trì tính tuyến tính của phép phân đoạn. Ràng buộc này có nghĩa là tất cả các đỉnh nằm giữa các đỉnh trái nhất và các đỉnh phải nhất của một phân hoạch cụ thể phải thuộc phân hoạch đó. Với ràng buộc này, các tác giả đã trình bày một thuật toán quy hoạch động để tìm chính xác nhát cắt chuẩn hoá đa đường cực tiểu trong thời gian đa thức:

$$C[i,k] = \min_{j<k}\left[C[i-1, j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]} \right]$$

$$B[i,k] = \arg\min_{j<k}\left[C[i-1, j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]} \right]$$

$$C[0,1] = 0, \quad C[0,k] = \infty, \quad 1 < k \le N$$

$$B[0,k] = 1, \quad 1 \le k \le N$$

$C[i,k]$ là giá trị nhát cắt chuẩn hoá của phân đoạn tối ưu của *k* câu đầu tiên vào phân đoạn *i*. Phân đoạn thứ *i*, $A_{j,k}$ bắt đầu từ đỉnh u_j và kết thức ở đỉnh u_k. $B[i,k]$ là bảng truy vết, theo đó ta sẽ tìm ra chuỗi đường biên phân đoạn tối ưu.

Độ phức tạp tính toán của thuật toán quy hoạch động ở trên là $O(KN^2)$, trong đó *K* là số lượng tập phân đoạn còn *N* là số lượng đỉnh trong đồ thị (số lượng câu trong văn bản).

Xây dựng đồ thị

Rõ ràng hiệu năng của thuật toán trên sẽ phụ thuộc rất nhiều vào cách thức biểu diễn của đồ thị, hàm đo độ tương tự giữa các cặp đỉnh và các tham số mô hình khác.

Trước tiên, ở giai đoạn tiền xử lí, các kĩ thuật xử lí văn bản được áp dụng. Các từ được rút gọn về dạng gốc (stemming) và các từ dừng (stop-word) bị loại bỏ.

Trong quá trình xây dựng đồ thị, do thuật toán thực hiện việc tính độ tương tự toàn cục tức là trên tất cả các cặp câu nên sẽ đưa ra một đồ thị đầy đủ. Điều này gây bất lợi lớn cho quá trình tính toán, đồng thời cũng sẽ làm giảm độ chính xác của việc phân đoạn, do đó một tham số ngưỡng sẽ được đưa vào để loại bỏ bớt các cạnh có trọng số quá thấp.

Trong quá trình tính toán độ tương tự, các câu sẽ được biểu diễn dưới dạng vectơ tần suất của các từ. Độ đo tương tự thường dùng là độ đo cosin của Hearst đã được giới thiệu ở phần trước. Trong phần này, để tránh vấn đề độ chính xác số học khi tính tổng một chuỗi các trọng số rất nhỏ, chúng ta sử dụng độ tương tự làm mũ giữa các vectơ của các câu:

$$w\left(s_i, s_j\right) = e^{\frac{s_i \square s_j}{\|s_i\| \times \|s_j\|}}$$

Ngoài ra, thuật toán còn sử dụng phép làm mịn độ đo tương tự. Khi so sánh hai câu, thuật toán xem xét độ tương tự giữa các láng giềng trực tiếp. Việc làm mịn đạt được bằng cách cộng số lượng từ xuất hiện trong các câu liền kề vào vectơ đặc trưng của câu hiện tại. Số lượng này được tính toán phù hợp với khoảng cách của chúng từ câu hiện tại:

$$\tilde{s}_i = \sum_{j=i}^{i+k} e^{-\alpha(j-i)} s_j$$

trong đó s_i là vectơ tần suất các từ và α là một tham số điều khiển độ mịn.

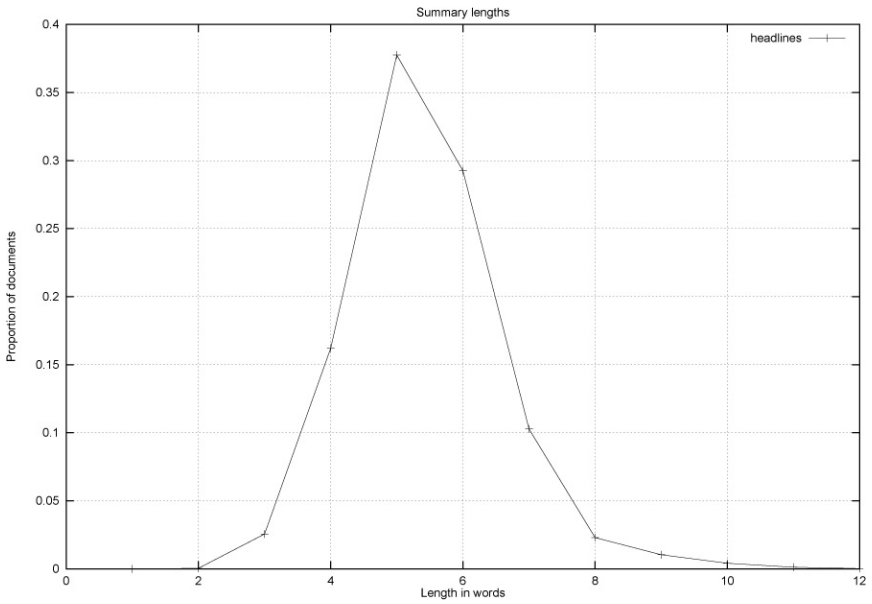
Trong các công thức ở trên, các câu chính là các đỉnh. Tuy nhiên, chúng ta có thể biểu diễn các đỉnh của đồ thị là các chuỗi từ cố độ dài cố định và không giao nhau do trong một số trường hợp việc xác định ranh giới của câu là rất khó khăn (văn bản nói). Khi đó kích thước của chuỗi có thể chọn như trong thuật toán của Hearst ở phần trước.

Ngoài các yếu tố trên, việc chọn trọng số cho các từ cũng có ảnh hưởng rất lớn tới chất lượng của việc phân đoạn [Ji, Sha 2003; Choi 2001]. Thuật toán của Hearst ở trên sử dụng tần suất của từ (*TF*) làm trọng số. Tuy nhiên thực tế cho thấy có khá nhiều từ phổ biến trong văn bản mà không có ý nghĩa mấy đến việc phân tách chủ đề. Ví dụ trong tài liệu nói về "Support Vector Machines" thì cụm từ SVM sẽ xuất hiện rất nhiều trong văn bản và nó không có ý nghĩa trong

việc phân đoạn văn bản. Do đó để giải quyết vấn đề này, trong thuật toán này, các tác giả sử dụng độ đo $TF * IDF$ để loại bỏ những từ quá phổ biến.

2.3. Sinh tiêu đề cho văn bản

So với toàn bộ văn bản, tiêu đề sẽ biểu diễn ngắn gọn thông tin trong văn bản và do đó giúp người đọc nhanh chóng nắm bắt được đại ý của toàn văn bản. Tự động sinh tiêu đề cho văn bản là một bài toán phức tạp, nó không chỉ đòi hỏi lựa chọn những từ có khả năng xuất hiện trong tiêu đề mà còn phải được sắp xếp theo một thứ tự phù hợp, đúng thứ tự và dễ hiểu. Bài toán này có nhiều khác biệt so với bài toán tóm tắt văn bản thông thường. Ở bài toán tóm tắt văn bản thông thường, độ dài của đoạn tóm tắt thường là 50, 100, 200 hay 400 từ (theo chuẩn của DUC), nhưng với bài toán sinh tiêu đề thì độ dài đó chỉ là từ 1 đến 12 từ [Banko] (Hình 2). Cũng vì lí do độ dài ngắn như vậy cho nên trong bài toán này, người ta thường dùng các phương pháp trích chọn ra các từ hoặc cụm từ mang ý nghĩa chính trong văn bản mà cụ thể là các danh từ/cụm danh từ hoặc động từ/cụm động từ [Roxana 2002].



Hình 2. Phân bố độ dài tiêu đề văn bản theo Reuters-1997

Hiện nay, phương pháp sinh tiêu đề cho văn bản được chia ra làm hai hướng chính:

- Sinh tiêu đề cho văn bản dựa trên việc trích chọn ra một từ/cụm từ "đặc trưng" nhất cho văn bản. Với phương pháp này thì độ dài của tiêu đề thường rất ngắn (chỉ từ 1 đến 3 từ) nhưng về mặt cú pháp thì luôn đảm bảo. Hơn nữa phương pháp này thường là học không giám sát cho

nên rất thích hợp với các trường hợp không có dữ liệu huấn luyện. [Roxana].

- Sinh tiêu đề cho văn bản được chia làm hai bước, bước thứ nhất sẽ là chọn ra các từ/cụm từ mang ý nghĩa chính trong văn bản. Bước thứ hai sẽ là sắp xếp các cụm từ để mang đúng cú pháp và dễ hiểu nhất. [Witbrock, Branavan].

Trong phần tiếp theo, luận văn sẽ lần lượt giới thiệu hai thuật toán điển hình đại diện cho hai phương pháp trên.

2.4. Các phương pháp sinh tiêu đề cho văn bản

2.4.1. Phương pháp trích chọn cụm từ

Phương pháp trích chọn cụm từ sẽ tiến hành phân tích các câu trong văn bản để tìm ra từ/cụm từ mang ý nghĩa tiêu biểu cho văn bản. Phương pháp này thường dựa vào các đặc trưng như: vị trí của cụm từ và sự phổ biến của cụm từ đó trong văn bản.

Trong [Roxana, 2002], các tác giả đã phân tích và sử dụng cụm danh từ để làm tiêu đề cho từng đoạn văn bản. Theo đó, phương pháp này bao gồm các bước sau:

- Phân đoạn văn bản thành các câu rời rạc.
- Gán nhãn từ loại cho các từ trong câu (POS Tagging).
- Tìm các danh từ/cụm danh từ trong câu.
- Tìm ra câu quan trọng nhất trong văn bản.
- Tìm ra chủ đề của câu quan trọng nhất ở bước trên và coi đó là tiêu đề của đoạn văn bản.

Trong phương pháp này, các tác giả có đưa ra khái niệm chủ đề của một câu. Chủ đề của một câu được định nghĩa là cụm danh từ mang ý nghĩa quan trọng nhất trong câu đó, thông thường được xác định theo "kinh nghiệm" (heuristic) đối với các ngôn ngữ tuân theo thứ tục SVO. Nếu câu không có cụm danh từ thì câu đó không được coi là câu quan trọng nhất trong văn bản. Cách tiếp cận để tìm ra câu quan trọng nhất trong văn bản là sử dụng độ đo cosin giữa các câu làm trọng số cho một đồ thị mà các đỉnh chính là các câu. Câu quan trọng nhất sẽ là câu tương ứng với đỉnh có tổng trọng số của các cạnh nối với đỉnh đó là cao nhất.

Phương pháp này tỏ ra khá hiệu quả và đã thực sự đạt được kết quả cao trong DUC 2002 và đây cũng là phương pháp luận văn lựa chọn để làm thực nghiệm do nó còn có thể áp dụng để phân đoạn văn bản. Ngoài ra phương pháp này không đòi hỏi dữ liệu có sẵn để huấn luyện nên sẽ đặc biệt thích hợp với sự khó khăn trong việc tìm kiếm và chuẩn bị dữ liệu trong nước.

2.4.2. Phương pháp hai pha

Trong phương pháp này, việc sinh tiêu đề cho văn bản được chia làm hai pha [Witbrock 1999, Hauptmann 2000-2001]:

- Pha 1: Chọn ra các từ có trọng số cao nhất trong văn bản và coi đó là các từ có ý nghĩa nhất trong văn bản. Các trọng số này thông thường được tính theo TF * IDF mà trong trường hợp này thì là TF do chỉ có một văn bản/đoạn văn bản.
- Pha 2: Các từ được chọn sẽ được sắp xếp lại theo các thức hợp lí nhất. Có 2 cách sắp xếp: cách thứ nhất dựa trên thứ tự nội tại trong văn bản; cách thứ hai là dựa trên thống kê sử dụng mô hình n-gram.

Tuy nhiên phương pháp này tồn tại 2 vấn đề cơ bản liên quan đến cả 2 pha ở trên:

- Pha 1: Các từ loại như giới từ, tính từ, mạo từ thường không mang mấy ý nghĩa trong việc chỉ ra ý chính của văn bản. Do đó các từ này thường phải bị loại đi. Để giải quyết vấn đề này thì ta có thể loại bỏ từ dừng, sử dụng nhãn từ loại để chỉ giữ lại danh từ, động từ hoặc cụm danh từ, cụm động từ.
- Pha 2: Nếu sử dụng cách sắp xếp dựa trên thứ tự nội tại trong văn bản thì một vấn đề rất dễ nhận ra là cú pháp của tiêu đề được sinh ra sẽ không được đảm bảo và tất nhiên là sẽ gây hiểu sai nghĩa của văn bản. Còn nếu sử dụng mô hình thống kê để tính xác suất xuất hiện của từ/cụm từ theo mô hình n-gram thì sẽ chỉ chọn được các từ tương đối phổ biến trong các tiêu đề có sẵn để làm tiêu đề mới, còn đối với các tiêu đề hiếm như văn bản nói về một căn bệnh mới với những thuật ngữ mới thì xác suất xuất hiện cho các từ đó sẽ bằng 0 và do đó sẽ không bao giờ được chọn vào tiêu đề của văn bản.

Phương pháp hai pha tỏ ra có hiệu quả hơn trong việc sinh tiêu đề cho văn bản, tuy nhiên vấn đề gặp phải trong pha thứ hai hiện vẫn chưa có một phương pháp để giải quyết triệt để.

2.5. Tóm tắt chương hai

Trong chương này, luận văn đã trình bày hai bước cơ bản để xây dựng mục lục cho một văn bản bao gồm phân đoạn văn bản và sinh tiêu đề cho các đoạn văn bản. Với mỗi bước, luận văn đã đi vào phân tích một số phương pháp và thuật toán tiêu biểu đồng thời chỉ ra điểm mạnh và điểm yếu của từng phương pháp. Các đề xuất cải tiến và cơ sở của các cải tiến sẽ được trình bày ở trong chương cuối. Trong chương tiếp theo, luận văn sẽ tiến hành phân tích cơ sở để tích hợp hai bước này để tạo ra một mục lục có tính hợp lí cao và các phương pháp đánh giá đối với từng bước.

Chương 3

XÂY DỰNG MỤC LỤC CHO VĂN BẢN

3.1. Mô hình tích hợp thuật toán

Như đã phân tích ở chương 1, bài toán xây dựng mục lục cho văn bản là một bài toán tóm tắt văn bản loại chỉ dẫn, theo đó trong "tóm tắt" sẽ có thông tin ngắn gọn cho từng đoạn văn bản và vị trí của đoạn văn bản tương ứng. Để có thể giải quyết bài toán này thì luận văn chọn hướng tiếp cận chia bài toán ra làm hai bài toán con là bài toán phân đoạn văn bản và bài toán sinh tiêu đề cho đoạn văn bản. Các bài toán này đã lần lượt được trình bày trong chương 2.

Về mặt nguyên tắc thì hai bài toán này có thể được giải quyết một cách độc lập, theo đó, sau khi văn bản được phân thành các đoạn độc lập với nhau thì ta sẽ áp dụng thuật toán sinh tiêu đề cho từng đoạn một. Tuy nhiên điều này sẽ gây lãng phí những thông tin đã thu thập được ở bước phân đoạn văn bản đồng thời có thể sẽ tạo ra những tiêu đề giống nhau.

Để giải quyết vấn đề trên, luận văn đề xuất một phương pháp để có thể sử dụng lại các đặc trưng đã thu thập được ở bước phân đoạn văn bản và sử dụng cho bước tiếp theo. Cơ sở của đề xuất này dựa trên nhận xét là khi ta phân đoạn văn bản thì đã dựa trên sự thay đổi chủ đề của các đoạn văn bản, điều đó có nghĩa là tiêu đề của văn bản đã ít nhiều được xác định tuy còn ở dạng "ẩn". Các đặc trưng được sử dụng ở đây là các đặc trưng về từ vựng. Cụ thể như sau:

- Tại bước phân đoạn văn bản, thay vì sử dụng tất cả các từ có trong mỗi câu, ta chỉ sử dụng các cụm danh từ, cụm động từ và do đó chuỗi từ vựng cho từng câu sẽ là các từ trong cụm danh từ và cụm động từ của câu đó.
- Với các chuỗi từ vựng (các vectơ biểu diễn câu) như trên, ta sẽ xác định được câu quan trọng nhất trong văn bản dựa trên đồ thị được xây dựng như mô tả như sau:
    - Mỗi đỉnh tương ứng với một chuỗi từ vựng.
    - Trọng số của các cạnh nối giữa các đỉnh là độ đo tương tự (cosin) giữa các chuỗi từ vựng tương ứng.
    - Trọng số của một đỉnh là tổng trọng số các cạnh liên kết với đỉnh đó.

Câu chủ đề là câu có chuỗi từ vựng tương ứng với đỉnh có trọng số cao nhất trong đồ thị.

- Đến đây, thuật toán được chia làm các hướng:
  - Sử dụng một thuật toán tinh giản câu (sentence compression) đối với câu chủ đề để thu được tiêu đề của văn bản. Phương pháp này được sử dụng trong công cụ thương mại của hãng BBN được nêu trong [Dorr 2003]. Thuật toán tinh giản câu sẽ thu được một câu chỉ còn cụm danh từ và cụm động từ.
  - Tìm chủ đề của câu chủ đề để làm tiêu đề của văn bản [Roxana 2002]. Chủ đề của câu được xác định là cụm danh từ chính trong câu. Cách xác định cụm danh từ chính được nêu trong [Givón 2001] và sử dụng trong bộ công cụ SUMMA của Roxana.

Trong luận văn này, tôi sử dụng phương pháp phân đoạn văn bản dựa trên chuỗi từ vựng [Hearst 1994] kết hợp với phương pháp sinh tiêu đề dựa trên chủ đề của câu chủ đề [Roxana 2002].

3.2. Đảm bảo tính hợp lí của mục lục

Như đã trình bày ở phần trước, trong mục lục chúng ta sẽ đưa ra tiêu đề và vị trí của các đoạn văn bản tương ứng. Tiêu đề này sẽ là cụm từ ngắn gọn mô tả mục đích chính của toàn đoạn văn. Tuy nhiên có rất nhiều trường hợp mà mục lục sinh ra sẽ có sự trùng lặp giữa các phân đoạn khác nhau, nghĩa là tiêu đề giống nhau cho hai đoạn văn bản khác nhau [Branavan 2007]. Hơn nữa, suy luận một cách "kinh nghiệm" cho thấy rằng, đối với các mục lục đa cấp thì tiêu đề của các mục con phải có sự liên hệ nào đó với mục cha và liên hệ đó có thể là một quan hệ toàn thể - bộ phận trên một ontology cho lĩnh vực tương ứng với văn bản.

Để giải quyết vấn đề này, luận văn đề xuất một thuật toán trung gian để sinh mục lục dựa trên thuật toán được nêu trong [Branavan 2007]. Cụ thể như sau:

- Mỗi đoạn văn bản thay vì đưa ra chỉ một tiêu đề thì sẽ đưa ra một danh sách *k* tiêu đề và được sắp xếp có thứ tự theo mức độ quan trọng của nó trong đoạn văn bản.
- Đối với phân đoạn tuyến tính, ta tính trọng số của mục lục bằng tổng trọng số của các tiêu đề thành viên. Sử dụng thuật toán đệ quy có nhánh cận duyệt qua tất cả các phương án xây dựng mục lục để tìm ra

phương án có tổng trọng số cao nhất mà không có sự trùng lặp tiêu đề giữa hai phân đoạn bất kì.

- Đối với phân đoạn đa cấp, cần đảm bảo hơn ½ số tiêu đề trong cấp con sẽ có ít nhất một từ có quan hệ toàn thể - bộ phận với ít nhất một từ trong tiêu đề của cấp cha.

Trên đây là một số đề xuất để đảm bảo tính hợp lí của mục lục được hình thành trong quá trình tìm hiểu dữ liệu của luận văn. Trên thực tế, luận văn mới dừng lại ở việc triển khai tránh sự trùng lặp giữa hai tiêu đề trong phân đoạn tuyến tính. Còn việc xử lí đối với phân đoạn đa cấp sẽ là hướng phát triển tiếp theo của luận văn.

3.3. Các phương pháp đánh giá

Hiện nay, vẫn chưa có một phương pháp đánh giá cụ thể cho bài toán xây dựng mục lục cho văn bản do đây là một bài toán mới. Trong luận văn này, tôi áp dụng các phương pháp đánh giá có sẵn cho từng bước của thuật toán. Đó là đánh giá cho bước phân đoạn văn bản và đánh giá cho việc sinh tiêu đề. Tuy nhiên, đối với bài toán tóm tắt văn bản nói chung và bài toán sinh tiêu đề nói riêng, người ta vẫn áp dụng một phương pháp phổ biến là dựa vào sự đánh giá của các chuyên gia ngôn ngữ. Vì thực tế cho thấy với mỗi một văn bản, tuỳ văn phong của từng người mà sẽ có cách tóm tắt khác nhau. Hơn nữa, hiện nay không tồn tại một phương pháp hiểu văn bản đủ hiệu quả để đánh giá xem một đoạn tóm tắt có phải là thực sự tốt hay không. Do vậy, trong luận văn này, ngoài việc trình bày các kết quả thực nghiệm và đánh giá thông qua các độ đo, tôi còn phân tích dựa trên ý kiến chuyên gia về sự phù hợp của tiêu đề đối với đoạn văn bản.

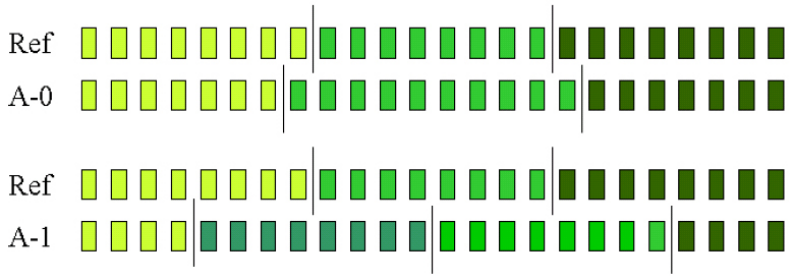
3.3.1. Đánh giá thuật toán phân đoạn

Trong bài báo năm 1994, Hearst sử dụng hai độ đo phổ biến trong học máy để đánh giá thuật toán là *độ chính xác* (precision) và *độ hồi tưởng* (recall) được định nghĩa như sau:

- Độ chính xác là tỉ lệ số đường biên mà mô hình chọn chính xác trên tổng số các đường biên được mô hình xác định trong văn bản.
- Độ hồi tưởng là tỉ lệ số đường biên mà mô hình chọn chính xác trên tổng số đường biên thực của văn bản.

Hai độ đo này cũng được sử dụng khá nhiều trong các công trình khác, tuy nhiên hai độ đo này gặp phải hai vấn đề lớn:

- Sự tác động qua lại vốn có của hai độ đo này, nghĩa là khi một độ đo tăng lên sẽ có khuynh hướng làm giảm độ đo còn lại. Ví dụ, khi ta đặt thêm nhiều đường biên hơn thì sẽ làm tăng độ hồi tưởng trong khi độ chính xác lại giảm đi. Một số công trình khác sử dụng độ đo F [Baeza, 1999] hoặc sử dụng đồ thị biểu diễn độ chính xác tương ứng với các mức khác nhau của độ hồi tưởng.
- Một vấn đề khác là hai độ đo này không "nhạy" với các trường hợp phân đoạn gần chính xác. Ví dụ, Hình 3 biểu diễn kết quả của 2 thuật toán phân đoạn khác nhau so với phân đoạn gốc của văn bản. Trong cả hai trường hợp, các thuật toán đều đoán sai vị trí đường biên, và do đó độ chính xác và độ hồi tưởng đều cho giá trị 0. Tuy nhiên, thuật toán A-0 cho kết quả gần chính xác (các đường biên dự đoán chỉ sai khác 1 đoạn so với thực tế), trong khi đó thuật toán A-1 cho kết quả sai hoàn toàn (thêm một phân đoạn, vị trí các đường biên cũng cách khá xa so với thực tế). Do đó trong trường hợp này, độ chính xác và độ hồi tưởng không thể chỉ ra được thuật toán A-0 tốt hơn thuật toán A-1 và do đó ta cần một phép đánh giá "nhạy" hơn để có thể giải quyết được vấn đề này.



Hình 3. Ví dụ đánh giá thuật toán phân đoạn

Độ đo P_k

Độ đo P_k được đề xuất lần đầu trong [Beeferman 97] và được áp dụng vào bài toán phân đoạn văn bản trong [Beeferman 99]. Độ đo này xem xét cả khoảng cách giữa đường biên do thuật toán xác định và đường biên thực tế. Mục đích của độ đo này là đo tỉ lệ lỗi của thuật toán. Do đó độ đo này càng nhỏ thì thuật toán càng chính xác.

Độ đo P_k xuất phát từ độ đo P_D được định nghĩa như sau:

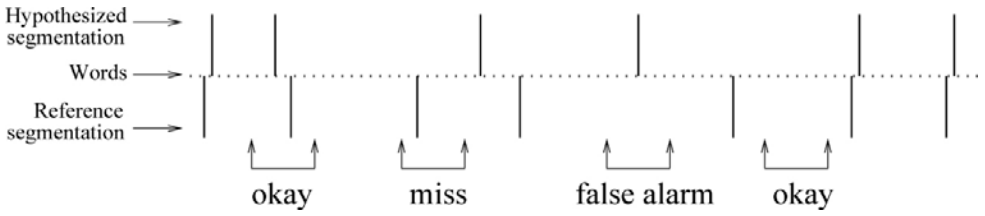
$$P_D(ref, hyp) = \sum_{1 \le i \le j \le N} D(i,j)\left[\delta_{ref}(i,j) \oplus \delta_{hyp}(i,j)\right]$$

trong đó:

- *ref* và *hyp* là các phân đoạn thực tế và phân đoạn do thuật toán sinh ra;
- *N* là số lượng câu;
- $\oplus$ là phép toán logic XNOR (cho giá trị 1 khi 2 số hạng giống nhau);
- $\delta_X(i,j)$ là hàm cho giá trị 1 nếu câu *i* và câu *j* nằm trong cùng phân đoạn và cho giá trị 0 nếu khác phân đoạn;
- $D(i,j)$ là phân phối xác suất khoảng cách trên một tập các khoảng cách có thể giữa các cặp câu chọn ngẫu nhiên.

Khi áp dụng thực tế vào bài toán này, khoảng cách $D(i,j)$ được cố định là *k*, thường được lấy là độ dài trung bình của các phân đoạn trong văn bản gốc tính theo câu. Khi đó độ đo được gọi là P_k được định nghĩa theo hình thức khác, là sự kết hợp của 2 xác suất có điều kiện gọi là xác suất *miss* và *false alarm* được xác định như trong Hình 4:

$$p(error \mid \text{ref}, \text{hyp}, k) =$$
$$p(miss \mid \text{ref}, \text{hyp}, different\ \text{ref}\ segments, k) \times p(different\ \text{ref}\ segments \mid \text{ref}, k) +$$
$$p(false\ alarm \mid \text{ref}, \text{hyp}, same\ \text{ref}\ segment, k) \times p(same\ \text{ref}\ segment \mid \text{ref}, k)$$



Hình 4. Cách xác định tham số cho độ đo P_k

Tuy nhiên độ đo P_k có một số nhược điểm sau:

- *miss* bị tính nhiều hơn *false alarm*.
- Khi một đường biên được thêm vào và tạo ra một phân đoạn có kích thước nhỏ hơn *k* thì nó không bị tính và độ đo.
- Khi kích thước của các phân đoạn có sự biến đổi mạnh thì thuật toán không bị "phạt" nhiều.
- Các lỗi xác định biên gần chính xác vẫn bị tính quá nhiều.
- Độ đo thực sự không mang tính độ đo theo phần trăm mà chỉ là một độ đo có giá trị trong khoảng 0 đến 1.

Độ đo WindowDiff

Trong [Hearst 2002] đề xuất một độ đo mới cho bài toán phân đoạn văn bản gọi là WindowDiff, đây là một sự mở rộng của độ đo P_k. Trong độ đo này, phép toán $\oplus$ được thay thế bằng sự khác nhau giữa số lượng đường biên giữa 2 vị trí i và $i+k$ trong cả *ref* và *hyp*. Nếu không có sự sai khác thì các vị trí i và $i+k$ nằm trong cùng phân đoạn của *ref* và *hyp*. Ý nghĩa của nó là giải quyết vấn đề khi có một phân đoạn nhỏ được thêm vào trong *hyp* mà P_k không giải quyết được.

$$WindowDiff\left(ref, hyp\right) = \frac{1}{N-k}\sum\left(\left|b\left(ref_i, ref_{i+k}\right) - b\left(hyp_i, hyp_{i+k}\right)\right|\right)$$

trong đó $b\left(x_i, x_j\right)$ biểu diễn số lượng đường biên giữa 2 vị trí i và j trong văn bản x và N là số lượng câu trong văn bản.

Các kết quả thực nghiệm trong [Hearst 2002] cho thấy độ đo này tương đối ổn định khi kích thước của phân đoạn biến đổi và tạo ra được sự cân bằng giữa *miss* và *false alarm*. Tuy nhiên, độ đo này có thể cho kết quả lớn hơn 1 nên không còn là độ đo theo phần trăm nữa. Do đó, độ đo này chỉ dùng để so sánh giữa các thuật toán mà không thể dùng để đánh giá trực tiếp chất lượng của thuật toán.

3.3.2. Đánh giá thuật toán sinh tiêu đề

Hiện nay, cách đánh giá thuật toán sinh tiêu đề phổ biến là so sánh tiêu đề sinh ra tự động với tiêu đề có sẵn của văn bản mẫu. Việc so sánh được thực hiện sau khi cả hai tiêu đề đã trải qua các bước tiền xử lý bao gồm:

- Loại bỏ từ dừng.
- Đưa từ về từ gốc (stemming).

Sau đó, việc đánh giá thuật toán sinh tiêu đề sẽ được tính dựa trên ba độ đo phổ biến trong lĩnh vực học máy là *độ chính xác P* (precision), *độ hồi tưởng R* (recall) và độ đo F_1 (F_β với $\beta = 1$).

Nếu gọi tập hợp các từ trong tiêu đề do thuật toán sinh ra là T_{gen} và tập hợp các từ trong tiêu đề gốc do con tác giả đặt là T_{org} thì các độ đo được tính lần lượt như sau:

$$P = \frac{\left|T_{gen} \cap T_{org}\right|}{\left|T_{gen}\right|}; \qquad R = \frac{\left|T_{gen} \cap T_{org}\right|}{\left|T_{org}\right|}; \qquad F_1 = \frac{2 \times P \times R}{P + R}$$

Trong đó $\left|T\right|$ kí hiệu lực lượng của tập hợp T hay số phần tử của tập hợp T.

3.4. Tóm tắt chương ba

Trong chương này, luận văn đã trình bày các đề xuất để tích hợp hai bước phân đoạn văn bản và sinh tiêu đề cho văn bản trong quá trình xây dựng mục lục cho văn bản nhằm tránh dư thừa tài nguyên. Tiếp đó luận văn cũng đưa ra các đề xuất cụ thể về việc đảm bảo tính hợp lí của một mục lục được sinh ra dựa trên các yếu tố "kinh nghiệm" thông qua việc tham khảo mục lục của các tài liệu sẵn có. Cuối cùng, luận văn trình bày các phương pháp đánh giá thông dụng sử dụng cho hai quá trình phân đoạn văn bản và sinh tiêu đề cho văn bản. Trong chương tiếp theo, luận văn sẽ trình bày các thử nghiệm trên một văn bản khoa học cụ thể để chứng minh tính khả thi và triển vọng của bài toán xây dựng mục lục cho văn bản.

Chương 4

THỬ NGHIỆM VÀ ĐÁNH GIÁ

4.1. Môi trường thử nghiệm

Quá trình thử nghiệm của luận văn được thực hiện trên hệ thống máy chủ IBM x3800 Series được cài đặt hệ điều hành Debian 4.0 (Etch) với các phiên bản phần mềm Java 1.6.0_01, GCC 4.1.2.

Các công cụ phần mềm được sử dụng trong quá trình thử nghiệm được liệt kê và mô tả trong Bảng 2.

Bảng 2. Danh sách các công cụ phần mềm sử dụng để thử nghiệm

STT	Tên phần mềm	Mô tả
1	**LT CHUNK**	**Tác giả:** Edinburgh Language Technology Group **Site:** http://www.ltg.ed.ac.uk/ **Công dụng:** Thực hiện gán nhãn từ loại cho các từ trong một văn bản, đồng thời nhận diện ra các cụm danh từ và cụm động từ.
2	**TextTiling**	**Tác giả:** Marti Hearst **Site:** http://people.ischool.berkeley.edu/~hearst/ **Công dụng:** Thực hiện phân đoạn văn bản sử dụng mối liên kết từ vựng. Kết quả là một văn bản được phân đoạn tuyến tính.
3	**C99**	**Tác giả:** Freddy Choi **Site:** http://www.lingware.co.uk/homepage/freddy.choi/ **Công dụng:** Phân đoạn văn bản sử dụng mối liên kết từ vựng kết hợp với đồ thị dotplotting. Kết quả là một văn bản được phân đoạn tuyến tính.
4	**SUMMA**	**Tác giả:** Marie-Francine Moens **Site:** http://www.cs.kuleuven.be/~liir/ **Công dụng:** Sinh tiêu đề cho một đoạn văn bản dựa trên khái niệm chủ đề của câu.

Trong quá trình thực hiện thử nghiệm, tôi có tiến hành chỉnh sửa công cụ SUMMA để phù hợp hơn với bài toán xây dựng mục lục cho văn bản và thử nghiệm các đề xuất đã nêu trong chương 3 về đảm bảo tính hợp lí của mục lục.

4.2. Dữ liệu thử nghiệm

Văn bản được sử dụng để thử nghiệm là bài báo "Generic Topic Segmentation of Document Texts" [Moens 2001]. Bài báo này bao gồm 1353 từ với 63 câu được chia làm 5 mục lớn (không tính các phần tiêu đề, tóm tắt, lời cảm ơn và tài liệu tham khảo). Chi tiết về các phần được trình bày trong Bảng 3.

Bảng 3. Cấu trúc văn bản thử nghiệm

Mục	Tiêu đề	Câu bắt đầu	Câu kết thúc	Mô tả
1	**Introduction**	1	6	Giới thiệu về bài toán phân đoạn văn bản.
2	**Research problem**	7	16	Giới thiệu vấn đề cần nghiên cứu và nhiệm vụ của bài báo.
3	**Methods**	17	52	Trình bày các phương pháp sử dụng trong quá trình phân đoạn văn bản.
3.1	Content terms and their distribution	17	22	Trình bày vấn đề về các khái niệm và sự phân bố của nó ảnh hưởng tới phân đoạn văn bản.
3.2	Lexical chains	23	29	Mô tả về phương pháp sử dụng chuỗi từ vựng để phân đoạn văn bản.
3.3	Topic segmentation	30	45	Trình bày các bước trong thuật toán phân đoạn văn bản mà bài báo trình bày, sử dụng sự phân bố của các khái niệm và chuỗi từ vựng.
3.4	Test corpora	46	52	Trình bày về tập dữ liệu thử nghiệm và sự khó khăn trong việc đánh giá mô hình.
4	**Related research**	53	60	Giới thiệu một số thuật toán phân đoạn văn bản khác và khiếm khuyết của các thuật toán đó.
5	**Conclusions**	61	63	Kết luận về bài báo: đóng góp và hướng phát triển.

Văn bản này được chia thành 5 mục lớn với mục số 3 được chia làm 4 mục con, do đó, với cách phân đoạn tuyến tính ta có thể coi văn bản được chia làm 8 mục. Trong phần này, luận văn sẽ chỉ giới hạn thử nghiệm bằng phương pháp phân đoạn tuyến tính.

Trong quá trình loại bỏ từ dừng, luận văn sử dụng tập từ dừng trong công cụ TextTiling của MartiHearst có sửa đổi để thêm nhiều từ dừng hơn. Danh sách các từ dừng được sử dụng được liệt kê trong Bảng 4.

Bảng 4. Danh sách từ dừng

said n't 'm a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co computer con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fify fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two un under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves

Trong quá trình gán nhãn từ loại sử dụng công cụ LT CHUNK, tập các nhãn từ loại được sử dụng là tập nhãn thu gọn được kế thừa từ tập nhãn Penn Treebank (http://www.cis.upenn.edu/~treebank/). Danh sách các nhãn cùng mô tả được trình bày trong Bảng 5 và Bảng 6.

Bảng 5. Tập nhãn từ loại (tập mở)

Nhãn từ loại	Mô tả	Ví dụ
JJ	adjective	green
JJR	adjective, comparative	greener

JJS	adjective, superlative	greenest
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
NN	common noun	table
NNS	noun plural	tables
NNP	proper noun	John
NNPS	plural proper noun	Vikings
VB	verb base form	take
VBD	verb past	took
VBG	gerund	taking
VBN	past participle	taken
VBP	verb, present, non-3d	take
VBZ	verb present, 3d person	takes
FW	foreign word	d'hoevre

Bảng 6. Tập nhãn từ loại (tập đóng)

Nhãn từ loại	Mô tả	Ví dụ
CD	cardinal number	1, third
CC	coordinating conjunction	and
DT	determiner	the
EX	existential there	there is
IN	preposition	in, of, like
LS	list marker	1)
MD	modal	could, will
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RP	particle	give up

TO	to (both "to go" and "to him")	to go, to him
UH	interjection	uhhuhhuhh
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-adverb	where, when

4.3. Quá trình thử nghiệm

Quá trình thử nghiệm được chia làm hai giai đoạn:

- Giai đoạn 1: Phân đoạn văn bản sử dụng các công cụ TextTiling và C99 để thu được các đoạn văn bản.
- Giai đoạn 2: Sinh tiêu đề cho từng đoạn văn bản bằng công cụ SUMMA.

Việc sinh tiêu đề cho văn bản được thực hiện cho cả các đoạn văn bản được phân đoạn tự động và các đoạn văn bản được phân sẵn của văn bản gốc.

4.4. Kết quả thử nghiệm

4.4.1. Kết quả phân đoạn văn bản

Kết quả phân đoạn văn bản được trình bày trong Bảng 7 và được biểu diễn trực quan hơn trong Hình 5. Trong Bảng 7 có 3 cột chính, mỗi cột tương ứng với từng cách phân đoạn, trong mỗi cột có 3 cột con gồm số thứ tự của đoạn văn bản, số thứ tự của câu đầu tiên và số thứ tự của câu cuối cùng của văn bản đó.

Bảng 7. Kết quả phân đoạn văn bản

Văn bản gốc			**C99**			**TextTiling**		
STT	**Đầu**	**Cuối**	**STT**	**Đầu**	**Cuối**	**STT**	**Đầu**	**Cuối**
1	1	6	1	1	6	1	1	7
2	7	16	2	7	17	2	8	13
3	17	22	3	18	19	3	14	20
4	23	29	4	20	28	4	21	28
5	30	45	5	29	35	5	29	34
			6	36	45	6	35	45

6	46	52	7	46	51	7	46	52
7	53	60	8	52	58			
8	61	63	9	59	63	8	53	63

Trong Hình 5, dòng đầu tiên là mô hình phân đoạn có sẵn, dòng thứ 2 là mô hình phân đoạn do công cụ C99 sinh ra và dòng thứ 3 là mô hình phân đoạn do công cụ TextTiling sinh ra. Các dấu "-" thể hiện cho các câu, các dấu "." thể hiện vị trí giữa các câu trong cùng một đoạn, còn các dấu "|" thể hiện đường biên phân tách giữa các đoạn.

|-.-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-|-.-.-|

|-.-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-.-.-|-.-|-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-|

|-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-.-.-.-.-|-.-.-.-.-.-|-.-.-.-.-.-.-.-.-.-|

Hình 5. Kết quả phân đoạn văn bản

4.4.2. Kết quả sinh tiêu đề

Các kết quả sinh tiêu đề cho văn bản lần lượt được trình bày trong Bảng 8, Bảng 9 và Bảng 10.

Bảng 8. Sinh tiêu đề cho phân đoạn gốc

Phần	Tiêu đề thực	Tiêu đề sinh
1	Introduction	Segmentation text
2	Research problem	Representation text
3.1	Content terms and their distribution	Terms
3.2	Lexical chains	Chains terms
3.3	Topic segmentation	Aim topics
3.4	Test corpora	Techniques corpora texts evaluation
4	Related research	Segmentation terms
5	Conclusions	Paper texts

Bảng 9. Sinh tiêu đề cho phân đoạn của C99

Phần	Tiêu đề	Câu bắt đầu	Câu kết thúc
1	Segmentation text	1	6
2	Representation subtopics texts	7	17

3	Synonyms	18	19
4	Terms	20	28
5	Information segmentation	29	35
6	Step topic	36	45
7	Techniques corpora texts evaluation	46	51
8	Addition topics	52	58
9	Referents	59	63

Bảng 10. Sinh tiêu đề cho phân đoạn của TextTiling

Phần	Tiêu đề	Câu bắt đầu	Câu kết thúc
1	Segmentation topics	1	7
2	Structure text	8	13
3	Cues texts	14	20
4	Terms	21	28
5	Information segmentation	29	34
6	Algorithms chains topic	35	45
7	Techniques corpora texts evaluation	46	52
8	Segmentation terms	53	63

4.5. Đánh giá thử nghiệm

Kết quả phân đoạn cho thấy chất lượng phân đoạn tương đối khả quan. Xét một cách trực quan khi quan sát Hình 5, ta thấy các điểm biên thứ 1, 4, 5, 6 hầu như không chệch so với văn bản gốc. Điều này được thể hiện qua độ đo P_k. Tuy nhiên cả hai thuật toán phân đoạn đều thống nhất chia đoạn văn bản thứ 5 "Topic Segmentation" ra làm hai đoạn tách rời với một bên là "segmentation" và một bên là "topic".

Kết quả sinh tiêu đề của luận văn đã được một số giáo viên tiếng Anh thuộc Trường Đại học Ngoại ngữ, ĐHQG Hà Nội thẩm định và cho rằng các kết quả đó là chấp nhận được tuy đôi chỗ còn mang ý nghĩa rất chung chung. Ví dụ như trường hợp sinh tiêu đề cho đoạn thứ 8 của văn bản gốc là "Paper Text" rất vô nghĩa so với "Conclusions". Tuy nhiên điều này cũng không khó hiểu do cách đặt tiêu đề của văn bản tuân theo cách đặt tiêu đề của các bài báo. Đây là vấn đề thuộc yếu tố văn phong. Với các thuật toán sinh tiêu đề không phụ thuộc miền ứng dụng thì vấn đề này là dễ hiểu.

4.5. Phương hướng cải tiến

Các kết quả thử nghiệm đã chứng mình bài toán xây dựng mục lục văn bản là khả thi và có triển vọng phát triển. Các thuật toán được trình bày trong luận văn tuy còn tương đối đơn giản và hầu hết là dựa trên luật nhưng đã tỏ ra rất hiệu quả trong thử nghiệm. Tuy nhiên ta vẫn có thể tăng cường chất lượng của thuật toán thông qua một số cải tiến sau:

- Đưa thêm các dấu hiệu nhận biết phân đoạn đặc trưng theo từng ngôn ngữ, ví dụ như trong tiếng Anh ta thường có "In this section", "As already discussed",… Các yếu tố này mang tính thống kê và do đó có thể thực hiện một mô hình thống kê trên một tập văn bản có sẵn để tìm ra tập các dấu hiệu chuyển chủ đề thông qua ngay câu đầu tiên của mỗi đoạn văn.
- Trong quá trình xây dựng các chuỗi token, nếu sử dụng từ gốc của các token thì khi đó sẽ làm giảm được nhiễu và tăng cường độ chính xác khi đo độ tương tự giữa các chuỗi token do chúng ta có thể xem xét được danh từ và tính từ với cùng một gốc thì tương đương nhau.
- Sử dụng các mô hình học có giám sát hoặc bán giám sát để học từ những dữ liệu có sẵn với các đặc trưng như: danh từ/cụm danh từ ở câu nào, vị trí như thế nào trong câu thì sẽ xuất hiện ở trong tiêu đề. Việc lựa chọn danh sách các từ cho tiêu đề dựa theo một mô hình học với các đặc trưng không liên quan đến bản thân từ sẽ giúp cho việc lựa chọn được những từ quan trọng và từ hiếm, chưa từng xuất hiện trong dữ liệu học.
- Để đảm bảo cú pháp của tiêu đề, cần thiết phải áp dụng một mô hình xác suất sinh để tạo ra một tiêu đề dễ hiểu và quen thuộc. Tuy nhiên, để đảm bảo được sự chính xác thì cần áp dụng mô hình cho các lĩnh vực khác nhau. Việc áp dụng cho từng loại văn bản cụ thể sẽ giúp tăng độ chính xác và chất lượng của mô hình.

4.6. Tóm tắt chương bốn

Trong chương này, luận văn đã trình bày quá trình thử nghiệm xây dựng mục lục cho một văn bản cụ thể nhằm chứng minh tính khả thi và triển vọng của bài toán xây dựng mục lục cho văn bản. Luận văn cũng đã trình bày mô hình trực quan để so sánh sự phân đoạn của các thuật toán. Đồng thời, luận văn cũng đưa ra một số đánh giá của cá nhân và của các chuyên gia về kết quả thử nghiệm và đưa ra một số phương hướng cải tiến giúp tăng cường chất lượng của mô

hình. Trong thời gian tiếp theo, tác giả sẽ tiếp tục thử nghiệm và cải tiến các thuật toán trên để đạt được kết quả cao hơn và hướng tới giải quyết triệt để bài toán phân đoạn đa cấp và sinh mục lục cho văn bản được phân đoạn đa cấp.

KẾT LUẬN

Luận văn đã tiến hành nghiên cứu và tìm hiểu bài toán xây dựng mục lục cho văn bản. Đây là một bài toán mới trong lĩnh vực xử lí ngôn ngữ tự nhiên và có liên hệ mật thiết với bài toán tóm tắt văn bản. Phương pháp giải quyết của luận văn là chia quá trình xây dựng mục lục thành hai quá trình nhỏ hơn là phân đoạn văn bản và sinh tiêu đề cho đoạn văn bản. Với mỗi quá trình này, luận văn đã tiến hành nghiên cứu, tìm hiểu và giới thiệu các phương pháp chính để giải quyết vấn đề đồng thời đánh giá ưu điểm cũng như khuyết điểm của các phương pháp. Luận văn đã tiến hành thử nghiệm trên một văn bản khoa học cụ thể để chứng minh tính khả thi của bài toán. Các kết quả thu được tương đối khả quan cho thấy triển vọng phát triển của bài toán.

Luận văn cũng đã đưa ra một số đề xuất về phương án tích hợp hai quá trình để giảm thiểu dư thừa dữ liệu cũng như thời gian tính toán. Thêm vào đó, luận văn cũng đã đề xuất một số cải tiến và hướng phát triển trong thời gian sắp tới để có thể đạt được những kết quả tốt hơn. Một số hướng phát triển tiếp theo của luận văn:

- Triển khai phân đoạn văn bản dựa trên chuỗi từ vựng với sự hỗ trợ từ WordNet.
- Cải tiến và đưa ra mô hình thuật toán mới cho phép phân đoạn văn bản đa cấp.
- Thử nghiệm các mô hình học có giám sát và bán giám sát trong việc sinh tiêu đề cho một văn bản.
- Triển khai thuật toán cải tiến dựa trên [Branavan 2007] để đảm bảo tính hợp lí và chất lượng của mục lục.

Đây là những hướng phát triển đã được nêu ra trong các chương của luận văn và có tính khả thi cao. Việc phát triển bài toán xây dựng mục lục cho văn bản có ý nghĩa lớn đối với các văn bản không có cấu trúc sẵn, trong đó đặc biệt là các văn bản dạng âm thanh.

TÀI LIỆU THAM KHẢO

1. Angheluta R., De Busser R.D., Moens M.F. (2002), "The Use of Topic Segmentation for Automatic Summarization", *In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, Philadelphia, USA.
2. Banko M., Mittal V.O., Witbrock M.J. (2000), "Headline Generation Based on Statistical Translation", *In Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics*, Hong Kong.
3. Beeferman D., Berger A., Lafferty J. (1999), "Statistical Models for Text Segmentation", *Machine Learning*, 34(1-3), pp. 177-210.
4. Branavan S.R.K., Deshpande P., Barzilay R. (2007), "Generating a Table-of-Contents", *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*: 544-551, Prague, Czech Republic.
5. Choi F. (2000), "Advances in domain independent linear text segmentation", *In Proceedings of NAACL '00*, pp. 26-33, Seattle, USA.
6. Church K.W. (1993), "Char align: A Program for Aligning Parallel Texts at the Character Level", *In Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, pp. 1-8, Ohio, USA.
7. Collins M., Roark B. (2004), "Incremental Parsing with the Perceptron Algorithm", *In Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, pp. 111-118, Barcelona, Spain.
8. Dorr B., Zajic D., Schwartz R. (2003), "Hedge Trimmer: A parse-and-trim approach to headline generation", *In Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization*: 1-8, Edmonton, Canada.
9. Elhada N., McKeown K.R. (2001), "Towards generating patient specific summaries of medical articles", *In Proceedings of NAACL Workshop on Automatic Summarization*, Pittsburgh, PA, USA.
10. Hearst M.A. (1994), "Multi-paragraph segmentation of expository text", *In Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics*, pp. 9-16, New Mexico, USA.
11. Hearst M.A. (1997), "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages", *Computational Linguistics*, 23(1), pp. 33-64.

12. Jin R., Hauptmann A.G. (2002), "A New Probability Model for Title Generation", *The 19th International Conference on Computational Linguistics*, Taiwan.

13. Jones K.S. (2007), "Automatic summarising: The state of the art", *Information Processing and Management*, doi:10.1016/j.ipm.2007.03.009.

14. Malioutov I., Barzilay R. (2006), "Minimum Cut Model for Spoken Lecture Segmentation", *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 25-32, Sydney, Australia.

15. Pevzner L., Hearst M.A. (2002), "A Critique and Improvement of an Evaluation Metric for Text Segmentation", *Computational Linguistics*, 28 (1), pp. 19-36.

16. Teufel S., Moens M. (2002), "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status", *Computational Linguistics*, 28(4), pp. 409-445.

17. Witbrock M.J., Mittal V.O. (1999), "Ultra-Summarization: A statistical Approach to Generating Highly Condensed Non-Extractive Summaries", *In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99)*, Poster Session, 315-316, USA.