ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN MAI VŨ

TÓM TẮT ĐA VĂN BẢN DỰA VÀO TRÍCH XUẤT CÂU

LUẬN VĂN THẠC SĨ

HÀ NỘI - 2009

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN MAI VŨ

TÓM TẮT ĐA VĂN BẢN DỰA VÀO TRÍCH XUẤT CÂU

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 THỤY

HÀ NỘI - 2009

Lời cảm ơn

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới PGS.TS. Hà Quang Thuy, người thầy đã chỉ bảo và hướng dẫn tận tình cho tôi trong suốt quá trình nghiên cứu khoa học và thực hiện luận văn này.

Tôi xin chân thành cảm ơn sự giúp đỡ và góp ý rất nhiệt tình của GS.TS. Kazuo Hashimoto trong quá trình nghiên cứu tại Đại học Tohoku, Nhật Bản.

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 anh chị em tại Phòng thí nghiệm Công nghệ tri thức và Tương tác người máy, Trường Đại học Công nghệ.

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!

Tác giả

Trần Mai Vũ

Lời cam đoan

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

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

Mục lục

Danh sách hình vẽ

Danh sách bảng

Bảng từ viết tắt

STT	Từ hoặc cụm từ	Viết tắt
1	Maximal Maginal Relevance	**MMR**
2	Question and Answering (Hệ thống hỏi đáp tự động)	**Q&A**
3	Document Understanding Conferences (Hội nghi chuyên về hiểu văn bản)	**DUC**
4	Term Frequency (Tần suất từ/cụm từ trong văn bản)	**TF**

Mở đầu

Sự phát triển nhanh chóng của mạng Internet cùng với những bước tiến mạnh mẽ của công nghệ lưu trữ, lượng thông tin lưu trữ hiện nay đang trở nên vô cùng lớn. Thông tin được sinh ra liên tục mỗi ngày trên mạng Internet, lượng thông tin văn bản khổng lồ trong đó đó đã và đang mang lại lợi ích không nhỏ cho con người, tuy nhiên, nó cũng khiến chúng ta khó khăn trong việc tìm kiếm và tổng hợp thông tin. Giải pháp cho vấn đề này là **tóm tắt văn bản tự động**. Tóm tắt văn bản tự động được xác định là một bài toán thuộc lĩnh vực khái phá dữ liệu văn bản; việc áp dụng tóm tắt văn bản sẽ giúp người dùng tiết kiệm thời gian đọc, cải thiện tìm kiếm cũng như tăng hiệu quả đánh chỉ mục cho máy tìm kiếm.

Từ nhu cầu thực tế như thế, bài toán tóm tắt văn bản tự động nhận được sự quan tâm nghiên cứu của nhiều nhà khoa học, nhóm nghiên cứu cũng như các công ty lớn trên thế giới. Các bài báo liên quan đến tóm tắt văn bản xuất hiện nhiều trong các hội nghị nổi tiếng như : DUC[1] 2001-2007, TAC[2] 2008, ACL[3] 2001-2007… bên cạnh đó cũng là sự phát triển của các hệ thống tóm tắt văn bản như : MEAD, LexRank, Microsoft Word (Chức năng AutoSummarize)…

Một trong những vấn đề thách thức và được sự quan tâm trong những năm gần đây đối với bài toán tóm tắt văn bản tự động đó là đưa ra kết quả tóm tắt cho một tập văn bản liên quan với nhau về mặt nội dung hay còn gọi là **tóm tắt đa văn bản**.

Bài toán tóm tắt đa văn bản được xác định là một bài toán có độ phức tạp cao. Đa số mọi người nghĩ rằng, tóm tắt đa văn bản chỉ là việc áp dụng tóm tắt đơn văn bản cho một văn bản được ghép từ các văn bản trong một tập văn bản cho trước. Tuy nhiên điều đó là hoàn toàn không chính xác, thách thức lớn nhất của vấn đề tóm tắt đa văn là do dữ liệu đầu vào có thể có sự nhập nhằng ngữ nghĩa giữa nội dung của văn bản này với văn bản khác trong cùng tập văn bản hay trình tự thời gian được trình bày trong

[1] Document Understanding Conference. http://duc.nist.gov
[2] Text Analysis Conference. http://www.nist.gov/tac
[3] Association for Computational Linguistics. http://aclweb.org

mỗi một văn bản là khác nhau, vì vậy để đưa ra một kết quả tóm tắt tốt sẽ vô cùng khó khăn [EWK].

Rất nhiều ứng dụng cần đến quá trình tóm tắt đa văn bản như: hệ thống hỏi đáp tự động (Q&A System), tóm tắt các báo cáo liên quan đến một sự kiện, tóm tắt các cụm dữ liệu được trả về từ quá trình phân cụm trên máy tìm kiếm… Hướng nghiên cứu ứng dụng bài toán tóm tắt đa văn bản vào việc xây dựng hệ thống hỏi đáp tự động đang là hướng nghiên cứu chính của cộng đồng nghiên cứu tóm tắt văn bản nhưng năm gần đây. Rất nhiều nghiên cứu cho thấy rằng, việc sử dụng phương pháp tóm tắt đa văn bản dựa vào câu truy vấn (Query-based multi-document summarization) đối với kho dữ liệu tri thức để đưa ra một văn bản tóm tắt trả lời cho câu hỏi của người sử dụng đạt được nhiều kết quả khả quan cũng như thể hiện đây là một hướng tiếp cận đúng đắn trong việc xây dựng các mô hình hỏi đáp tự động [Ba07,YYL07].

Với việc lựa chọn đề tài "**Tóm tắt đa văn bản dựa vào trích xuất câu**", chúng tôi tập trung vào việc nghiên cứu, khảo sát, đánh giá và đề xuất ra một phương pháp tóm tắt đa văn bản phù hợp với ngôn ngữ tiếng Việt, bên cạnh đó áp dụng phương pháp này vào việc xây dựng một mô hình hệ thống hỏi đáp tiếng Việt.

Ngoài phần mở đầu và kết luận, luận văn được tổ chức thành 5 chương như sau:

- **Chương 1: Khái quát bài toán tóm tắt** giới thiệu khái quát bài toán tóm tắt văn bản tự động nói chung và bài toán tóm tắt đa văn bản nói riêng, trình bày một số khái niệm và cách phân loại đối với bài toán tóm tắt.
- **Chương 2: Tóm tắt đa văn bản dựa vào trích xuất câu** giới thiệu chi tiết về hướng tiếp cận, thách thức và các vấn đề trong giải quyết bài toán tóm tắt đa văn bản dựa vào trích xuất câu.
- **Chương 3: Độ tương đồng câu và các phương pháp tăng cường tính ngữ nghĩa cho độ tương đồng câu** trình bày các nghiên cứu về các phương pháp tính độ tương đồng ngữ nghĩa câu tiêu biểu áp dụng vào quá trình trích xuất câu quan trọng của văn bản.

- **Chương 4: Một số đề xuất tăng cường tính ngữ nghĩa cho độ tương đồng câu và áp dụng vào mô hình tóm tắt đa văn tiếng Việt** phân tích, đề xuất một phương pháp tích hợp các thuật toán để giải quyết bài toán tóm tắt đa văn bản tiếng Việt và trình bày việc áp dụng phương pháp được đề xuất để xây dựng mô hình hệ thống hỏi đáp tiếng Việt đơn giản.
- **Chương 5: Thực nghiệm và đánh giá** trình bày quá trình thử nghiệm của luận văn và đưa ra một số đánh giá, nhận xét các kết quả đạt được.

Chương 1. Khái quát bài toán tóm tắt văn bản

1.1. Bài toán tóm tắt văn bản tự động

Vào năm 1958, Luhn của IBM đã trình bày phương pháp tóm tắt tự động cho các bài báo kĩ thuật sử dụng phương pháp thống kê thông qua tần suất và phân bố của các từ trong văn bản [Lu58]. Tuy nhiên mãi cho đến những năm cuối thế kỷ 20, với sự phát triển của Internet, lượng thông tin bùng nổ nhanh chóng, việc thu nhận những thông tin quan trọng cũng trở thành một vấn đề thiết yếu thì bài toán tóm tắt văn bản tự động mới được sự quan tâm thiết thực của nhiều nhà nghiên cứu.

Theo Inderjeet Mani, 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"* [MM99].

Việc đưa ra được một văn bản kết quả tóm tắt có chất lượng như là văn bản do con người làm ra mà không bị giới hạn bởi miền ứng dụng là được xác định là cực kỳ khó khăn. Vì vậy, các bài toán được giải quyết trong tóm tắt văn bản thường chỉ hướng đến một kiểu văn bản cụ thể hoặc một kiểu tóm tắt cụ thể.

1.2. Một số khái niệm của bài toán tóm tắt và phân loại tóm tắt

- Tỷ lệ nén(Compression Rate): là độ đo thể hiện bao nhiêu thông tin được cô đọng trong văn bản tóm tắt được tính bằng công thức:

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

SummaryLength: Độ dài văn bản tóm tắt

SourceLength: Độ dài văn bản nguồn

- Độ nổi bật hay liên quan(Salience or Relevance): là trọng số được gán cho thông tin trong văn bản thể hiện độ quan trọng của thông tin đó đối với toàn văn bản hay để chỉ sự liên quan của thông tin đó đối với chương trình của người sử dụng.

- Sự mạch lạc(coherence): Một văn bản tóm tắt gọi là mạch lạc nếu tất cả các thành phần nằm trong nó tuân theo một thể thống nhất về mặt nội dung và không có sự trùng lặp giữa các thành phần.

Phân loại bài toán tóm tắt.

Có nhiều cách phân loại tóm tắt văn bản khác nhau tuy nhiên sự phân loại chỉ mang tính tương đối, phụ thuộc vào việc tóm tắt trên cơ sở nào. Ở đây, luận văn đề cập đến phân loại tóm tắt dựa trên 3 cơ sở là: dựa vào định dạng, nội dung đầu vào, dựa vào định dạng, nội dung đầu ra, dựa vào mục đích tóm tắt.

- Tóm tắt dựa trên cơ sở định dạng, nội dung đầu vào sẽ trả lời cho câu hỏi "Cái gì sẽ được tóm tắt". Cách chia này sẽ cho ta nhiều cách phân loại con khác nhau. Cụ thể như:

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

- Định dạng văn bản: dựa vào từng định dạng văn bản khác nhau, tóm tắt cũng chia ra thành các loại khác nhau như: tóm tắt văn bản không theo khuôn mẫu (free-form) hay tóm tắt văn bản có cấu trúc. Với văn bản có cấu trúc, tóm tắt văn bản thường sử dụng một mô hình học dựa vào mẫu cấu trúc đã xây dựng từ trước để tiến hành tóm tắt.

- Số lượng dữ liệu đầu vào: tùy vào số lượng đầu vào của bài toán tóm tắt, người ta cũng có thể chia tóm tắt ra thành tóm tắt đa văn bản, tóm tắt đơn văn bản. Tóm tắt đơn văn bản khi đầu vào chỉ là một văn bản đơn, trong khi đó đầu vào của tóm tắt đa văn bản là một tập các tài liệu có liên quan đến nhau như: các tin tức có liên quan đến cùng một sự kiện, các trang web cùng chủ đề hoặc là cụm dữ liệu được trả về từ quá trình phân cụm.

- Miền dữ liệu: dựa vào miền của dữ liệu như cụ thể về một lĩnh vực nào đó, ví dụ như: y tế, giáo dục… hay là miền dữ liệu tổng quát, có thể chia tóm tắt ra thành từng loại tương ứng.

- Tóm tắt trên cơ sở mục đích thực chất là làm rõ cách tóm tắt, mục đích tóm tắt là gì, tóm tắt phục vụ đối tượng nào ...

- Nếu phụ thuộc vào đối tượng đọc tóm tắt thì tóm tắt cho chuyên gia khác cách tóm tắt cho các đối tượng đọc thông thường.

- Tóm tắt sử dụng trong tìm kiếm thông tin (IR) sẽ khác với tóm tắt phục vụ cho việc sắp xếp.

- **Dựa trên mục đích tóm tắt, còn có thể chia ra thành tóm tắt chỉ thị (Indicative) và tóm tắt thông tin (Informative)**. Tóm tắt chỉ thị (indicative) chỉ ra loại của thông tin, ví dụ như là loại văn bản chỉ thị "tối mật". Còn tóm tắt thông tin chỉ ra nội dung của thông tin.

- **Tóm tắt trên cơ sở truy vấn (Query-based) hay tóm tắt chung (General)**. Tóm tắt general mục đích chính là tìm ra một đoạn tóm tắt cho toàn bộ văn bản mà nội dung của đoạn văn bản sẽ bao quát toàn bộ nội dung của văn bản đó. Tóm tắt trên cơ sở truy vấn thì nội dung của văn bản tóm tắt sẽ dựa trên truy vấn của người dùng hay chương trình đưa vào, loại tóm tắt này thường được sử dụng trong quá trình tóm tắt các kết quả trả về từ máy tìm kiếm.

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

- **Dựa vào ngôn ngữ:** Tóm tắt cũng có thể phân loại dựa vào khả năng tóm tắt các loại ngôn ngữ:

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

- **Dựa vào định dạng đầu ra của kết quả tóm tắt:** như bảng, đoạn, từ khóa.

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

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

Hiện nay, các hệ thống sử dụng tóm tắt theo trích xuất được sử dụng phổ biến và cho kết quả tốt hơn tóm tắt theo tóm lược. Nguyên nhân tạo ra sự khác biệt này là do các vấn đề trong bài toán tóm tắt theo tóm lược như: biểu diễn ngữ nghĩa, suy luận và sinh ra ngôn ngữ tự nhiên được đánh giá là khó và chưa có nhiều kết quả nghiên cứu khả quan hơn so với hướng trích xuất câu của bài toán tóm tắt theo trích xuất. Trong thực tế, theo đánh giá của Dragomir R. Radev (Đại học Michigan, Mỹ) chưa có một hệ thống tóm tắt theo tóm lược đạt đến sự hoàn thiện, các hệ thống tóm tắt theo tóm lược hiện nay thường dựa vào thành phần trích xuất có sẵn. Các hệ thống này thường được biết đến với tên gọi **tóm tắt theo nén văn bản**.

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

Mặc dù dựa vào nhiều cơ sở có nhiều loại tóm tắt khác nhau tuy nhiên hai loại tóm tắt là **tóm tắt đơn văn bản** và **tóm tắt đa văn bản** vẫn được sự quan tâm lớn của các nhà nghiên cứu về tóm tắt tự động.

1.3. Tóm tắt đơn văn bản

Bài toán tóm tắt văn bản đơn cũng giống như các bài toán tóm tắt khác, là một quá trình tóm tắt tự động với đầu vào là một văn bản, đầu ra là một đoạn mô tả ngắn gọn nội dung chính của văn bản đầu vào đó. Văn bản đơn có thể là một trang Web,

một bài báo, hoặc một tài liệu với định dạng xác định (ví dụ : .doc, .txt)… Tóm tắt văn bản đơn là bước đệm cho việc xử lý tóm tắt đa văn bản và các bài toán tóm tắt phức tạp hơn. Chính vì thế những phương pháp tóm tắt văn bản ra đời đầu tiên đều là các phương pháp tóm tắt cho văn bản đơn.

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

Tóm tắt theo trích xuất

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

Tóm tắt theo tóm lược

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

cho ra kết quả tốt tuy nhiên các phương pháp dạng này thường chỉ áp dụng trong một miền nhất định [MR95].

1.4. Tóm tắt đa văn bản

Tóm tắt đa văn bản có thể được coi như là một mở rộng của tóm tắt đơn văn bản. Mục đích của tóm tắt đa văn bản:

Là quá trình trích xuất nội dung từ một tập các văn bản có liên quan đến nhau, trong quá trình đó các thông tin dư thừa sẽ được loại bỏ và những thông tin quan trọng sẽ được biểu diễn dưới hình thức cô đọng, súc tích và giàu cảm súc đến người sử dụng hoặc chương trình cần dùng [MM99].

Tóm tắt đa văn bản được xác định là một bài toán có độ phức tạp cao, ngoài những thách thức đã được biết đến đối với tóm tắt đơn văn bản như sự cô đọng của thông tin và mạch lạc về nội dung, tóm tắt đa văn bản còn có những thách thức như cần phải xác định những thông tin trùng lặp giữa các văn bản, xác định thông tin quan trọng trong nhiều văn bản hay việc sắp xếp các thông tin trong văn bản tóm tắt.

Do tóm tắt đa văn bản là một mở rộng của tóm tắt đơn văn bản, cho nên cũng như tóm tắt văn bản đơn các phương pháp giải quyết tóm tắt đa văn bản cũng đi theo hai hướng tiếp cận là dựa vào trích xuất và dựa vào tóm lược. Tuy nhiên, do những hạn chế của phương pháp giải quyết bằng tóm tắt theo tóm lược đã được nêu ở trên, các phương pháp giải quyết tóm tắt đa văn bản hầu như tập trung vào **phương pháp tóm tắt đa văn bản dựa vào trích xuất câu**. Chính từ tình hình thực tế đấy, luận văn đã tập trung nghiên cứu, khảo sát các kỹ thuật tóm tắt đa văn bản liên quan đến phương pháp tóm tắt văn bản dựa vào trích xuất câu để giải quyết bài toán tóm tắt đa văn bản tiếng Việt.

1.5. Tóm tắt chương một

Trong chương này luận văn giới thiệu khái quát bài toán tóm tắt văn bản tự động các vấn đề liên quan và cách phân loại đối với bài toán tóm tắt văn bản tự động. Trong chương tiếp theo, luận văn sẽ làm rõ các vấn đề của bài toán tóm tắt đa văn bản nói chung và bài toán tóm tắt đa văn bản dựa vào trích xuất câu nói riêng.

Chương 2. Tóm tắt đa văn bản dựa vào trích xuất câu

2.1. Hướng tiếp cận của bài toán tóm tắt đa văn bản

Như chúng ta đã biết ở trên tóm tắt văn bản nói chung và tóm tắt đa văn bản nói riêng là bài toán thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Trong phân tích xử lý ngôn ngữ tự nhiên có các mức độ sâu xử lý khác nhau được sắp xếp theo thứ tự như sau: đầu tiên là mức hình thái (Morphological), tiếp theo là mức cú pháp (Syntactic), tiếp đến là mức ngữ nghĩa (Semantic) và cuối cùng là mức ngữ dụng (Pragmatic). Tương tự như các độ sâu xử lý của xử lý ngôn ngữ tự nhiên, phương pháp tiếp cận để giải quyết bài toán tóm tắt đa văn bản cũng có thể được phân loại dựa vào độ sâu xử lý được thực hiện trong quá trình tóm tắt. Tuy nhiên phương pháp tiếp cận để giải quyết bài toán tóm tắt đa văn bản chỉ có ba mức, là các mức: hình thái, cú pháp và ngữ nghĩa.

Mức hình thái: tại mức xử lý này, trong các văn bản, đơn vị được sử dụng để so sánh là các ngữ, câu hay đoạn văn (paragraph). Các phương pháp tại mức này thường sử dụng độ đo tương đồng dựa trên mô hình không gian vector (Vector space model) áp dụng trọng số TF.IDF cho các từ và các câu. Phương pháp tóm tắt MMR [CG98] là phương pháp nổi bật tại mức xử lý này.

Mức cú pháp: đơn vị được sử dụng để so sánh tại mức xử lý này là sử dụng việc phân tích những cấu trúc ngữ pháp tương ứng giữa các văn bản với nhau. Các phương pháp tại mức này tập trung vào việc phân tích cấu trúc ngữ pháp giữa các câu hay các ngữ trong từng đoạn văn thuộc các văn bản. Phương pháp do Barzilay và các đồng tác giả khác đề xuất năm 1999 [BME99] thuộc mức xử lý này.

Mức ngữ nghĩa: tại mức xử lý này tập trung nhiều vào việc phân tích các tên thực thể, mối quan hệ giữa các thực thể cũng như các sự kiện nảy sinh thực thể để xác định được độ quan trọng của thông tin. Phương pháp của McKeown và Radev đề xuất năm 1995[MR95] là một dạng của tóm tắt tại mức xử lý này.

Dựa vào các đặc trưng của từng phương pháp tiếp cận, Inderjeet Mani đã đưa ra bảng so sánh, đánh giá ba mức tiếp cận để giải quyết bài toán tóm tắt đa văn bản [Ma01].

Mức xử lý	Đặc tính	Ưu điểm	Nhược điểm
Mức hình thái	Sử dụng nhiều các độ đo tương đồng giữa các từ vựng	Sử dụng rất phổ biến, xử lý dư thừa tốt	Không thể mô tả các đặc trưng khác, khả năng tổng hợp thông tin kém.
Mức cú pháp	So sánh giữa các cây cú pháp của câu hay ngữ trong văn bản	Có khả năng phát hiện các khái niệm tượng đồng trong các ngữ,cho phép tổng hợp thông tin.	Không thể mô tả các đặc trưng khác, đòi hỏi phải mở rộng các luật so sánh giữa các cây cú pháp
Mức ngữ nghĩa	So sánh giữa các mẫu tài liệu đã được ấn định.	Có khả năng mô tả nhiều đặc trưng khác nhau.	Các mẫu phải được tạo trước đối với từng miền.

Bảng 2.1. Bảng so sánh các phương pháp tiếp cận tóm tắt đa văn bản [Ma01].

2.2. Các thách thức của quá trình tóm tắt đa văn bản

Một trong những thách thức lớn nhất của tóm tắt đa văn bản chính là sự nhập nhằng nội dung giữa các văn bản. Có ba nguyên nhân gây ra nhập nhằng nội dung trong tóm tắt đa văn bản đó là: đồng tham chiếu xuyên văn bản, nhập nhằng về thời gian xuyên văn bản, sự trùng lặp nội dung giữa các văn bản.

Trùng lặp đại từ và đồng tham chiếu

Thông thường, chúng ta đề cập đến một tên thực thể chính là nói đến tên ban đầu của thực thể đấy và sau đó thường hay sử dụng một đại từ thay thế nói về thực thể

trên. Xác định chính xác được thực thể mà đại từ chỉ đến được gọi là việc **xác định trùng lặp đại từ** (Pronominal Anaphora resolution).

Việc xác định đúng hai hay nhiều hơn các thực thể của nhiều văn bản khác nhau cùng chỉ đến một thực thể được gọi là vấn đề **xác định đồng tham chiếu xuyên văn bản** (Cross Document Co-Reference). Vấn đề này cần phải được giải quyết tốt thì kết quả đầu ra của tóm tắt đa văn bản mới cho ra kết quả tốt và dễ hiểu.

Nhập nhằng mặt thời gian

Các văn bản trong cụm tài liệu có thể được chỉ đến bởi nhiều từ hay cụm từ chỉ thời gian ví dụ: hôm qua, hôm nay… Việc xác định rõ ràng các mốc thời gian tương ứng là một điều kiện cần để sắp xếp các câu hay các văn bản theo đúng trình tự hợp lý. Một số hệ thống có khả năng xác định được mốc thời gian và thay thế các mốc thời gian tương đối thành các mốc thời gian tuyệt đối bằng việc phân tích nội dung của văn bản.

Để đảm bảo tính có thể đọc được đối với văn bản tóm tắt của hệ thống tóm tắt đa văn bản thì ba yếu tố: Xác định trùng lặp đại từ, xác định đồng tham chiếu xuyên văn bản và nhập nhằng về mặt thời gian cần phải được giải quyết tốt. Mặc dù, trong tóm tắt đơn văn bản hai yếu tố đầu tiên vẫn xuất hiện tuy nhiên giải quyết hai vấn đề này không phức tạp như giải quyết trong tóm tắt đa văn bản. Bên cạnh đó, vấn đề nhập nhằng thời gian không xuất hiện trong tóm tắt văn bản đơn, do các văn bản đơn đầu vào coi như đã đảm bảo về mặt trật tự, yếu tố này do chính người tạo ra văn bản tạo nên [Ji98]. Mặc dù vậy đối với tóm tắt đa văn bản, vấn đề này trở nên cực kỳ khó khăn, các nghiên cứu xoay quanh vấn đề này chỉ tập trung vào các loại dữ liệu có đi kèm với thời gian như tin tức hay chuỗi các sự kiện. Một trong các phương pháp giải quyết tốt vấn đề này được Barzilay, Elhadad và McKeown đưa ra vào năm 2002 [BME02]. Còn đối với các tập dữ liệu không rõ ràng về mặt thời gian, các nhà nghiên cứu mặc định như các văn bản tương đồng về mặt thời gian.

Sự chồng chéo nội dung giữa các tài liệu

Một câu hỏi mà nhiều người đặt ra đối với tóm tắt đa văn bản đó là:

- Liệu có thể ghép các văn bản lại với nhau rồi sử dụng tóm tắt đơn văn bản?

- Câu trả lời ở đây là **không**!

Bằng cách đó chúng ta sẽ không tạo ra được một văn bản tóm tắt tốt do không loại bỏ được sự chống chéo về mặt nội dung cũng như xác định được mối quan hệ giữa các văn bản.

Mối quan hệ giữa các văn bản có rất nhiều loại khác nhau. Dragomir Radev đã liệt kê ra 24 loại quan hệ giữa các văn bản [Ra00] như trong bảng 2.2. Các mối quan hệ tồn tại ở nhiều mức khác nhau: mức từ (W), mức ngữ (P), mức đoạn hoặc mức câu (S), mức toàn tài liệu (D).

Đây là một taxonomy của các mối quan hệ xuyên tài liệu được gọi là **Cross-document Structure Theory (CST)**. Việc sử dụng tốt CST sẽ tạo hiệu quả cực kỳ hữu ích cho việc xác định sự trùng lặp giữa các văn bản trong bài toán tóm tắt đa văn bản.

#	Relationship type	Level	Description
1	Identity	Any	The same text appears in more than one location
2	Equivalence (paraphrasing)	S, D	Two text spans have the same information content
3	Translation	P, S	Same information content in different languages
4	Subsumption	S, D	One sentence contains more information than another
5	Contradiction	S, D	Conflicting information
6	Historical background	S	Information that puts current information in context
7	Cross-reference	P	The same entity is mentioned
8	Citation	S, D	One sentence cites another document
9	Modality	S	Qualified version of a sentence
10	Attribution	S	One sentence repeats the information of another while adding an attribution
11	Summary	S, D	Similar to Summary in RST: one textual unit summarizes another
12	Follow-up	S	Additional information which reflects facts that have happened since the last account
13	Elaboration	S	Additional information that wasnt included in the last account
14	Indirect speech	S	Shift from direct to indirect speech or vice-versa
15	Refinement	S	Additional information that is more specific that the one previously included
16	Agreement	S	One source expresses agreement with another
17	Judgment	S A	qualified account of a fact
18	Fulfilment	S A	prediction turned true
19	Description	S	Insertion of a description
20	Reader profile	S	Style and background-specific change
21	Contrast	S	Contrasting two accounts or facts
22	Parallel	S	Comparing two accounts of facts
23	Generalization	S	Generalization
24	Change of perspective	S, D	The same source presents a fact in a different light

Bảng 2.2. Taxonomy mối quan hệ xuyên văn bản [Ra00]

Tỷ lệ nén

Bên cạnh các vấn đề nhập nhằng về mặt nội dung thì tỷ lệ nén cũng là một vấn đề được đặt ra khi nói đến tóm tắt đa văn bản. Trong tóm tắt đơn văn bản, tỷ lệ 10% so với chiều dài của văn bản gốc có thể đủ đối với một văn bản tóm tắt. Tuy nhiên đối với một cụm tài liệu n tài liệu với tỷ lệ 10% ta có một văn bản có độ dài 0.1n độ dài trung bình văn bản. Với n là biến, văn bản tóm tắt có thể sẽ trở nên lớn hơn nhiều so với nhu cầu của người sử dụng muốn đọc. Chính vì vậy đối với tóm tắt đa văn bản, tỷ lệ nén cần có sự liên quan đến kích thước của cụm tài liệu đó. Đối với tóm tắt đa văn bản dựa

vào trích xuất câu để đưa ra một văn bản tóm tắt có độ dài phù hợp với yêu cầu của người sử dụng, tỷ lệ nén thường được thay thế bằng **số lượng câu** của văn bản tóm tắt.

2.3. Đánh giá kết quả tóm tắt

Đánh giá kết quả tóm tắt văn bản là một việc làm khó khăn trong thời điểm hiện tại. Việc sử dụng ý kiến đánh giá của các chuyên gia ngôn ngữ được xem là cách đánh giá tốt nhất, tuy nhiên, cách làm này lại tốn rất nhiều chi phí. Bên cạnh các phương pháp đánh giá thủ công do các chuyên gia thực hiện, vấn đề đánh giá tự động kết quả tóm tắt cũng nhận được nhiều sự chú ý hiện nay. NIST[1] kể từ năm 2000 đã tổ chức hội nghị DUC mỗi năm một lần để thực hiện việc đánh giá với quy mô lớn các hệ thống tóm tắt văn bản.Việc đánh giá tự động này nhằm mục đích là tìm ra được một độ đo đánh giá tóm tắt gần với những đánh giá của con người nhất.

Độ hồi tưởng (recall) tại các tỷ lệ nén khác nhau chính là thước đo đánh giá hợp lý, mặc dù nó không chỉ ra được sự khác nhau về hiệu suất của hệ thống. Vì vậy độ đo về sự bao phủ được tính theo công thức:

C = R × E

Ở đây, R là độ hồi tưởng câu được trả về bởi công thức

R = Số đơn vị bao phủ/ Tổng số đơn vị trong mô hình tóm tắt.

E là tỷ lệ hoàn thành nằm trong khoảng từ 0 đến 1 (1 là hoàn thành tất cả, ¾ là một phần, ½ là một số, ¼ là khó, 0 là không có)

DUC 2002 đã sử dụng một phiên bản để điều chỉnh chiều dài của thước đo bao phủ, C':

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

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

---

[1] National Institute of Standards and Technology. http://nist.goc

Phương pháp **ROUGE**

BiLingual Evaluation Understudy (BLEU) [KST02] là một phương pháp của cộng động dịch máy đưa ra để đánh giá tự động các hệ thống dịch máy. Phương pháp này có hiểu qua nhanh, độc lập với ngôn ngữ và sự liên quan với các đánh giá của con người. Recall Oriented Understudy of Gisting Evaluation (ROUGE) [LH03] là một phương pháp do Lin và Hovy đưa ra vào năm 2003 cũng dựa trên các khái niệm tương tự. Phương pháp này sử dụng n-gram để đánh giá sự tương quan giữa các kết quả của mô hình tóm tắt và tập dữ liệu đánh giá. Phương pháp này đã cho ra kết quả khả quan và được sự đánh giá cao của cộng đồng nghiên cứu tóm tắt văn bản.

2.4. Tóm tắt đa văn bản dựa vào trích xuất câu

Tóm tắt đa văn bản dựa vào trích xuất câu là phương pháp giải quyết bài toán tóm tắt đa văn bản theo hướng tiếp cận ở mức hình thái. Phương pháp này có ưu điểm là xử lý tốt các dự thừa do chồng chéo về mặt nội dung giữa các văn bản trong cụm và cho ra hiệu quả cao đối với văn bản tóm tắt. Chính vì ưu điểm này nên tóm tắt đa văn bản dựa vào trích xuất câu được sự quan tâm,phát triển và sử dụng rộng rãi của cộng động tóm tắt văn bản tự động [HMR05, FMN07, BKO07]. Mặc dù có nhiều phương pháp được công bố nhưng hầu hết các phương pháp đều tập trung vào giải quyết hai vấn đề chính, đó là:

- Xác định và loại bỏ sự trùng lặp, chồng chéo về mặt nội dung giữa các văn bản.
- Sắp xếp các câu trong các văn bản theo độ nổi bật(quan trọng) về mặt nội dung hoặc độ liên quan đến một truy vấn do người sử dụng hay chương trình cung cấp.

2.4.1. Loại bỏ chồng chéo và sắp xếp các văn bản theo độ quan trọng

Loại bỏ chồng chéo và sắp xếp độ quan trọng giữa các văn bản trong cụm văn bản là một trong những vấn đề quan trọng nhất của bài toán tóm tắt đa văn bản. Một trong các phương pháp phổ biến để tính được độ quan trọng này là phương pháp MMR (Maximal Maginal Relevance) do Jaime Carbonell và Jade Goldstein đề xuất năm

1998 [CG98]. Đầu vào của phương pháp này là một cụm văn bản đã được sắp xếp sẵn và đầu ra là cụm văn bản đã được sắp xếp lại theo thứ tự về ngữ nghĩa. Phương pháp này sắp xếp các văn bản dựa vào việc xác định một độ đo làm rõ ranh giới về ngữ nghĩa giữa các văn bản trong cụm. Mỗi một văn bản có độ đo này cực đại nếu độ đo về sự tương đồng giữa văn bản với câu truy vấn cao và cực tiểu được sự tương đồng giữa văn bản này và các văn bản khác đã được chọn trước đấy. Công thức để tính độ đo này như sau:

$$MMR \overset{def}{=} Arg \max_{D_i \in R \setminus S} [\lambda * (Sim_1(D_i, Q) - (1-\lambda) * \max_{D_j \in S} Sim_2(D_i, D_j))]$$

Trong đó:

λ: là tham số nằm trong ngưỡng [0,1] để quyết định việc đóng góp giữa 2 độ đo. Nếu λ=1 thì độ quan trọng của văn bản chỉ phụ thuộc vào độ đo tương đồng giữa văn bản và câu truy vấn, còn nếu λ=0 thì độ đo sự tương đồng giữa văn bản này và văn bản khác sẽ đạt giá trị cực đại trong biểu thức trên.

C: cụm văn bản.

D_i: văn bản thuộc cụm C.

Q: là câu truy vấn (hay câu hỏi người dùng đưa vào).

R=IR(C,Q,θ) : là tập các văn bản của C đã được sắp xếp thứ tự theo sự liên quan với câu truy vấn Q dựa vào một ngưỡng xác định θ.

S: là tập các văn bản của R đã được chọn .

R\S: là tập các văn bản chưa được chọn của R.

Sim_1,Sim_2: là độ đo về sự tương đồng giữa hai văn bản.

2.4.2. Phương pháp sắp xếp câu

Xác định độ quan trọng câu là bước xuất hiện hầu hết trong các phương pháp tóm tắt đơn văn bản cũng như tóm tắt đa văn bản hiện nay. Độ đo quan trọng này có thể được xây dựng bằng cách kết hợp nhiều độ đo độ tương đồng câu khác nhau với các phương pháp cải tiến từ phương pháp MMR để làm tăng độ quan trọng đối với

mức ngữ nghĩa câu [HMR05, FMN07, BKO07]. Công thức của phương pháp MMR được cải tiến cho mức ngữ nghĩa câu:

$$Score(s_i) = \arg\max_{s_i}[\lambda * sim(s,q) - (1-\lambda) * \max sim(s_i, s_j)]$$

Trong đó:

λ: là tham số nằm trong ngưỡng [0,1] để quyết định việc đóng góp giữa 2 độ đo.

q: là câu truy vấn (hay câu hỏi người dùng đưa vào).

s_i: là một câu trong cụm văn bản.

s_j: các câu khác nằm trong cụm văn bản

sim: độ đo về sự tương đồng giữa hai câu

Nhận xét

Cả hai vấn đề cần giải quyết trong bài toán tóm tắt đa văn bản dựa vào trích xuất câu đều tập trung vào việc xác định được sự tương đồng giữa hai văn bản nói chung và giữa hai câu nói riêng. Trên thực tế, các phương pháp áp dụng và cải tiến cho tóm tắt đa văn bản dựa vào đều tập trung vào vấn đề là tăng cường tính ngữ nghĩa cho độ đo tương đồng giữa hai câu hay hai văn bản [HMR05, FMN07, BKO07]. Trong chương 3, luận văn sẽ đi sâu vào giới thiệu chi tiết đến các phương pháp tăng cường tính ngữ nghĩa cho độ tương đồng câu.

2.5. Tóm tắt chương hai

Trong chương này luận văn đã giới thiệu chi tiết đến hướng tiếp cận, các vấn đề đặt ra đối với bài toán tóm tắt đa văn bản và một số phương pháp để giải quyết các vấn đề trên. Trong chương tiếp theo, luận văn tiếp tục tập trung vào việc giới thiệu các phương pháp nhằm tương cường tính ngữ nghĩa cho độ tương đồng giữa hai câu.

Chương 3. Độ tương đồng câu và các phương pháp tăng cường tính ngữ nghĩa cho độ tương đồng câu

3.1. Độ tương đồng

Trong toán học, một độ đo là một hàm số cho tương ứng với một "chiều dài", một "thể tích" hoặc một "xác suất" với một phần nào đó của một tập hợp cho sẵn. Nó là một khái niệm quan trọng trong giải tích và trong lý thuyết xác suất.

Ví dụ, độ đo đếm được định nghĩa bởi $\mu(S)$ = số phần tử của S

Rất khó để đo sự giống nhau, sự tương đồng. Sự tương đồng là một đại lượng (con số) phản ánh cường độ của mối quan hệ giữa hai đối tượng hoặc hai đặc trưng. Đại lượng này thường ở trong phạm vi từ -1 đến 1 hoặc 0 đến 1. Như vậy, một độ đo tương đồng có thể coi là một loại scoring function (hàm tính điểm).

Ví dụ, trong mô hình không gian vector, ta sử dụng độ đo cosine để tính độ tương đồng giữa hai văn bản, mỗi văn bản được biểu diễn bởi một vector.

3.2. Độ tương đồng câu

Phát biểu bài toán độ tính tương đồng câu như sau: Xét một tài liệu d gồm có n câu: $d = s_1, s_2, ..., s_n$. Mục tiêu của bài toán là tìm ra một giá trị của hàm $S(s_i, s_j)$ với $S \in (0,1)$, và i, j = 1, ..., n. Hàm $S(s_i, s_j)$ được gọi là độ đo tương đồng giữa hai câu s_i và s_j. Giá trị càng cao thì sự giống nhau về nghĩa của hai câu càng nhiều.

Ví dụ: Xét hai câu sau: “Tôi là nam” và “Tôi là nữ”, bằng trực giác có thể thấy rằng hai câu trên có sự tương đồng khá cao.

Độ tương đồng ngữ nghĩa là một giá trị tin cậy phản ánh mối quan hệ ngữ nghĩa giữa hai câu. Trên thực tế, khó có thể lấy một giá trị có chính xác cao bởi vì ngữ nghĩa chỉ được hiểu đầy đủ trong một ngữ cảnh cụ thể.

3.3. Các phương pháp tính độ tương đồng câu

Bài toán độ tương đồng ngữ nghĩa câu được sử dụng phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và có nhiều kết quả khả quan. Một số phương pháp được sử dụng để tính độ đo này như [SD08, LLB06, RFF05, STP06]:

- Phương pháp sử dụng thống kê: độ đo cosine, độ đo khoảng cách euclid …

- Phương pháp sử dụng các tập dữ liệu chuẩn về ngôn ngữ để tìm ra mối quan hệ giữa các từ: Wordnet, Brown Corpus, Penn TreeBank…

Các phương pháp tính độ tương đồng câu sử dụng kho ngữ liệu Wordnet được đánh giá cho ra kết quả cao. Tuy nhiên, kho ngữ liệu Wordnet chỉ hỗ trợ ngôn ngữ tiếng Anh, việc xây dựng kho ngữ liệu này cho các ngôn ngữ khác đòi hỏi sự tốn kém về mặt chi phí, nhân lực và thời gian. Nhiều phương pháp được đề xuất để thay thế Wordnet cho các ngôn ngữ khác, trong đó việc sử dụng phân tích chủ đề ẩn [Tu08] hay sử dụng mạng ngữ nghĩa Wikipedia để thay thế Wordnet [SP06, ZG07, ZGM07] được xem như là các phương án khả thi và hiệu quả. Các phương pháp này tập trung vào việc bổ sung các thành phần ngữ nghĩa hỗ trợ cho độ đo tương đồng Cosine.

3.3.1. Phương pháp tính độ tương đồng câu sử dụng độ đo Cosine

Trong phương pháp tính độ này, các câu sẽ được biểu diễn theo một mô hình không gian vector. Mỗi thành phần trong vector chỉ đến một từ tương ứng trong danh sách mục từ chính. Danh sách mục từ chính thu được từ quá trình tiền xử lý văn bản đầu vào, các bước tiền xử lý gồm: tách câu, tách từ, gán nhãn từ loại, loại bỏ những câu không hợp lệ (không phải là câu thực sự) và biểu diễn câu trên không gian vectơ.

Không gian vector có kích thước bằng số mục từ trong danh sách mục từ chính. Mỗi phần tử là độ quan trọng của mục từ tương ứng trong câu. Độ quan trọng của từ j được tính bằng TF như sau:

$$w_{i,j} = \frac{tf_{i,j}}{\sqrt{\sum_j tf_{i,j}^2}}$$

Trong đó, $tf_{i,j}$ là tần số xuất hiện của mục từ i trong câu j.

Với không gian biểu diễn tài liệu được chọn là không gian vector và trọng số TF, độ đo tương đồng được chọn là cosine của góc giữa hai vector tương ứng của hai câu S_i và S_k. Vector biểu diễn hai câu lần lượt có dạng:

$\mathbf{S_i} = \langle \mathbf{w_1^i}, \ldots, \mathbf{w_t^i} \rangle$, với w_t^i là trọng số của từ thứ t trong câu i

$\mathbf{S_k} = \langle \mathbf{w_1^k}, \ldots, \mathbf{w_t^k} \rangle$, với w_t^k là trọng số của từ thứ t trong câu k

Độ tương tự giữa chúng được tính theo công thức:

$$Sim(S_i, S_j) = \frac{\sum_{j=1}^{t} w_j^i w_j^k}{\sqrt{\sum_{j=1}^{t} \left(w_j^i\right)^2 \bullet \sum_{j=1}^{t} \left(w_j^k\right)^2}}$$

Trên các vector biểu diễn cho các câu lúc này chưa xét đến các quan hệ ngữ nghĩa giữa các mục từ, do đó các từ đồng nghĩa sẽ không được phát hiện, dẫn đến kết quả xét độ tương tự giữa các câu chưa tốt. Ví dụ như cho hai câu sau:

S_1 : Cần trao đổi ý kiến kỹ trước khi lấy biểu quyết.

S_2 : Hội đàm đã diễn ra trong bầu không khí thân mật và hiểu biết lẫn nhau.

Nếu không xét đến quan hệ ngữ nghĩa giữa các từ thì hai câu trên không có mối liên hệ gì cả và độ tương đồng bằng 0. Những thực chất, ta thấy rằng, từ "nhân loại" và từ "loài người" là đồng nghĩa, hai câu trên đều nói về loài người, do đó giữa hai câu có một sự liên quan nhất định và với công thức tính độ tương tự như trên thì độ tương tự giữa hai câu này phải khác 0.

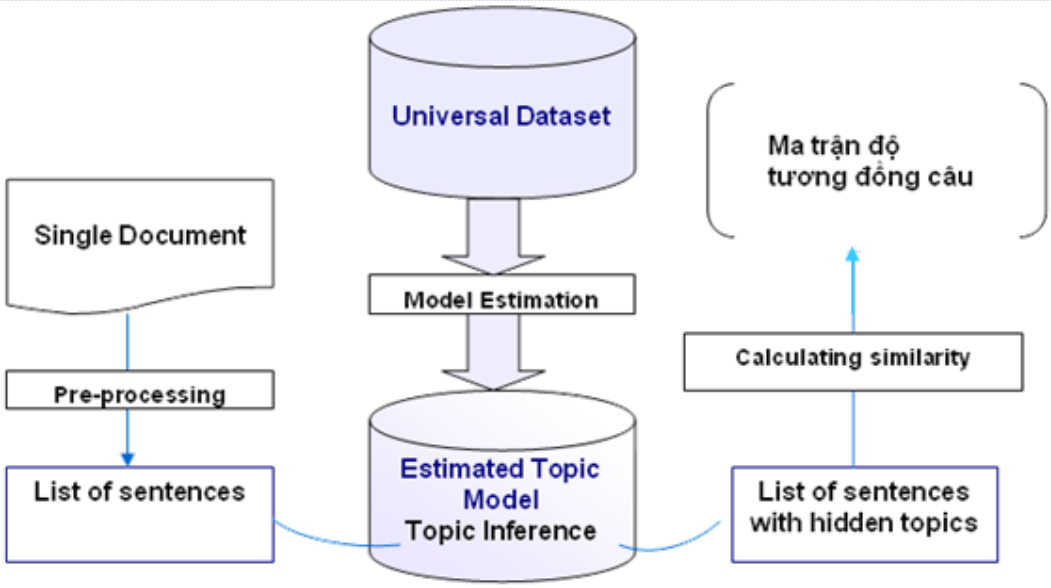
3.3.2. Phương pháp tính độ tương đồng câu dựa vào chủ đề ẩn

Phương pháp tiếp cận bài toán tính độ tương đồng câu sử dụng chủ đề ẩn dựa trên cơ sở các nghiên cứu thành công gần đây của mô hình phân tích topic ẩn LDA (Latent Dirichlet Allocation). Ý tưởng cơ bản của mô hình là với mỗi lần học, ta tập hợp một tập dữ liệu lớn được gọi là "Universal dataset" và xây dựng một mô hình học

trên cả dữ liệu học và một tập giàu các topic ẩn được tìm ra từ tập dữ liệu đó [Tu08, HHM08].

Mô hình độ tương đồng câu sử dụng chủ đề ẩn

Dưới đây là mô hình chung tính độ tương đồng câu với chủ đề ẩn:



Hình 3.1. Tính độ tương đồng câu với chủ đề ẩn

Mục đích của việc sử dụng chủ đề ẩn là tăng cường ngữ nghĩa cho các câu hay nói cách khác nghĩa của các câu sẽ được phân biệt rõ hơn thông qua việc thêm các chủ đề ẩn. Đầu tiên chọn một tập "universal dataset" và phân tích chủ đề cho nó. Quá trình phân tích chủ đề chính là quá trình ước lượng tham số theo mô hình LDA. Kết quả lấy ra được các chủ đề trong tập "universal dataset", các chủ đề này được gọi là chủ đề ẩn. Quá trình trên được thực hiện bên ngoài mô hình tính độ tương đồng câu với chủ đề ẩn.

Trong Hình 3.2, với đầu vào là một văn bản đơn, sau các bước tiền xử lý văn bản sẽ thu được một danh sách các câu. Tiếp theo, suy luận chủ đề cho các câu đã qua tiền xử lý, kết quả thu được một danh sách các câu được thêm chủ đề ẩn. Từ đây, có thể lần lượt tính toán độ tương đồng giữa các câu đã được thêm chủ đề ẩn.

Suy luận chủ đề và tính độ tương đồng các câu

Với mỗi câu, sau khi suy luận chủ đề cho câu sẽ nhận được các phân phối xác suất của topic trên câu và phân phối xác suất của từ trên topic. Tức là với mỗi câu i, LDA sinh ra phân phối topic $\vec{\vartheta}_i$ cho câu. Với mỗi từ trong câu, $z_{i,j}$ – topic index (từ j của câu i) - được lấy mẫu dựa theo phân phối topic trên. Sau đó, dựa vào topic index $z_{i,j}$ ta làm giàu các câu bằng cách thêm từ. Vector tương ứng với câu thứ i có dạng như sau: [Tu08]**Error! Reference source not found.**

$$s_i = \{t_1, t_2, ..., t_K, w_1, ..., w_{|V|}\}$$

Ở đây, t_i là trọng số của topic thứ i trong K topic đã được phân tích (K là một tham số hằng của LDA); w_i là trọng số của từ thứ i trong tập từ vựng V của tất cả các câu.

Mỗi câu có thể có nhiều phân phối xác suất topic. Với hai câu thứ i và j, chúng ta sử dụng độ đô cosine để tính độ tương đồng giữa hai câu đã được làm giàu với chủ đề ẩn.

$$sim_{i,j}(topic - parts) = \frac{\prod_{k=1}^{K} t_{i,k} \times t_{j,k}}{\sqrt{\sum_{k=1}^{K} t_{i,k}^2} \sqrt{\sum_{k=1}^{K} t_{j,k}^2}}$$

$$sim_{i,j}(word - parts) = \frac{\prod_{t=1}^{|V|} w_{i,t} \times w_{j,t}}{\sqrt{\sum_{t=1}^{|V|} w_{i,t}^2} \sqrt{\sum_{t=1}^{|V|} w_{j,t}^2}}$$

Cuối cùng, tổ hợp hai độ đo trên để ra độ tương đồng giữa hai câu:

$$sim(s_i, s_j) = \lambda \times sim(\text{topic} - \text{parts}) + (1 - \lambda) \times sim(\text{word - parts})$$

Trong công thức trên, λ là hằng số trộn, thường nằm trong đoạn [0,1]. Nó quyết định việc đóng góp giữa 2 độ đo tương đồng. Nếu $\lambda = 0$, độ tương đồng giữa hai câu không có chủ đề ẩn. Nếu $\lambda = 1$, đo độ tương đồng giữa hai câu chỉ tính với chủ đề ẩn [Tu08].

3.3.3. Phương pháp tính độ tương đồng câu dựa vào Wikipedia

Giới thiệu mạng ngữ nghĩa Wikipedia

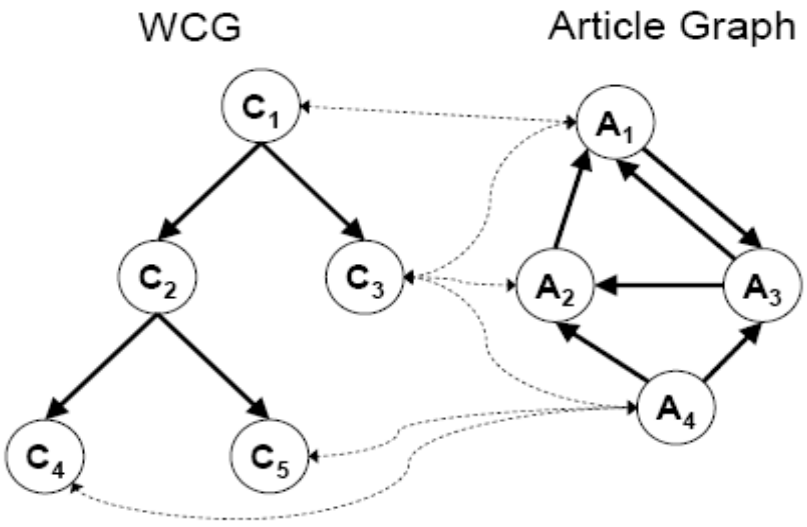
Wikipedia[1] là một bách khoa toàn thư nội dung mở bằng nhiều ngôn ngữ trên Internet. Wikipedia được viết và xây dựng do rất nhiều người dùng cùng cộng tác với nhau. Dự án này, nói chung, bắt đầu từ ngày 15 tháng 1 năm 2001 để bổ sung bách khoa toàn thư Nupedia bởi những nhà chuyên môn; hiện nay Wikipedia trực thuộc Quỹ Hỗ trợ Wikimedia, một tổ chức phi lợi nhuận. Wikipedia hiện có hơn 200 phiên bản ngôn ngữ, trong đó vào khoảng 100 đang hoạt động. 15 phiên bản đã có hơn 50.000 bài viết: tiếng Anh, Đức, Pháp, Ba Lan, Nhật, Ý, Thụy Điển, Hà Lan, Bồ Đào Nha, Tây Ban Nha, Hoa, Nga, Na Uy, Phần Lan, Esperanto và tiếng Việt, tổng cộng Wikipedia hiện có hơn 4,6 triệu bài viết, tính cả hơn 1,2 triệu bài trong phiên bản tiếng Anh (English Wikipedia).

Kiến trúc Wikipedia

Các trang thông tin của Wikipedia được lưu trữ trong một cấu trục mạng.Chi tiết hơn, các bài viết của Wikipedia được tổ chức dạng một mạng các khái niệm liên quan với nhau về mặt ngữ nghĩa và các mục chủ đề (category) được tổ chức trong một cấu trúc phân cấp(taxonomy) được gọi là đồ thị chủ đề Wikipedia (Wikipedia Category Graph - WCG).

Đồ thị bài viết(Article graph): Giữa các bài viết của Wikipedia có các siêu liên kết với nhau, các siêu liên kết này được tạo ra do quá trình chỉnh sửa bài viết của người sử dụng. Nếu ta coi mỗi bài viết như là một nút và các liên kết từ một bài viết đến các bài viết khác là các cạnh có hướng chạy từ một nút đến các nút khác thì ta sẽ có một đồ thị có hướng các bài viết trên Wikipedia (phía bên phải của hình 3.5).

[1] http://www.wikipedia.org



Hình 3.2. Mối quan hệ giữa đồ thị bài viết và đồ thị chủ đề Wikipedia

Đồ thị chủ đề (Category graph): Các chủ đề của Wikipedia được tổ chức giống như cấu trúc của một taxonomy (phía bên trái của hình 3.2). Mỗi một chủ đề có thể có một số lượng tùy ý các chủ đề con, mỗi một chủ đề con này thường được xác định bằng mối quan hệ thượng hạ vị (Hyponymy) hay mối quan hệ bộ phận tổng thể (Meronymy).

Ví dụ: Chủ đề vehicle có các chủ đề con là aircraft và watercraft

Do đó, đồ thị chủ đề (WCG) giống như là một mạng ngữ nghĩa giữa các từ tương tự như Wordnet. Mặc dù đồ thị chủ đề không hoàn toàn được xem như là một cấu trúc phân cấp do vẫn còn tồn tại các chu trình, hay các chủ đề không có liên kết đến các chủ đề khác tuy nhiên số lượng này là khá ít. Theo khảo sát của Torsten Zesch và Iryna Gurevych [ZG07] vào tháng 5 năm 2006 trên Wikipedia tiếng Đức thì đồ thị chủ đề chứa 99,8% số lượng nút chủ đề và chỉ tồn tại 7 chu trình.

Độ tương đồng giữa các khái niệm trong mạng ngữ nghĩa Wikipedia

Phương pháp tính độ tương đồng giữa các khái niệm trong mạng ngữ nghĩa Wikipedia được khá nhiều các nghiên cứu đưa ra như Ponzetto và cộng sự trong các năm 2006, 2007 [SP06, PSM07], Torsten Zesch và cộng sự năm 2007 [ZG07, ZGM07],…Các nghiên cứu này tập trung vào việc áp dụng và cải tiến một số độ đo

phổ biến về tính độ tương đồng từ trên tập ngữ liệu Wordnet cho việc tính độ tương đồng giữa các khái trên mạng ngữ nghĩa Wikipedia.

Cũng giống như trên Wordnet các độ đo này được chia thành hai loại độ đo, nhóm độ đo dựa vào **khoảng cách giữa các khái niệm** (Path based measure) như Path Length (PL, năm 1989), Leacock & Chodorow (LC, năm 1998), Wu and Palmer (WP, năm 1994) [ZG07, SP06] và nhóm độ đo dựa vào **nội dung thông tin** (Information content based measures) như Resnik (Res, năm 1995), Jiang and Conrath (JC, năm 1997), Lin (Lin, năm 1998) [ZG07]. Trong các độ đo này, trừ độ đo Path Length khi giá trị càng nhỏ thì độ tương đồng càng cao, còn lại các đô đo khác giá trị tính toán giữa 2 khái niệm càng lớn thì độ tương đồng càng cao.

- Độ đo Path Length (PL)

Độ đo PL được Rada và cộng sự đề xuất năm 1989 sử dụng độ dài khoảng cách ngắn nhất giữa hai khái niệm trên đồ thị (tính bằng số cạnh giữa hai khái niệm) để thể hiện sự gần nhau về mặt ngữ nghĩa.

$$dist_{PL} = l(n_1, n_2)$$

- n_1, n_2: là hai khái niệm cần tính toán
- $l(n_1,n_2)$: khoảng cách ngắn nhất giữa hai khái niệm

- Độ đo Leacock & Chodorow (LC)

Độ đo LC được Leacock và Chodorow đề xuất năm 1998 chuẩn hóa độ dài khoảng cách giữa hai node bằng độ sâu của đồ thị

$$sim_{LC}(n_1, n_2) = -\log \frac{l(n_1, n_2)}{2 \times depth}$$

- n_1, n_2: là hai khái niệm cần tính toán
- depth: là độ dài lớn nhất trên đồ thị
- $l(n_1,n_2)$: khoảng cách ngắn nhất giữa hai khái niệm

- Độ đo WP được Wu và Palmer đề xuất năm 1994:

$$sim_{WP} = \frac{2\,depth(lcs)}{l(n_1, lcs) + l(n_2, lcs) + 2\,depth(lcs)}$$

- c_1, c_2: là hai khái niệm cần tính toán
- lcs: Khái niệm thấp nhất trong hệ thống cấp bậc quan hệ is-a hay nó là cha của hai khái niệm n1 và n2
- depth(lcs): là độ sâu của khái niệm cha

- Độ đo Resnik được Resnik đề xuất 1995. Resnik đã coi độ tương đồng ngữ nghĩa giữa hai khai niệm được xem như nội dung thông tin trong nút cha gần nhất của hai khái niệm

$$res(c_1, c_2) = ic(lcs_{c_1,c_2})$$

Với c_1, c_2: là hai khái niệm cần tính toán và ic được tính như công thức ở dưới:

$$ic(n) = 1 - \frac{\log(hypo(n) + 1)}{\log(C)}$$

- hypo(n) là số các khái niệm có quan hệ thượng hạ vi (hyponym) với khái niệm n và C là tổng số các khái niệm có trên cây chủ đề

- Độ đo JC được Jiang và Conrath đề xuất năm 1997:

$$dist_{JC}(n_1, n_2) = IC(n_1) + IC(n_2) - 2IC(lcs)$$

- n_1, n_2: là hai khái niệm cần tính toán
- IC được tính như công thức ở trên

- Độ đo Lin được Lin đề xuất năm 1998:

$$sim_{Lin}(n_1, n_2) = 2 \times \frac{IC(lcs)}{IC(n_1) + IC(n_2)}$$

- n_1, n_2: là hai khái niệm cần tính toán
- IC được tính như công thức ở trên

Độ tương đồng câu dựa vào mạng ngữ nghĩa Wikipedia

Do các giá trị độ tương đồng được nêu ở trên đều không bị ràng buộc bởi khoảng 0,1, trong khi đó việc tính độ tương đồng câu theo phương pháp cosine đòi hỏi các thành phần thuộc khoảng này. Vào năm 2006, Li và cộng sự [LLB06] đã đưa ra hai công thức cải tiến độ tương đồng từ mà không làm mất tính đơn điệu.

- Đối với độ đo PL, f là một hàm đơn điệu giảm, vì vậy:

$$f_i(l) = \mathrm{e}^{\alpha l}$$

- Đối với các độ đo khác, f là một hàm đơn điệu tăng, vì vậy:

$$f_2(h) = \frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}}$$

Trong hai hàm số trên, α và β là hai tham số được chọn là α =0.2 và β=0.45

Sau khi tính được độ tương tự từ, ta đưa ra được vector ngữ nghĩa s_i cho mỗi câu. Giá trị của từng thành phần có trong vector là giá trị cao nhất về độ tương tự từ giữa từ trong tập từ chung tương ứng với thành phần của vector với mỗi từ trong câu [LLB06].

Sự giống nhau về ngữ nghĩa giữa 2 câu là hệ số cosine giữa 2 vector :

$$S_s = \frac{s_1 . s_2}{\| s_1 \| . \| s_2 \|}$$

3.4. Tóm tắt chương ba

Trong chương này, luận văn đã giới thiệu khái niệm về độ tương đồng câu, phương pháp xây dựng độ tương câu và một số giải pháp nhằm tăng cường tính ngữ nghĩa cho độ tương đồng câu. Trong chương tiếp theo, luận văn đi sâu vào đề xuất của tác giả cho việc tính độ tương đồng câu trong tiếng Việt và mô hình tóm tắt đa văn bản tiếng Việt.

Chương 4. Một số đề xuất tăng cường tính ngữ nghĩa cho độ tương đồng câu và áp dụng vào mô hình tóm tắt đa văn tiếng Việt

4.1. Đề xuất tăng cường tính ngữ nghĩa cho độ tương đồng câu tiếng Việt

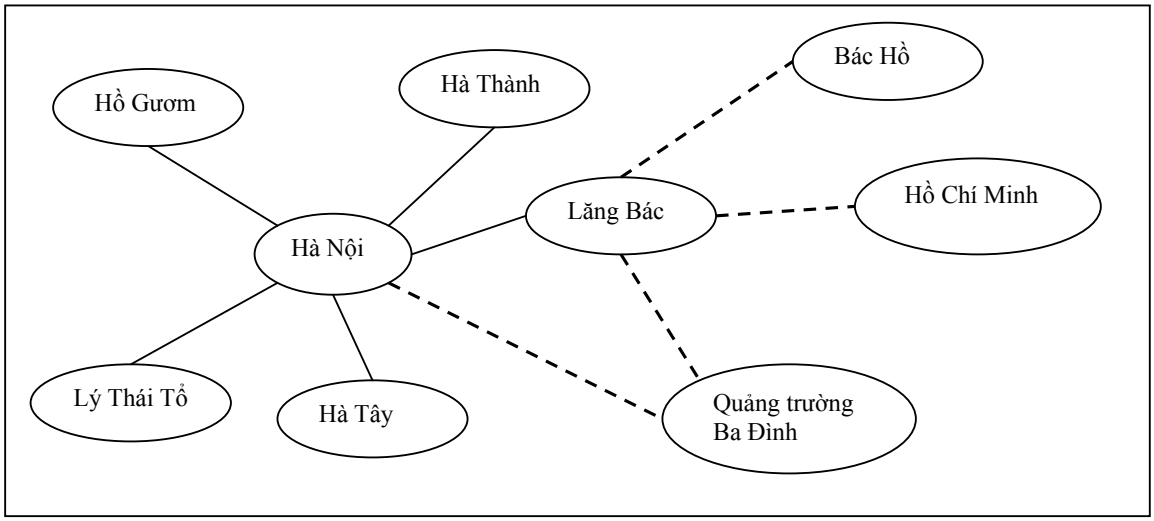
Việc xây dựng các độ đo tương đồng ngữ nghĩa có độ chính xác cao thường đòi hỏi cần có các kho ngữ liệu ngôn ngữ học thể hiện được mối quan hệ ngữ nghĩa giữa các từ, các khái niệm hay các thực thể như Wordnet hoặc Brown Corpus. Trong khi đó, đối với xử lý ngôn ngữ tự nhiên tiếng Việt hiện nay, các kho ngữ liệu ngôn ngữ học như vậy vẫn chưa được xây dựng hoàn chỉnh. Chính vì vậy, việc tìm ra phương pháp để xây dựng các kho ngữ liệu tương tự với chi phí thấp nhất trở thành một vấn đề đặt ra đối với cộng đồng xử lý ngôn ngữ tự nhiên tiếng Việt.

Cùng với việc nghiên cứu áp dụng hai phương pháp đã được đề cập ở mục 3.3.2 và mục 3.3.4 cho tiếng Việt là phân tích chủ đề ẩn và xây dựng mạng ngữ nghĩa Wikipedia, tác giả cũng đã nghiên cứu và đề xuất ra một phương pháp cho phép xây dựng đồ thị quan hệ giữa các thực thể (entities) dựa vào phương pháp học bán giám sát Bootstrapping trên máy tìm kiếm.

4.1.1. Đồ thị thực thể và mô hình xây dựng đồ thị quan hệ thực thể

Web ngữ nghĩa hay tìm kiếm thực thể là những đề tài lớn đang được nhiều nhà nghiên cứu quan tâm. Một trong những vấn đề đang được chú trọng hiện nay đó là làm thế nào để có thể từ một tập các thực thể, một tập các khái niệm hoặc một tập các thuật ngữ chuyên ngành có thể tìm kiếm và mở rộng ra được một tập lớn hơn, hoàn chỉnh hơn các thực thể, các khái niệm hay các thuật ngữ chuyên ngành khác mà có tương đồng ngữ nghĩa với tập gốc ban đầu.

Ví dụ: Trong Hình 4.1, yêu cầu đặt ra đối với bài toán mở rộng thực thể là tìm ra các mối quan hệ, các thực thể mới từ các thực thể có sẵn như mối quan hệ giữa Lăng Bác – Bác Hồ, Lăng Bác – Hồ Chí Minh, Lăng Bác – Quảng trường Ba Đình, Hà Nội – Quảng trường Ba Đình…

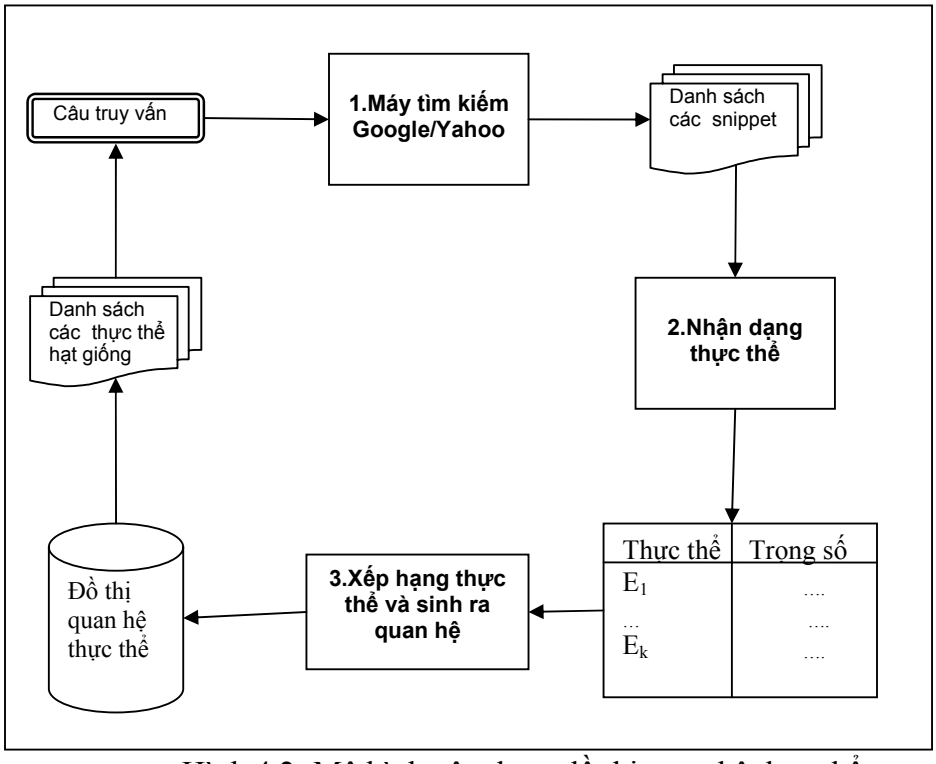


Hình 4.1. Mở rộng mối quan hệ và tìm kiếm các thực thể liên quan

Từ ý tưởng của bài toán mở rộng thực thể cũng như thông qua việc nghiên cứu khảo sát 2 mạng ngữ nghĩa Wordnet và Wikipedia, chúng tôi quan tâm tới việc xây dựng đồ thị thể hiện mối quan hệ giữa các thực thể với nhau và sử dụng đồ thị này như một mạng ngữ nghĩa để xây dựng độ đo tương đồng ngữ nghĩa câu. Mỗi một quan hệ giữa hai thực thể được xem như là một cạnh nối trực tiếp giữa hai nốt thực thể.

Dựa vào hai nghiên cứu về mở rộng thực thể dựa vào máy tìm kiếm của R.Wang và W.Cohen đưa ra năm 2007 [WC07] và độ đo tương đồng giữa các khái niệm dựa vào máy tìm kiếm của Bollegala đề xuất năm 2006 [BMI06], chúng tôi đưa ra mô hình xây dựng đồ thị quan hệ thực thể dựa vào máy tìm kiếm áp dụng giải thuật học bán giám sát Bootstrapping.

Dưới đây là mô hình xây dựng đồ thị quan hệ thực thể dựa vào máy tìm kiếm theo đề xuất của chúng tôi:



Hình 4.2: Mô hình xây dựng đồ thị quan hệ thực thể

Mô hình xây dựng đồ thị quan hệ thực thể gồm 3 pha chính:

- **Pha tương tác với các máy tìm kiếm(Google/Yahoo):**

Đưa một số thực thể từ đồ thị quan hệ thực thể đưa vào danh sách các thực thể hạt giống. Pha xử lý này nhận đầu vào một truy vấn được lấy ra từ tập các thực thể hạt giống (Seed) và đưa truy vấn này vào các máy tìm kiếm. Ví dụ: Hà Nội, Hồ Gươm,… Các máy tìm kiếm như Google/Yahoo sẽ trả về các snippet tương ứng với các câu truy vấn đưa vào.

- **Pha nhận dạng thực thể (NER):**

Tại pha xử lý này, các snippet sẽ được đưa qua công cụ nhận dạng thực thể để phát hiện các thực thể mới tồn tại trong snippet. Tại bước này, các công cụ nhận dạng thực thể đóng một vai trò quan trọng trong quá trình xây dựng đồ thị quan hệ thực thể. Trong Tiếng Anh đã có khá nhiều các công cụ sử dụng các giải thuật học máy cho

phép nhận dạng tên thực thể với độ chính xác cao như: Lingpipe Api[1], OpenNLP[2]…Tuy nhiên, trong tiếng Việt chưa tồn tại công cụ nào như vậy, tác giả đã sử dụng một số luật nhận dạng tên thực thể dựa vào biểu thức chính quy như: chọn các chuỗi ký tự mà mỗi từ được viết hoa và có độ dài lớn hơn hai từ… Sau khi có được tập các tên thực thể mới pha xử lý tiếp tục thống kê tần số xuất hiện của các tên thực thể đã có.

- **Pha nhận xếp hạng thực thể và sinh ra quan hệ:**

Trong pha này, tập các tên thực thể mới được sắp xếp lại theo tần số xuất hiện, dựa vào một ngưỡng lựa chọn đã xác định trước pha xử lý sẽ chọn ra các tên thực thể có tần số xuất hiện vượt ngưỡng cho phép để ghép với thực thể đầu vào thành một quan hệ. Các thực thể mới và mối quan hệ sẽ được thêm vào đồ thị có sẵn được lưu trữ trong cơ sở dữ liệu.

Mô hình này sẽ được lặp liên tục cho đến khi không có một quan hệ mới nào được sinh ra. Các thực thể mồi trong vòng lặp lần đầu tiên được đưa vào bằng tay. Các thực thể đã được từng đưa vào pha truy vấn máy tìm kiếm sẽ được đánh dấu để không đưa vào trong các lần sau.

4.1.2. Độ tương đồng ngữ nghĩa câu dựa vào đồ thị quan hệ thực thể

Thông qua việc nghiên cứu và xem xét sự tương quan giữa đồ thị quan hệ thực thể do tác giả đề xuất và hai mạng ngữ nghĩa Wordnet và Wikipedia cùng một số độ đo tương đồng ứng dụng trên hai mạng ngữ nghĩa đã được đề xuất ở mục 3.3.3, chúng tôi đã đề xuất một độ tương đồng ngữ nghĩa dựa vào đồ thị thực thể.

Sự tương quan giữa đồ thị quan hệ thực thể và mạng ngữ nghĩa Wordnet, Wikipedia

[1] Lingpipe Api. http://alias-i.com/lingpipe
[2] OpenNLP. http://opennlp.sourceforge.net

	Wordnet	Wikipedia	Đồ thị thực thể
Đồ thị quan hệ giữa các khái niệm	Có	Có	Có
Cây phân cấp chủ đề	Có	Có	Không
Nội dung thông tin tại các khái niệm	Có	Có	Không
Loại quan hệ giữa các khái niệm	Bao gồm hầu hết các quan hệ giữa hai từ/thực thể/khái niệm	Quan hệ thượng hạ vị, quan hệ bộ phận tổng thể, quan hệ tương đồng	Quan hệ tương đồng
Ngôn ngữ	Tiếng Anh	265 ngôn ngữ	Tiếng Anh, Tiếng Việt

Bảng 4.1: Sự tương quan giữa đồ thị quan hệ thực thể, Wordnet và Wikipedia

Độ tương đồng ngữ nghĩa dựa vào đồ thị quan hệ thực thể

Dựa vào sự xem xét tương quan được nêu ở bảng 4.1, chúng tôi nhận thấy việc xây dựng độ tương đồng ngữ nghĩa dựa vào đồ thị quan hệ thực thể chỉ có thể áp dụng nhóm các độ đo tương đồng dựa vào khoảng cách giữa các khái niệm (Path length measures). Độ đo tương đồng thực thể được chúng tôi đề xuất dựa trên độ đo LC (Leacock & Chodorow) như đã được trình bày ở chương 3:

$$sim_{LC}(n_1, n_2) = -\log \frac{l(n_1, n_2)}{2 \times depth}$$

trong đó:

- **n_1, n_2**: là hai thực thể cần tính toán trên đồ thị
- **depth**: là độ dài lớn nhất trên đồ thị được tính từ các thực thể mồi lúc khởi tạo hệ thống đến thực thể (nút) có khoảng cách xa nhất so với các nút này.

- **$l(n_1,n_2)$**: khoảng cách ngắn nhất giữa hai thực thể.

Áp dụng công thức tính độ tương đồng câu tại mục 3.3.3 của Li và các cộng sự trong năm 2006 [LLB06] để xây dựng độ tương đồng câu cho đồ thị quan hệ thực thể.

Nhận xét:

Mặc dù, đồ thị quan hệ thực thể không có nhiều thông tin trong mỗi nút thực thể cũng như việc phân loại chủ đề cho các thực thể trong đồ thị. Mặc dù vậy, đây là một phương pháp tự động giảm thiểu được chi phí xây dựng kho ngữ liệu cũng như có thể tạo ra được một đồ thị có số lượng nút thực lớn và mở rộng nhanh.

Độ đo tương đồng ngữ nghĩa cậu dựa vào đồ thị quan hệ thực thể chỉ hạn chế trong việc áp dụng các độ đo khoảng cách tuy nhiên nó có thể dễ dàng kết hợp với các độ đo tương đồng ngữ nghĩa khác thông qua các hàm trộn giữa các độ đo.

4.2. Độ tương đồng ngữ nghĩa câu tiếng Việt

Thông thường, để xây dựng các độ đo tương đồng ngữ nghĩa tốt, phương pháp phổ biến là sử dụng việc kết hợp nhiều độ đo lại với nhau thông qua một hàm tính hạng tuyến tính. Công thức biểu diễn việc kết hợp các độ đo như sau:

$$SimTotal(s_1, s_2) = \sum_i \alpha_i * sim_i(s_1, s_2)$$

Với điều kiện: $\sum_i \alpha_i = 1$

Trong đó:

- **s_1, s_2**: là hai câu cần tính độ tương đồng
- **i**: là số lượng các độ đo tương đồng kết hợp lại
- **sim_i**: là các độ đo tương đồng thành phần
- **α_i**: là các hằng số trộn nằm trong ngưỡng [0,1] thể hiện sự đóng góp của các độ đo tương đồng thành phần với độ đo SimTotal. Các tham số này

phải thỏa mãn điều kiện, tổng tất cả các hằng số trong công thức bằng 1 (Các hằng số này sẽ được ước lượng trong quá trình thực nghiệm).

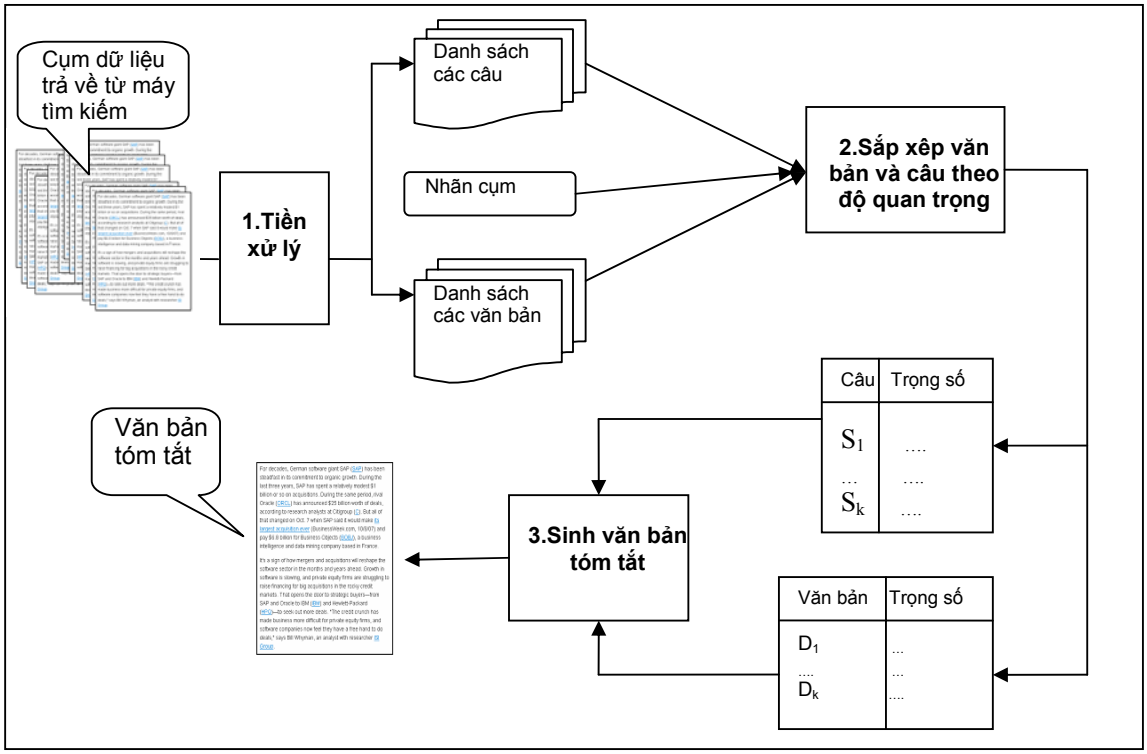
Dưới đây là các độ đo được sử dụng để tiến hành đánh giá, tìm ra độ đo tương đồng ngữ nghĩa phù hợp nhất với tiếng Việt. Trong các độ đo này, độ 5 và 6 là các độ đo kết hợp.

STT	Độ đo	Mô tả	Hằng số trộn được chọn qua thực nghiệm
1	Cosine [Cos]	Độ tương đồng Cosine	-
2	Hidden topic [Hidden]	Độ tương đồng dựa vào chủ đề ẩn kết hợp cosine	α_{Cos}=0.6 α_{Hidden}=0.4
3	Wikipedia [Wiki]	Độ tương đồng dựa vào mạng ngữ nghĩa Wikipedia	-
4	Entity Graph [EntG]	Độ tương đồng dựa vào đồ thị quan hệ thực thể	-
5	Hidden topic & Wikipedia & Entity Graph [All_1]	Độ tương đồng kết hợp 3 độ đo 1,2,3	α_{Cos}=0.3 α_{Hidden}=0.3 α_{Wiki}=0.2 α_{EntG}=0.2
6	Hidden topic & Wikipedia & Entity Graph & Dictionary [All_2]	Độ tương đồng kết hợp 3 độ đo 1,2,3 và độ tương đồng dựa vào từ điển đồng nghĩa	α_{Cos}=0.3 α_{Hidden}=0.2 α_{Wiki}=0.2 α_{EntG}=0.2 $\alpha_{Dictionary}$=0.1

Bảng 4.2. Danh sách các độ đo tương đồng ngữ nghĩa câu

4.3. Mô hình tóm tắt đa văn bản tiếng Việt

Từ những nghiên cứu được nêu ở các mục trên, tác giả đã đưa ra một mô hình tóm tắt đa văn bản cho các cụm dữ liệu trang web tiếng Việt trả về từ máy tìm kiếm.



Hình 4.3. Mô hình tóm tắt đa văn bản tiếng Việt

Mô hình tóm tắt đa văn bản tiếng Việt nhận đầu vào là các cụm dữ liệu trang web tiếng Việt được trả về từ quá trình phân cụm trên máy tìm kiếm. Mỗi cụm dữ liệu có nhãn của cụm và các trang web có nội dung liên quan đến nhãn cụm. Mỗi một trang web được coi như là một tài liêu. Mô hình tóm tắt gồm ba pha chính:

- **Pha tiền xử lý dữ liệu**

Pha xử lý này nhận đầu vào tập các trang web thuộc một cụm dữ liệu. Các quá trình được thực hiện theo các bước sau:

- Loại bỏ các trang web có nội dung trùng lặp.
- Lọc nhiễu, loại bỏ các thẻ HTML, lấy nội dung chính của trang Web.
- Tách từ, tách câu các văn bản có được bằng công cụ JvnTextpro của tác giả Nguyễn Cẩm Tú.
- Tách từ đối với nhãn cụm.

- **Pha sắp xếp văn bản và câu theo độ quan trọng**

Pha này nhận dữ liệu đầu vào là các văn bản và nhãn cụm đã qua tiền xử lý, đầu ra là danh sách các câu, các văn bản đã được sắp xếp lại theo độ quan trọng về mặt ngữ nghĩa.

Việc sắp xếp các văn bản và câu theo độ quan trọng bên cạnh việc loại bỏ sự chồng chéo giữa các văn bản là một bước quan trọng trong mô hình tóm tắt đa văn bản. Trong mô hình này, phương pháp được sử dụng để sắp xếp lại văn bản và câu là sự kết hợp của các nghiên cứu được nêu ra tại mục 2.4.1 và 2.4.2 với các độ đo tương đồng ngữ nghĩa được nêu ở mục 4.2.

- **Pha sinh văn bản tóm tắt**

Trong pha sinh văn bản tóm tắt, các câu được sắp xếp đã được sắp xếp ở pha trên sẽ được sắp xếp lại. Trọng số độ quan trọng của câu sẽ được bổ sung thêm trọng số của văn bản chưa câu đấy, việc này sẽ giúp văn bản tóm tắt không có sự chồng chéo về mặt nội dung. ScoreTotal là công thức tính lại độ quan trọng của câu:

$$ScoreTotal(s_k) = (\lambda * Score(s_k) + (1-\lambda) * \underset{s_k \in D_i}{Score}(D_i))$$

- $\mathbf{S_k}$: là câu cần tính độ quan trọng.
- $\mathbf{D_i}$: là văn bản chưa s_k.
- **Score($\mathbf{s_k}$), Score($\mathbf{D_i}$)**: là trọng số độ quan trọng của s_k và D_i được tính ở pha trước.
- **λ**: là các hằng số trộn nằm trong ngưỡng [0,1] thể hiện sự đóng góp của hai độ đo Score(s_k) và Score(D_i) (Các hằng số này sẽ được ước lượng trong quá trình thực nghiệm).

Sau khi đã có độ quan trọng câu, các câu sẽ được sắp xếp theo thứ tự từ lớn đến nhỏ theo độ đo ScoreTotal, trích số lượng các câu có độ quan trọng cao nhất theo tỷ lệ cho trước. Các câu sau khi được trích ra sẽ được sắp xếp vào trong một văn bản theo trình tự ưu tiên sau đây:

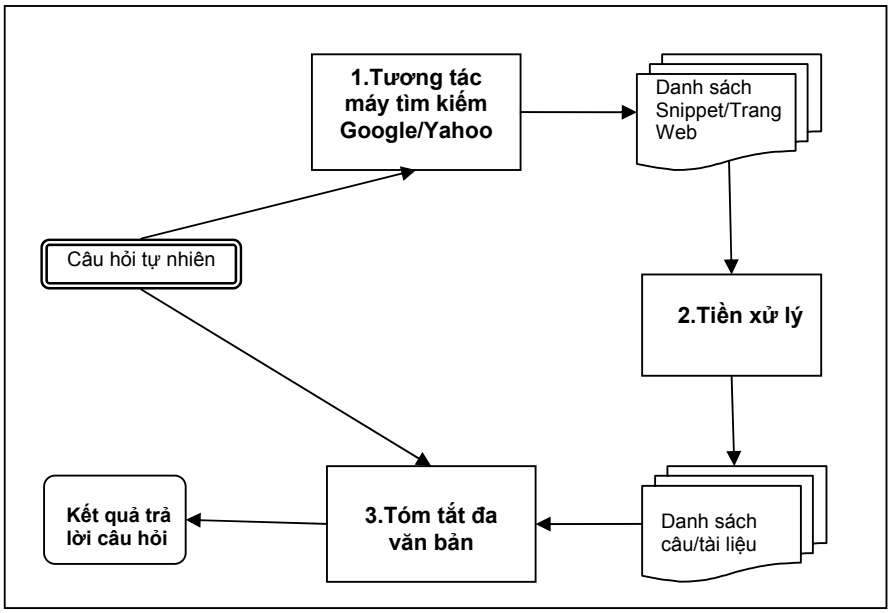
- Ưu tiên các câu thuộc văn bản có độ đo Score(D_i) cao hơn sẽ được xếp lên đầu văn bản.

- Ưu tiên theo thứ tự câu từ trên xuống dưới trong cùng một văn bản.

4.4. Mô hình hỏi đáp tự động tiếng Việt áp dụng tóm tắt đa văn bản

Một trong những vấn đề nhận được sự quan tâm của cộng đồng nghiên cứu tóm tắt đa văn bản là việc ứng dụng tóm tắt đa văn bản để xây dựng hệ thống hỏi đáp tự động(Question Answering System). Các nghiên cứu này sử dụng tóm tắt đa văn bản để tìm ra các câu trả lời trong một tập dữ liệu tri thức nền. Bên cạnh việc sinh các văn bản trả lời cho câu hỏi, các nghiên cứu này cũng giúp cho việc đánh giá các mô hình tóm tắt đa văn bản được dễ dàng và khách quan hơn. Thay vì cần có các chuyên gia ngôn ngữ học để đánh giá độ chính xác của các văn bản sinh ra từ mô hình tóm tắt, việc đánh giá bây giờ chỉ còn là việc xác định xem câu trả lời có trả lời chính xác câu hỏi đưa vào hay không.

Qua quá trình khảo sát kết quả trả về từ các máy tìm kiếm như Google, Yahoo đối với các một số câu hỏi tự nhiên, tác giả nhận thấy sự tồn tại của các câu trả lời trong danh sách các snippet hay các trang web trả về. Chính từ nhận định trên, tác giả đã đề xuất mô hình hỏi đáp tự động tiếng Việt dựa trên việc tóm tắt đa văn bản các kết quả trả về từ máy tìm kiếm để tìm ra kết quả trả lời.



Hình 4.4. Mô hình hỏi đáp tự động tiếng Việt áp dụng tóm tắt đa văn bản

Mô hình hỏi đáp tự động tiếng Việt gồm 3 pha chính:

- **Pha tương tác với máy tìm kiếm:**

Pha này nhận câu hỏi tự nhiên của người sử dụng, tiến hành tách từ và biến đổi thành câu truy vấn đưa vào các máy tìm kiếm Google và Yahoo. Các snippet, trang web tiếng Việt trả về từ máy tìm kiếm sẽ được tải về và đưa qua pha tiền xử lý.

- **Pha tiền xử lý:** Các bước xử lý tại pha này:
  - Lọc nhiễu, loại bỏ các thẻ HTML, lấy nội dung chính của trang Web.
  - Tách từ, tách câu các văn bản có được từ trang web và snippet
- **Tóm tắt đa văn bản:**

Pha này sử dụng mô hình tóm tắt đa văn bản tiếng Việt được nêu ở mục 4.3 với đầu vào là câu hỏi tự nhiên được xem như nhãn cụm và tập các văn bản trích xuất từ trang web qua pha tiền xử lý được xem như cụm dữ liệu. Kết quả đầu ra của mô hình tóm tắt sẽ là câu có trọng số cao nhất qua trình sắp xếp, câu này được xem như là câu trả lời cho câu hỏi.

4.5. Tóm tắt chương bốn

Trong chương này, luận văn đã trình bày các đề xuất của tác giả trong việc xây dựng độ tương đồng ngữ nghĩa câu cho tiếng Việt, mô hình tóm tắt đa văn bản và mô hình hỏi đáp tự động áp dụng tóm tắt đa văn bản. Trong chương tiếp theo, luận văn sẽ trình bày các thực nghiệm để chứng minh tính khả thi và triển vọng của bài toán tóm tắt đa văn bản cho tiếng Việt và mô hình hệ thống hỏi đáp tiếng Việt.

Chương 5. Thực nghiệm và đánh giá

5.1. Môi trường thực nghiệm

Quá trình thực nghiệm của luận văn được thực hiện trên máy tính có cấu hình:

- Chip: Intel Core 2 Duo 2.53 Ghz x 2

- Ram: 3 GB

- Hệ điều hành: Windows Vista

- Phần mềm lập trình: MyEclipse 7.5, Java 1.6

Các công cụ phần mềm và nguồn mở được liệt kê trong bảng dưới đây:

STT	Tên phần mềm	Mô tả
1	**JSum**	**Tác giả: Trần Mai Vũ** **Công dụng:** Công cụ có 2 nhóm chức năng chính là: - Xây dựng mạng ngữ nghĩa Wikipedia và đồ thị quan hệ thực thể - Tóm tắt đa văn bản dựa trên các độ đo tương đồng ngữ nghĩa như: suy luận chủ đề ẩn, mạng ngữ nghĩa wikipedia, đồ thị thực thể, ontology
2	**VQA**	**Tác giả: Trần Mai Vũ và Nguyễn Đức Vinh** **Công dụng:** Hệ thống hỏi đáp tiếng Việt dựa trên 2 phương pháp: tóm tắt đa văn bản và trích xuất quan hệ ngữ nghĩa [VVU09]
3	**JVnTextpro**	**Tác giả: Nguyễn Cẩm Tú** **Công dụng:** Tách từ, tách câu đối với các văn bản tiếng Việt

4	**JGibbsLDA**	**Tác giả: Nguyễn Cẩm Tú** **Công dụng:** Xây dựng và phân tích chủ đề ẩn
5	**Mulgara**	**Tác giả: Northrop Grumman Corporation** **Website:** http://www.mulgara.org **Công dụng:** Lưu trữ các mạng ngữ nghĩa Wikipedia và đồ thị quan hệ thực thể trên nền tảng công nghệ semantic web
6	**Lingpipe**	**Tác giả: Alias-i** **Website:** http://alias-i.com/lingpipe **Công dụng:** Nhận dạng tên thực thể (NER) trong tiếng Anh

Bảng 5.1. Các công cụ phần mềm sử dụng trong quá trình thực nghiệm

5.2. Quá trình thực nghiệm

5.2.1. Thực nghiệm phân tích chủ đề ẩn

- **Dữ liệu phân tích chủ đề ẩn:**
  - Bộ dữ liệu 125 topic (vnexp-lda4-125topics) đã được phân tích bằng JGibbsLDA trên kho dữ liệu các bài báo thu thập từ trang web Vnexpress

Sau quá trình phân tích chủ đề ẩn các câu sẽ được xác định nằm trong các chủ đề đã xác định trước trong bộ dữ liệu chủ đề ẩn.

Ví dụ:

STT	Câu	Các chủ đề trong câu
1	Cắt giảm thuế	Topic_48 Topic_97
2	Tiếp tục giảm thuế nhiều mặt hàng nhập khẩu	Topic_97
3	Những mặt hàng nằm trong diện cắt giảm thuế trong thời gian tới gồm rượu, bia, thuốc lá, cà	Topic_16 Topic_33 Topic_54 Topic_62 Topic_97 Topic_106

	phê, dầu thực vật, thịt chế biến...	Topic_123
4	Theo yêu cầu của Chính phủ Liên bộ Tài chính – Công thương tiếp tục thực hiện lộ trình giá thị trường đối với mặt hàng chiến lược có sự kiểm soát của Nhà Nước, nhằm khuyến khích cạnh tranh, hạn chế độc quyền.	Topic_13 Topic_33 Topic_41 Topic_47 Topic_67 Topic_78 topic_105 Topic_105 Topic_115 Topic_122

Bảng 5.3. Kết quả phân tích chủ đề ẩn

Dễ dàng nhận thấy các câu trên có nội dung liên quan đến chủ đề "Thuế" đều thấy xuất hiện Topic_97 quá trình phân tích chủ đề.

Dưới đây là 20 từ có phân phối xác suất cao trong Topic_97:

Topic 97:	
1. thương_mại 0.051798	11. kinh_tế 0.010271
2. wto 0.038748	12. hiệp_định 0.010070
3. đàm_phán 0.028651	13. phát_triển 0.009695
4. gia_nhập 0.021578	14. tự_do 0.009162
5. thành_viên 0.017416	15. tổ_chức 0.007909
6. nhập_khẩu 0.015039	16. dệt 0.007175
7. cam_kết 0.014520	17. asean 0.007131
8. thuế 0.013109	18. đạt 0.007117
9. xuất_khẩu 0.011164	19. bộ_trưởng 0.006872
10. vấn_đề 0.010848	20. nông_nghiệp 0.006757

Bảng 5.4: 20 từ có phân phối xác suất cao trong Topic ẩn 97

5.2.2. Thực nghiệm xây dựng đồ thị quan hệ thực thể

- **Dữ liệu xây dựng đồ thị quan hệ thực thể:**
  - Dữ liệu mồi: 200 thực thể tiếng Việt và 200 thực thể tiếng Anh thuộc các lĩnh vực: Địa danh, tổ chức, nhân vật.

Thực nghiệm là kết quả của quá trình thực thi mô hình xây dựng đồ thị quan hệ thực thể được đề xuất tại mục 4.1.1 đã được cài đặt. Trong thực nghiệm này, đồ thị

quan hệ thực thể được xây dựng cho 2 ngôn ngữ tiếng Anh và tiếng Việt. Phương pháp nhận dạng tên thực thể(NER) được áp dụng mô hình này:

Đối với tiếng Anh: mô hình học máy CRF, sử dụng bộ công cụ Lingpipe Api.

Đối với tiếng Việt: sử dụng biểu thức chính quy.

Ngôn ngữ	Số lượng thu được	Số lượng quan hệ	Thời gian thực thi
Tiếng Anh	48.365 thực thể	72.619 quan hệ	5 ngày
Tiếng Việt	21.693 thực thể	32.774 quan hệ	5 ngày

Bảng 5.5. Kết quả dữ liệu thu được của mô hình xây dựng đồ thị quan hệ thực thể

5.2.3. Thực nghiệm đánh giá các độ đo tương đồng

- **Dữ liệu Wikipedia:**
  - 99.679 bài viết trên Wikipedia Tiếng Việt (23/10/2009)
  - Download tại địa chỉ: http://download.wikimedia.org/viwiki/20091023
- **Dữ liệu từ điển:**
  - Từ điển đồng nghĩa: gồm 2393 nhóm từ đồng nghĩa được phát triển dựa trên "Từ điển đồng nghĩa" của Nguyễn Văn Tu, NXB Đại học và Trung học chuyên nghiệp, 1985.
- **Dữ liệu đánh giá độ đo tương đồng ngữ nghĩa câu:**
  - Sử dụng 20 cụm: mỗi cụm gồm 3-5 cặp câu, được đánh giá bằng tay theo thứ tự về độ tương đồng về mặt ngữ nghĩa (Thứ tự càng thấp độ tương đồng càng cao).

Ví dụ:

Số thứ tự	Câu thứ nhất	Câu thứ hai	Xếp hàng bằng tay
1	Tôi thích Hà Nội	Anh yêu Hồ Gươm	1
2	Tôi thích Hà Nội	Em mến người Hà Thành	2

3	Tôi thích Hà Nội	Cô ấy ngắm nhìn Tháp rùa	3
4	Tôi thích Hà Nội	Bạn ấy thích Hà Giang	4

Bảng 5.6. Một cụm dữ liệu dùng để đánh giá độ tương đồng ngữ nghĩa

Trong thực nghiệm này, các độ đo tương đồng được đánh giá nêu trong bảng 4.2. Các bước thực nghiệm:

- Tính độ đo tương đồng giữa các cặp câu bằng các độ đo khác nhau, sắp xếp theo thứ tự càng gần về mặt ngữ nghĩa thì thứ tự càng thấp.
- Độ chính xác được tính bằng số lượng các câu giữ đúng thứ tự xếp hạng bằng tay đã được gán cho tập dữ liệu thực nghiệm.

Số thứ tự của câu	**Cos**	**EntG**	**Wiki**	**Hidden**	**All_1**	**All_2**
1	3	2	2	2	2	1
2	2	3	1	1	1	2
3	3	4	4	4	3	3
4	1	1	3	3	4	4

Bảng 5.7. Kết quả đánh giá các độ đo trên cụm dữ liệu ở bảng 5.2

Trong việc đánh giá trên 10 cụm tiếng Anh, tác giả chỉ sử dụng hai độ đo tương đồng là Cosine và đồ thị quan hệ thực để đánh giá.

Ngôn ngữ	**Cos**	**Hidden**	**Wiki**	**EntG**	**All_1**	**All_2**
Tiếng Việt	56%	72%	76%	69%	81%	89%
Tiếng Anh	68%	~	~	83%	~	~

Bảng 5.8. Độ chính xác đánh giá trên 20 cụm dữ liệu tiếng Việt và 10 cụm tiếng Anh

Kết quả thực nghiệm cho thấy việc độ đo tương đồng ngữ nghĩa All_2 cho kết quả tốt hơn các độ đo khác. Trong các thực nghiệm tiếp theo, tác giả sử dụng All_2 làm đô đo tương đồng ngữ nghĩa chính.

5.2.4. Thực nghiệm đánh giá độ chính xác của mô hình tóm tắt đa văn bản

- **Dữ liệu đánh giá độ mô hình tóm tắt đa văn bản:**
  - Sử dụng 5 cụm trả về từ quá trình phân cụm trên máy tìm kiếm tiếng Việt VnSen: mỗi cụm gồm 8-10 văn bản. Các văn bản trong cụm và 20 câu quan trọng nhất trong văn bản sẽ được sắp xếp bằng tay dựa vào độ tương đồng của giữa văn bản/câu với nhãn cụm.

Độ chính xác được tính bằng số lượng các văn bản/câu giữ đúng thứ tự xếp hạng bằng tay đã được gán cho tập dữ liệu thực nghiệm.

Cụm	Số lượng văn bản	Số lượng câu	Nhãn cụm	Độ chính xác thứ tự văn bản	Độ chính xác thứ tự của 20 câu quan trọng
1	10	216	Lãi suất tiết kiệm	80%	80%
2	8	116	Cắt giảm thuế	87.5%	85%
3	8	127	Công cụ tìm kiếm Google	87.5%	80%
4	8	101	Laptop giá rẻ	75%	75%
5	8	86	Dịch tiêu chảy	75%	70%

Bảng 5.9. Đánh giá kết quả thứ tự văn bản và thứ tự của 20 câu quan trọng nhất

Đối với cụm văn bản có nhãn "Lãi suất tiết kiệm", với tỷ lệ trích xuất là 10 câu, kết quả tóm tắt trả về theo đánh giá trực quan là tương đối tốt.

Văn bản tóm tắt
[8][7] Hôm qua, Dong A Bank thông báo tăng lãi suất tiền gửi tiết kiệm VND dành cho khách hàng cá nhân với mức tăng bình quân 0,06% mỗi tháng.
[9][2] "Lãi suất ngân hàng đang cao. Ai cũng muốn bán tháo cổ phiếu lấy tiền gửi tiết kiệm

nhưng không được, tôi phải vất vả lắm mới bán thành công", chị Phúc cười vui vẻ.
[1][1] Lãi suất tiết kiệm đụng mốc 15%
[10][1] Đổ xô đến ngân hàng gửi tiền ngắn hạn
[10][25] Tuy nhiên, nhiều nhà băng cũng ước đoán lượng gửi tiền với kỳ hạn ngắn sẽ chiếm ưu thế hơn so với gửi tiết kiệm lâu dài.
[10][4] Còn tại Ngân hàng Phương Đông, chị Linh đã chuẩn bị sẵn 70 triệu đồng từ cuối tuần để gửi tiết kiệm linh hoạt 12 tháng.
[2][23] Một lãnh đạo của ngân hàng VP nhận định: “Trong tuần này sẽ có nhiều biến động về lãi suất vì các ngân hàng theo dõi động thái của nhau để điều chỉnh kịp thời mức lãi suất. Chỉ có như vậy mới có thể giữ chân được khách hàng”.
[7][19] Mỗi tháng doanh nghiệp thanh toán lãi tháng cho nhà băng gần 10 triệu đồng.
[7][11] Lãi suất cho vay của các ngân hàng đang được điều chỉnh, cộng với tình hình một số nhà băng ngừng cho vay đã tác động tức thời đến các doanh nghiệp đang có nhu cầu vay tiền vào thời điểm này.
[7][1] Lâm thế kẹt vì ngân hàng điều chỉnh cho vay

Bảng 5.10. Kết quả tóm tắt trả về theo tỷ lệ trích xuất là 10 câu (*hai chỉ số đầu dòng tương ứng là thứ tự của văn bản trong cụm và thứ tự của câu trong văn bản*).

5.2.5. Thực nghiệm đánh giá độ chính xác của mô hình hỏi đáp

- **Dữ liệu đánh giá hệ thống hỏi đáp:**
  - Dữ liệu: 500 câu hỏi dịch có lựa chọn và chỉnh sửa từ bộ dữ liệu của TREC (Lấy từ bộ công cụ OpenEphyra). Các câu hỏi được đưa kiểm tra trước trên các máy tìm kiếm xem có xuất hiện câu trả lời trong các snippet trả về hay không.

Đô tương đồng	Số trả lời đúng	Độ chính xác	Thời gian trả lời trung bình
Cos	67	13.4%	30 giây

Hidden	238	47.6%	2 phút
Wiki	142	28.4%	25 phút
EntG	167	33.4%	15 phút
All_1	318	63.6%	35 phút
All_2	376	75.2%	40 phút

Bảng 5.11. Độ chính xác của mô hình hỏi đáp dựa vào tóm tắt đa văn bản cho snippet

Đô tương đồng	**Số trả lời đúng**	**Độ chính xác**	**Thời gian trả lời trung bình**
Cos	101	21.6%	2 phút
Hidden	356	71.2%	15 phút
Wiki	104	20.8%	45phút
EntG	125	25.0%	1 giờ 15 phút
All_1	359	71.8%	2 giờ 30 phút
All_2	389	77.8%	3 giờ
*Tốc độ trên không tính thời gian download trang web			

Bảng 5.12. Độ chính xác của mô hình hỏi đáp dựa vào tóm tắt đa văn bản cho trang web

Câu hỏi	**Câu trả lời**
Người đầu tiên tìm ra châu mỹ ?	Ai cũng biết Cô-lôm-bô là người đầu tiên tìm ra châu Mỹ
Nhạc sĩ sáng tác bài hát người hà nội ?	Người Hà Nội là một bài hát do nhạc sĩ Nguyễn Đình Thi sáng tác
Cà chua có tác dụng gì đối với sức khỏe ?	Cà chua có tác dụng phòng chống ung thư vú, ung thư dạ dày

Bác Hồ sang pháp năm nào ?	Mùa hè năm 1911, Bác đặt chân lên đất Pháp, đối với Bác
Người sáng lập ra google ?	Tờ Financial Times đã bình chọn hai nhà đồng sáng lập ra công cụ tìm kiếm Google, Sergey Brin và Larry Page, đều 32 tuổi là Người đàn ông của năm
…	…

Bảng 5.13. Danh sách một số câu kết quả trả lời của hệ thống hỏi đáp

Kết luận

Những vấn đề đã được giải quyết trong luận văn

Luận văn tiến hành nghiên cứu giải quyết bài toán tóm tắt đa văn bản tiếng Việt dựa vào trích xuất câu. Bài toán này được xác định là một bài toán có độ phức tạp cao và là nền tảng của nhiều ứng dụng thực tế. Phương pháp giải quyết của luận văn tập trung vào việc tăng cường tính ngữ nghĩa cho độ đo tương đồng giữa hai câu trong quá trình trích xuất câu quan trọng của tập dữ liệu đầu vào.

Dựa vào các nghiên cứu về chủ đề ẩn, mạng ngữ nghĩa Wikipedia và một phương pháp do tác giả luận văn đề xuất, luận văn đã đưa ra một độ đo tương đồng ngữ nghĩa câu để xây dựng mô hình tóm tắt đa văn bản tiếng Việt.

Hơn nữa, luận văn cũng đã trình bày mô hình hệ thống hỏi đáp tiếng Việt áp dụng tóm tắt đa văn bản sử dụng dữ liệu trên các máy tìm kiếm nổi tiếng như Google, Yahoo làm tri thức nền. Quá trình thực nghiệm đạt được kết quả khả quan, cho thấy tính đúng đắn của việc lựa chọn cũng như kết hợp các phương pháp, đồng thời hứa hẹn nhiều tiềm năng phát triển hoàn thiện.

Công việc nghiên cứu trong tương lai

- Phát triển và mở rộng đồ thị quan hệ thực thể, nghiên cứu và xây dựng cây phân cấp chủ đề thực thể cho đồ thị.
- Nghiên cứu và áp dụng một số giải thuật tính toán độ tương đồng ngữ nghĩa trên mạng ngữ nghĩa để cải tiến mô hình tóm tắt đa văn bản tiếng Việt.
- Cải tiến quá trình lưu trữ và đánh chỉ mục để tăng tốc cho các việc tìm kiếm và tính toán trên đồ thị, qua đó tăng tốc độ trả lời câu hỏi cho mô hình hỏi đáp tiếng Việt.
- Xây dựng và triển khai hệ thống hỏi đáp tiếng Việt cho người sử dụng.

Các công trình khoa học và sản phẩm đã công bố

[VVU09] Vu Tran Mai, Vinh Nguyen Van, Uyen Pham Thu, Oanh Tran Thi and Thuy Quang Ha (2009). *An Experimental Study of Vietnamese Question Answering System*, International Conference on Asian Language Processing (IALP 2009): 152-155, Dec 7-9, 2009, Singapore.

[VUH08] Trần Mai Vũ, Phạm Thị Thu Uyên, Hoàng Minh Hiền, Hà Quang Thụy (2008). *Độ tương đồng ngữ nghĩa giữa hai câu và áp dụng vào bài toán sử dụng tóm tắt đa văn bản để đánh giá chất lượng phân cụm dữ liệu trên máy tìm kiếm VNSEN*, Hội thảo Công nghệ Thông tin & Truyền thông lần thứ nhất (ICTFIT08): 94-102, ĐHKHTN, ĐHQG TP Hồ Chí Minh, Thành phố Hồ Chí Minh, 2008.

Sản phẩm phần mềm

[VTTV09] Trần Mai Vũ, Vũ Tiến Thành, Trần Đạo Thái, Nguyễn Đức Vinh (2009). *Máy tìm kiếm giá cả*, http://vngia.com

Tài liệu tham khảo

Tiếng Việt

[MB09] Lương Chi Mai và Hồ Tú Bảo (2009). *Báo cáo Tổng kết đề tài KC.01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt"* và *Về xử lý tiếng Việt trong công nghệ thông tin (2006),* Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ Việt Nam, 2009.

Tiếng Anh

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

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

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

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

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

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

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

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

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

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

[EWK] Website: http://en.wikipedia.org/wiki/Multi-document_summarization.

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

[GMC00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, Mark Kantrowitz (2000). *Multi-Document Summarization By Sentence Extraction*, 2000.

[HHM08] Phan Xuan Hieu, Susumu Horiguchi, Nguyen Le Minh (2008). *Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections*, In The 17th International World Wide Web Conference, 2008.

[HMR05] B. Hachey, G. Murray, D. Reitter (2005). *Query-Oriented Multi-Document Summarization With a Very Large Latent Semantic Space,* In The Embra System at DUC, 2005.

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

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

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

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

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

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

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

[Mi04] Nguyen Le Minh (2004). *Statistical Machine Learning Approaches to Cross Language Text Summarization*, PhD Thesis, School of Information Science Japan Advanced Institute of Science and Technology, September 2004.

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

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

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

[PSM07] Ponzetto, Simone Paolo, and Michael Strube (2007). *Knowledge Derived from Wikipedia For Computing Semantic Relatedness,* Journal of Artificial Intelligence Research, **30**: 181-212, 2007.

[Ra00] Dragomir Radev (2000). *A common theory of information fusion from multiple text sources, step one: Cross-document structure,* In 1st ACL SIGDIAL Workshop on Discourse and Dialogue, Hong Kong, October 2000.

[RFF05] Francisco J. Ribadas, Manuel Vilares Ferro, Jesús Vilares Ferro (2005). *Semantic Similarity Between Sentences Through Approximate Tree Matching*. IbPRIA (2): 638-646, 2005.

[RJS04] Dragomir R. Radev, Hongyan Jing, Malgorzata Sty´s, and Daniel Tam (2004). *Centroid-based summarization of multiple documents,* Information Processing and Management, **40**:919–938, December 2004.

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

[Sen07] Pierre Senellart (2007). *Understanding the Hidden Web*, PhD thesis, Université Paris-Sud, Orsay, France, December 2007.

[SP06] Strube, M. & S. P. Ponzetto (2006). *WikiRelate! Computing semantic relatedness using Wikipedia,* In Proc. of AAAI-06, 2006.

[STP06] Krishna Sapkota, Laxman Thapa, Shailesh Bdr. Pandey (2006). *Efficient Information Retrieval Using Measures of Semantic Similarity*, Conference on Software, Knowledge, Information Management and Applications: 94-98, Chiang Mai, Thailand, December 2006.

[Su05] Sudarshan Lamkhede. *Multi-document summarization using concept chain graphs*, Master Thesis, Faculty of the Graduate School of the State University of New York at Buffalo, September 2005.

[Tu08] Nguyen Cam Tu (2008). *Hidden Topic Discovery Toward Classification And Clustering In Vietnamese Web Documents,* Master Thesis, Coltech of Technology, Viet Nam National University, Ha Noi, Viet Nam, 2008.

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

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

[YYL07] J.-C. Ying, S.-J. Yen, Y.-S. Lee, Y.-C. Wu, J.-C. Yang (2007). *Language Model Passage Retrieval for Question-Oriented Multi Document Summarization*, DUC 07, 2007.

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

[ZGM07] Torsten Zesch, Iryna Gurevych, and Max Muhlhauser (2007). *Comparing Wikipedia and German Word-net by Evaluating Semantic Relatedness on Multiple Datasets,* In Proceedings of NAACL-HLT, 2007.