ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Đỗ Thị Thanh Tuyền

MÔ HÌNH TÌM KIẾM VĂN BẢN TIẾNG VIỆT DỰA TRÊN NGỮ NGHĨA

Chuyên ngành: **Khoa học máy tính**
Mã số: **62 48 01 01**

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:
1. PGS. TS. Nguyễn Tuấn Đăng
2. PGS. TS. Vũ Đức Lung

PHẢN BIỆN ĐỘC LẬP:
1. PGS. TS. Đỗ Thanh Nghị
2. TS. Ngô Quốc Việt

TP. HỒ CHÍ MINH - NĂM 2020

MỤC LỤC

MỞ ĐẦU

1. Lý do lựa chọn đề tài

Truy hồi tập văn bản phù hợp với ngữ nghĩa của truy vấn là mục tiêu quan trọng nhất của lĩnh vực truy hồi thông tin văn bản. Các nghiên cứu truy hồi văn bản được tiến hành theo hai hướng chính là chú giải ngữ nghĩa ("semantic annotation") [47], [56], [31], [22], [21], [24], [45], [11], [28] và mở rộng truy vấn ("query expansion") [66], [26], [40], [39], [27], [63], [58] hiện tại tập trung vào việc giải quyết nghĩa từ vựng (gồm từ và thuật ngữ) trong so khớp văn bản và truy vấn. Nghĩa của từ vựng có thể được biểu diễn bằng một nhãn nghĩa, hoặc một vector ngữ nghĩa ("word embedding") hoặc một biến tiềm ẩn ("latent variable"). Vì thế, thách thức đặt ra là giải quyết vấn đề truy hồi văn bản ở cấp độ ngữ nghĩa (là một trong các cấp độ phân tích ngôn ngữ tự nhiên). Ngữ nghĩa, theo ngôn ngữ học, bao gồm cả nghĩa của từ vựng và quan hệ phụ thuộc giữa các từ ngữ trong câu. Có ba vấn đề trong thách thức này gồm (1) xác định ngôn ngữ biểu diễn ngữ nghĩa (NN-BD-NN) cho các phát biểu trong ngôn ngữ tự nhiên để làm cơ sở cho các phép so khớp khi tính toán độ liên quan giữa văn bản và truy vấn, (2) biến đổi một phát biểu trong ngôn ngữ tự nhiên thành một phát biểu trong NN-BD-NN và (3) đề xuất mô hình có các thành phần và cơ chế xử lý phù hợp với NN-BD-NN. Khi đó, thay vì xử lý văn bản và truy vấn trong ngôn ngữ tự nhiên, mô hình sẽ xử lý trên ngữ nghĩa tương ứng của chúng để tính toán độ tương đồng.
Dựa trên những khảo sát về truy hồi thông tin về ngữ nghĩa, luận án xác định hướng tiếp chú giải ngữ nghĩa bằng cách kết hợp Ontology cho nghĩa từ tiếng Việt, Phân tích quan hệ phụ thuộc trong câu tiếng Việt và Mô hình truy hồi văn bản.

2. Mục đích của luận án

Mục đích của luận án là đề xuất giải pháp truy hồi văn bản trên cơ sở phân tích ngữ nghĩa của câu tiếng Việt. Để đạt được mục đích này, luận án phải giải quyết được ba vấn đề

chính: (1) Xác định NN-BD-NN, (2) Phân tích ngữ nghĩa câu tiếng Việt theo NN-BD-NN và (3) Truy hồi văn bản theo NN-BD-NN của văn bản và truy vấn.

3. Nội dung nghiên cứu

Luận án đề ra những nội dung cụ thể:

1. Khảo sát các dạng biểu diễn ngữ nghĩa, đề xuất NN-BD-NN và chứng tỏ NN-BD-NN thỏa yêu cầu biểu diễn ngữ nghĩa theo hướng tiếp cận ngữ nghĩa học chân – ngụy (truth-conditional semantics).
2. Nghiên cứu phương pháp phân tích ngữ nghĩa của tiếng Việt để biến đổi câu tiếng Việt theo dạng NN-BD-NN. Nội dung này gồm có ba bài toán con: (a) xây dựng hệ thống nhãn nghĩa cho từ tiếng Việt; (b) xác định nhãn nghĩa của từ tiếng Việt và (c) phân tích quan hệ phụ thuộc của các từ trong câu tiếng Việt.
3. Đề xuất mô hình truy hồi văn bản dựa trên NN-BD-NN.

4. Đối tượng nghiên cứu

Đối tượng nghiên cứu thứ nhất là hệ thống nhãn nghĩa cho từ tiếng Việt dùng để chú giải nghĩa cho từng từ tiếng Việt. Đối tượng nghiên cứu thứ hai là các quan hệ phụ thuộc giữa các từ trong câu tiếng Việt. Đối tượng nghiên cứu thứ ba là các câu tiếng Việt. Đối tượng nghiên cứu thứ tư là văn bản có một chủ đề cụ thể, không có hiện tượng chuyển mạch ý.

5. Phạm vi nghiên cứu

- Phạm vi nghiên cứu về xử lý ngôn ngữ tự nhiên:
  - Phân tích ngữ nghĩa dựa trên ngữ pháp phụ thuộc [51] và các ràng buộc ngữ nghĩa được xác định từ kết quả phân tích phụ thuộc thủ công cho các câu thực tế. Việc phân tích áp dụng cho ngữ đoạn hoặc câu riêng lẻ mà không phân tích ngữ nghĩa diễn ngôn của văn bản.
  - Ngữ nghĩa của câu hoặc ngữ đoạn được phân tích là nghĩa của từ và mối quan hệ phụ thuộc giữa chúng trong câu [52, 53] không xử lý hàm ý, ẩn ý.
  - Văn bản đảm bảo tính liền lạc (cohesion), không có hiện tượng chuyển mạch ý.

- Phạm vi nghiên cứu về truy hồi thông tin:
    - Nghiên cứu mô hình truy hồi văn bản dựa trên mô hình căn bản với ba thành phần chính là thành phần phân tích văn bản, thành phần lập chỉ mục và thành phần so khớp phù hợp với NN-BD-NN.
    - Cấu trúc chỉ mục được đề xuất ở mức logic, không đặt vấn đề cài đặt, tối ưu và nén chỉ mục.

6. Ý nghĩa khoa học và thực tiễn của đề tài

Luận án có những đóng góp khoa học chính:

1. Đề xuất mô hình truy hồi văn bản có cơ chế so khớp ở mức ngữ nghĩa của câu.
2. Đề xuất mô hình ontology cho nghĩa từ vựng tiếng Việt (VLO) và phương pháp xây dựng VLO thủ công.
3. Đề xuất phương pháp phân tích ngữ nghĩa của câu tiếng Việt qua ba giai đoạn: (a) phân tích cú pháp phụ thuộc, (b) gán nhãn nghĩa từ vựng và (c) điều chỉnh các quan hệ phụ thuộc dựa trên kết quả phân tích cú pháp phụ thuộc và các ràng buộc nghĩa được lưu trữ trong VLO.
4. Đề xuất phương pháp tính toán độ tương đồng trên NN-BD-NN theo độ đo Jaccard-Tanimoto.

7. Cấu trúc của luận án

Ngoài phần Mở đầu và Kết luận – kiến nghị, luận án được trình bày qua năm chương như sau:

- **Chương 1** trình bày tổng quan về những nghiên cứu liên quan trong truy hồi văn bản theo ngữ nghĩa với hướng tiếp cận ngôn ngữ học tính toán và các cơ sở cho việc nghiên cứu ở các chương sau.
- **Chương 2** đề xuất NN-BD-NN, cấu trúc biểu diễn ngữ nghĩa của văn bản và trình bày mô hình truy hồi thông tin văn bản đã được nghiên cứu để áp dụng trên NN-BD-NN.

- **Chương 3** giới thiệu về Cơ sở tri thức ngữ nghĩa từ vựng tiếng Việt (VLO), phương pháp xây dựng và tác dụng của nó trong giải pháp truy hồi văn bản theo ngữ nghĩa.
- **Chương 4** trình bày phương pháp phân tích ngữ nghĩa cụm từ tiếng Việt dựa vào kết quả phân tích cú pháp phụ thuộc tiếng Việt kết hợp với các ràng buộc ngữ nghĩa trong VLO. Kết quả phân tích ngữ nghĩa được dùng xác định ngữ nghĩa của câu theo NN-BD-NN.
- **Chương 5** trình bày kết quả đánh giá mô hình truy hồi thông tin văn bản tiếng Việt dựa trên ngữ nghĩa.

CHƯƠNG 1. TỔNG QUAN

1.1 TRUY HỒI THÔNG TIN

1.1.1 Lịch sử nghiên cứu

Thuật ngữ "information retrieval" mới được C. N. Mooers đưa ra lần đầu tiên [44]. Theo Mark Sanderson, tác giả như H. F. Mitchell, B. Nanus, H. L. Brownson đã nghiên cứu truy hồi văn bản từ thập niên 1950 [61]. Hiện tại, thuật ngữ "information retrieval" có thể được diễn giải một cách chính xác theo quan điểm của C. D. Manning và các đồng tác giả (2008) [38] như sau:

> *"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."*

1.1.2 Một số mô hình truy hồi thông tin căn bản

Mô hình truy xuất thông tin căn bản có hai vấn đề cơ sở là cấu trúc biểu diễn văn bản và phương pháp phân tích văn bản.

1.1.2.1 Mô hình Vector (Salton, Wong and Yang 1975).

Mô hình vector có đặc điểm sau [60]:

- Dùng cấu trúc Bag of Word – BOW. Mỗi từ được gọi là term.
- Phương pháp phân tích văn bản hoặc truy vấn là tách các term và xác định trọng số của chúng dựa trên tần số và chỉ số IDF.
- So khớp văn bản và truy vấn theo độ tương đồng giữa hai vector của văn bản và truy vấn. Độ tương đồng có thể là một chuẩn (metric) hoặc một độ đo bất kỳ, chẳng hạn Euclide, Cosine, Jaccard, v.v.

Mô hình vector được cải tiến nhờ phương pháp phân tích ngữ nghĩa tiềm ẩn Latent Semantic Analysis – LSA[33], [16] hoặc sử dụng word embeddings[42]. Các word

embeddings có thể được ước lượng[41], [36], [55] từ một khối lượng văn bản rất lớn. Kết quả của sự cải tiến là độ phủ tăng.

1.1.2.2 Mô hình xác suất.

Mô hình xác suất có các đặc điểm [23]:

- Văn bản được biểu diễn bằng một phân phối đa thức của các term.
- Phương pháp phân tích tài liệu là xác định phân phối đa thức của các term có trong tài liệu đó.
- Phương pháp so khớp văn bản và tài liệu là tính xác suất tài liệu có liên quan đến truy vấn.

1.1.2.3 Mô hình chủ đề

Mô hình chủ đề có đặc điểm như sau:

- Văn bản được biểu diễn bằng một vector với số chiều có thể chọn. Mỗi chiều tương ứng với một chủ đề (topic).
- Phương pháp phân tích tài liệu được thực hiện qua hai bước là xác định tập chủ đề của tập tài liệu và tính xác suất của chúng.
- Phương pháp so khớp văn bản và truy vấn là tính toán xác suất mà văn bản tạo ra truy vấn.

Một số phương pháp phân tích tài liệu trong mô hình chủ đề gồm PLSA[29], LDA[9].

1.2 TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

Năm 1964, B. Raphael theo tác giả này, thuật ngữ "*semantic*" được quan niệm là "… *'meaning' of material*" [7] đã đưa ra khái niệm "Semantic Information Retrieval. Có hai nhóm phương pháp giải quyết vấn đề này là chú giải ngữ nghĩa và mở rộng truy vấn tự động.

1.2.1 Chú giải ngữ nghĩa

Gonzalo (1998) [27], Mihalcea (2000) [40], Ozcan (2004) [54], Giunchiglia (2008) [25], Wolf (2009) [68], Ngô Minh Vương (2018) [47] Chú giải ngữ nghĩa ở bằng các synset

trong WordNet [43]. Gonzalo (2000) [26] chú giải bằng các nhóm từ đồng nghĩa được rút trích tự động (sense cluster). Soner (2012) [31], Fernandez (2011) [22], Castells (2006) [11], Hliaoutakis (2006) [28] và Vallet (2005) [66] chú giải bằng các thực thể trong một ontology cho lĩnh vực riêng. Rodriguez (2016) [56], Egozi (2011) [21] và Gabrilovich (2007) [24] chú giải bằng các khái niệm tiềm ẩn. Trong các nghiên cứu này, vấn đề truy hồi được giải quyết trên kết quả chú giải thay vì trên từ ngữ trong văn bản và truy vấn theo mô hình vector, mô hình LSI và mô hình xác suất.

Rindflesch (1993) [58], Matsumura (2000) [39] sử dụng thêm quan hệ phụ thuộc được định nghĩa riêng, Moreda (2007) [45] dùng các quan hệ tham tố trong khung vị từ trong ngôn ngữ học [2] để loại bỏ bớt tài liệu không liên quan trong kết quả truy hồi.

Koopman (2016) [32] dùng một ontology trong lĩnh vực y khoa để xác định đồ thị các thực thể trong từ văn bản y khoa và tính toán trọng số các thực thể theo đồ thị này.

Amir (2017) [6] câu thành các bộ ba <S, V, O> (S – chủ từ, V – động từ, O – tân từ) để giải quyết bài toán so khớp câu.

Các nghiên cứu trong hướng tiếp cận này tập trung chủ yếu vào khía cạnh nghĩa của từ ngữ trong câu, khía cạnh quan hệ phụ thuộc trong câu chỉ được giải quyết theo một vài quan hệ phụ thuộc nên khái niệm ngữ nghĩa trong các nghiên cứu này chưa được đáp ứng đầy đủ.

1.2.2 Mở rộng truy vấn tự động

Fernandez (2011) [22], Tomassen (2010) [65] và Szymanski (2012) [64] xác định các từ khóa trong truy vấn và xác định các từ khóa liên quan của chúng nhờ từ điển đồng nghĩa hoặc ontology. Một truy vấn mở rộng là truy vấn ban đầu được nối thêm các từ khóa liên quan. Bài toán truy hồi được thực hiện trên văn bản và truy vấn được mở rộng thay vì truy vấn gốc để tăng độ phủ trong kết quả tìm kiếm.

Các nghiên cứu trong cách tiếp cận này cho thấy việc mở rộng truy vấn chỉ tập trung vào khía cạnh nghĩa của từ ngữ trong câu, khía cạnh quan hệ phụ thuộc được bỏ qua vì truy vấn mở rộng không thể có ngữ nghĩa của truy vấn gốc vì những từ liên quan được thêm vào truy vấn đã làm ngữ nghĩa của câu trở nên khác đi.

1.3 CÁC CƠ SỞ CHO TRUY HỒI VĂN BẢN TIẾNG VIỆT

1.3.1 Phân tích hình thái

Phân tích hình thái trong tiếng Việt đã được một số kết quả sau:

- Bài toán tách từ (Word Segmentation): Kết quả nghiên cứu của Cam-Tu Nguyen (2006) có F_1 đạt được là 94.23% [48]. Kết quả nghiên cứu của Phuong Hong Le có độ chính xác là 95.6%
- Bài toán gán nhãn từ loại (POS Tagging): Kết quả nghiên cứu của Phuong Hong Le (2010) có độ chính xác là 93.4% [34].

1.3.2 Phân tích cú pháp phụ thuộc

Có hai hướng tiếp cận:

- Hướng tiếp cận học máy cần treebank đủ lớn. Kết quả của *Dat Quoc Nguyen* (2016) độ chính xác đạt 0.739 [49] với dependency treebank được chuyển đổi từ constituent treebank của D. Q. Nguyen (2014) [50]. Ngữ liệu huấn luyện chứa các đặc trưng từ và từ loại, chưa có đặc trưng nghĩa của từ.
- Hướng tiếp cận theo hệ luật văn phạm sử dụng các luật văn phạm theo ngữ pháp cấu trúc ngữ đoạn hướng tâm – Head-driven Phrase Structure Grammar (HPSG) [57],[35]. Hiện tại chưa có kết quả công bố cho tiếng Việt.

1.3.3 Phân tích ngữ nghĩa của câu

Có hai phương tiện để biểu diễn ngữ nghĩa của câu:

1. Logic hình thức: Blackburn (2003) [8], Delmonte (2009) [17] và Kamp (2011) [30]. Phương pháp tính toán hiện tại chỉ áp dụng cho những câu đơn giản trong đó mỗi mệnh đề chỉ có một động từ.
2. Kết quả phân tích phụ thuộc của câu: Oepen và các đồng tác giả (2014) [52, 53], Schuster và Manning (2016) [62] phân tích câu thành các quan hệ phụ thuộc theo Stanford Dependencies [15].

1.4 VẤN ĐỀ NGỮ NGHĨA TRONG TRUY XUẤT VĂN BẢN TIẾNG VIỆT

1.4.1 Ngữ nghĩa của từ

Khái niệm 1.1 Nghĩa của từ vựng

Nghĩa của một từ vựng là một ký hiệu không trùng lắp được gắn cho mỗi sự vật, mỗi tính chất hoặc mỗi hành vi được diễn tả bởi từ vựng đó trong một văn cảnh xác định. Nếu dùng các ký hiệu khác nhau để gắn cho những nội dung giải nghĩa khác nhau của từ vựng trong một từ điển thì những ký hiệu này chính là những nghĩa của từ vựng đó.

1.4.2 Ngữ nghĩa của ngữ đoạn

Khái niệm 1.2. Cụm từ

Cụm từ là một dãy các từ liên tiếp nhau và có các mối quan hệ ngữ pháp và ngữ nghĩa với nhau để tạo nên cấu trúc của một ngữ đoạn hoặc một câu. Trường hợp chỉ có một từ thì cũng được xem là một cụm từ.

Khái niệm 1.3 Ngữ nghĩa của cụm từ

Ngữ nghĩa của một cụm từ là nghĩa của từng từ vựng trong các mối quan hệ phụ với những từ khác trong chính cụm từ đó. Các mối quan hệ phụ thuộc này bao gồm các quan hệ tham tố của khung vị từ và các quan hệ bổ nghĩa. Ngữ nghĩa của một cụm từ có thể được đại diện bởi một từ có vai trò trung tâm.

1.4.3 Ngữ nghĩa của văn bản

Khái niệm 1.4 - Văn bản

Văn bản là một tập có phân biệt thứ tự các cụm từ. Văn bản có nghĩa xác định dựa trên ngữ nghĩa của các cụm từ theo trình tự xuất hiện của chúng và các liên từ tạo nên cấu trúc diễn ngôn của văn bản.

Khái niệm 1.5 Ngữ nghĩa của văn bản

Ngữ nghĩa của văn bản không đơn giản là kết quả cộng gộp ngữ nghĩa của các cụm từ. Ngữ nghĩa của văn bản là ngữ nghĩa của từng câu trong cấu trúc diễn ngôn của văn bản.

1.4.4 Truy hồi thông tin văn bản theo ngữ nghĩa

Khái niệm 1.6 Truy hồi thông tin văn bản theo ngữ nghĩa

Truy hồi thông tin văn bản theo ngữ nghĩa theo cách tiếp cận ngôn ngữ học tính toán là truy hồi thông tin văn bản trong đó quá trình so khớp được thực hiện trên ngữ nghĩa của văn bản và ngữ nghĩa của cụm từ truy vấn.

1.5 CÁC VẤN ĐỀ CẦN NGHIÊN CỨU

Để giải quyết bài toán đặt ra, luận án xác định cần phải có các điều kiện sau:

1) Một từ điển các nhãn nghĩa từ vựng để khi thực hiện so khớp sẽ so khớp trên các nhãn nghĩa. Kết quả so khớp trên các nhãn thể hiện kết quả so khớp về nghĩa từ vựng.
2) Một tập hợp các ràng buộc giữa các nghĩa từ vựng, có vai trò như những ràng buộc ngữ nghĩa trong các văn phạm có ràng buộc ngữ nghĩa HPSG [57] và văn phạm gia tố do [5] để có được kết quả phân tích cú pháp đảm bảo đúng ngữ nghĩa.
3) Phương pháp phân ngữ nghĩa dựa trên kết quả phân tích cú pháp theo ngữ pháp phụ thuộc để xác định tất cả quan hệ phụ thuộc đúng ngữ nghĩa trong câu.
4) Ngôn ngữ để biểu diễn ngữ nghĩa, làm nền tảng cho việc so khớp ở mức ngữ nghĩa.
5) Mô hình truy hồi phù hợp để xử lý việc so khớp ở mức ngữ nghĩa nhằm điều chỉnh độ chính xác và độ phủ trong kết quả truy hồi.

Vì thế, luận án đã tiến hành:

a) Đề xuất NN-BD-NN, biểu diễn ngữ nghĩa của văn bản và nghiên cứu mô hình truy hồi văn bản trên biểu diễn ngữ nghĩa của văn bản và truy vấn. Các nội dung này được trình bày trong **Chương 2**.

b) Nghiên cứu xây dựng một cơ sở tri thức ngữ nghĩa từ vựng tiếng Việt (Vietnamese Lexicon Ontology - VLO) cho điều kiện 1) và 2). Nội dung này được trình bày trong **Chương 3** của luận án.

c) Nghiên cứu phương pháp phân tích ngữ nghĩa của cụm từ, được trình bày trong **Chương 4** của luận án, để đáp ứng điều kiện 3).

CHƯƠNG 2. MÔ HÌNH TRUY HỒI VĂN BẢN

2.1 BIỂU DIỄN NGỮ NGHĨA

2.1.1 Ngôn ngữ biểu diễn ngữ nghĩa

Ngôn ngữ biểu diễn ngữ nghĩa (NN-BD-NN) là một ngôn ngữ hình thức dùng để biểu diễn ngữ nghĩa của cụm từ, được định nghĩa dựa hướng tiếp cận True-Conditional Semantics [10, 46].

Định nghĩa 2.1 Ngôn ngữ biểu diễn ngữ nghĩa

Ngôn ngữ biểu diễn ngữ nghĩa, ký hiệu là L_S, gồm các thành phần:

1. *Tập từ vựng V gồm các ký hiệu có dạng S·I và một ký hiệu đặc biệt ROOT. Trong đó:*
   - *S là nghĩa từ vựng theo **Định nghĩa 2.1***
   - *I là một số nguyên chỉ lần thứ I từ vựng tương ứng xuất hiện trong phát biểu.*
2. *Tập các quan hệ R là các quan hệ hai ngôi, xác định trên tập V. Giả sử $x, y \in V$, có 7 quan hệ trong R có ý nghĩa như sau:*
   a. *hasMod<x,y> là quan hệ bất đối xứng, cho biết y là modifier của x, nghĩa là y bổ sung thêm thuộc tính hoặc tính chất cho x trong ngữ đoạn hoặc câu.*
   b. *hasPComp<x, y> là quan hệ bất đối xứng, cho biết y là complement của x, nghĩa là y bổ sung thông tin về khung cảnh hay các mối liên hệ của x trong ngữ đoạn hoặc câu.*
   c. *hasActor<x, y> là quan hệ bất đối xứng, cho biết y là nhân tố của hành động x.*
   d. *hasDObj<x, y> là quan hệ bất đối xứng, cho biết y là tham tố trực tiếp của hành động x.*
   e. *hasIDObj<x, y> là quan hệ bất đối xứng, cho biết y là tham tố gián tiếp của hành động x.*

f. root<ROOT, x> là quan hệ bất đối xứng, cho biết x là thành tố trung tâm của một mệnh đề.

3. *Cho* $x \cdot xi, y \cdot yi, z \cdot zi, t \cdot ti \in V$ *và* $r_1, r_2 \in R$, *thì:*

$$r_1 < x \cdot xi, y \cdot yi >= r_2 < z \cdot zi, t \cdot ti >$$

nếu và chỉ nếu:

$$r_1 = r_2, x = z, y = t, xi = zi, yi = it$$

4. *Phép toán: chỉ có một phép toán nối, ký hiệu bằng khoảng trắng " ", là một phép toán hai ngôi xác định trên hai quan hệ* $r_1 < x_1, y_1 >$ *và* $r_2 < x_2, y_2 >$ *với* $r_2, r_2 \in R$ *và* $x_1, x_2, y_1, y_2 \in V$ *cho biết hai quan hệ này cùng được nhắc đến trong một phát biểu. Phép toán nối có tính chất giao hoán.*

5. *Mệnh đề trong* L_S

   a. $r < x, y >$ *là một mệnh đề với* $r \in R$ *và* $x, y \in V$

   b. *Nếu p và q là hai mệnh đề thì t = p q = q p là một mệnh đề.*

   c. *Cho hai mệnh đề p và q, p=q nếu và chỉ nếu tất cả quan hệ có trong mệnh đề p đều có trong mệnh đề q và tất cả quan hệ có trong mệnh đề q đều có trong mệnh đề p.*

6. *Giá trị chân lý của mệnh đề trong* L_S

   a. *Một mệnh đề* $r < x, y >$ *có giá trị chân lý đúng (true) trong ngữ cảnh* t *nếu và chỉ nếu thực sự có quan hệ phụ thuộc* r *giữa nghĩa* x *và nghĩa* y *trong ngữ cảnh* t. *Ngược lại thì* $r < x, y >$ *có giá trị chân lý sai (false).*

   b. *Cho* p *và* q *là hai mệnh đề, mệnh đề* $u = p\, q = q\, p$ *có giá trị chân lý đúng (true) trong ngữ cảnh* t *nếu và chỉ nếu* p *và* q *cùng đúng trong ngữ cảnh* t. *Ngược lại thì* u *có giá trị chân lý sai (false).*

Theo **Định nghĩa 2.1**, có ba tính chất quan trọng của ngôn ngữ hình thức L_S là:

Tính chất 1 - Tính không nhập nhằng về từ vựng

Với mọi $x \cdot xi, y \cdot yi, z \cdot zi \in V$ *với x,y,z là các nghĩa từ vựng tùy ý, xi,yi,zi là các số nguyên tùy ý,* $x \neq y$, *thì các điều sau đúng với mọi* $r \in R$:

- $r < x \cdot xi, z \cdot zi >\neq r < y \cdot yi, z \cdot zi >$
- $r < z \cdot zi, x \cdot xi >\neq r < z \cdot zi, y \cdot yi >$

- $r < x \cdot xi, z \cdot zi >= r < x \cdot xi, z \cdot zi >$
- $r < z \cdot zi, x \cdot xi >= r < z \cdot zi, x \cdot xi >$

Tính chất này có thể phát biểu rằng từ vựng trong L_S không có hiện tượng đồng nghĩa và đồng âm.

Tính chất 2 - Tính không nhập nhằng về cấu trúc

Cho $a_1, a_1, .. a_n$ là các mệnh đề trong L_S, nếu hai mệnh đề $p = a_1 a_1 .. a_n$ và $q = a_1 a_1 .. a_n$ thì p = q

Tính chất này có thể phát biểu rằng một mệnh đề trong L_S không có hai nghĩa khác nhau.

Tính chất 3 – Tính không phụ thuộc vào thứ tự các thành tố trong mệnh đề

Cho mệnh đề $p = a_1\, a_2 .. a_n$ với a_i là các mệnh đề trong L_S. Gọi $q = a_{q1}\, a_{q2} .. a_{qn}$ là một hoán vị của các mệnh đề a_i trong p thì $p = q$.

Các **Tính chất 1** và **Tính chất 2** cho phép thực hiện việc so khớp hai mệnh đề trong một điều kiện lý tưởng là không có nhập nhằng về từ vựng và cấu trúc. **Tính chất 3** đảm bảo cho việc vector hóa trong mô hình vector hay giả thiết độc lập về thứ tự trong mô hình xác suất không ảnh hưởng đến ngữ nghĩa của mệnh đề.

Định lý 2.1 Biến đổi một phát biểu từ nhiên thành một phát biểu trong $\boldsymbol{L_S}$

Trong một ngữ cảnh xác định, cho:

*$s = (w_1, w_2, .., w_n)$ là một phát biểu có **n** từ vựng trong ngôn ngữ tự nhiên.*

$Dep_S = \{r_i < w_{ai} \cdot ai, w_{bi} \cdot bi > | i = 1..m; ai, bi \in N\}$ là kết quả phân tích quan hệ phụ thuộc của s

$c_{ai} \cdot ki, c_{bi} \cdot li \in V$ với c_{ai} và c_{bi} là nghĩa của từ w_{ai} và w_{bi} trong ngữ cảnh đang xét và $ki, li \in N$ là lần thứ ki và lần thứ li các nghĩa tương ứng c_{ai} và c_{bi} xuất hiện tính từ w_1 đến w_{ai} và w_1 đến w_{bi} trong phát biểu s.

Khi đó, phép biến đổi F sau sẽ biến đổi phát biểu s thành phát biểu t trong L_S sao cho s và t cùng đúng hoặc cùng sai trong mọi ngữ cảnh.

- $F(root < ROOT, w_{bi} \cdot bi >) = root < ROOT, c_{bi} \cdot li >$
- $F(r_i < w_{ai} \cdot ai, w_{bi} \cdot bi >) = hasMod < c_{ai} \cdot ki, c_{bi} \cdot li >$ *nếu* r_i *là các loại*

quan hệ phụ thuộc con của quan hệ modifier.

- $F(r_i(w_{ai} \cdot ai, w_{bi} \cdot bi)) = hasPComp(c_{ai} \cdot ki, c_{bi} \cdot li)$ *nếu* r_i *là các loại quan hệ phụ thuộc con của quan hệ complement.*
- $F(r_i(w_{ai} \cdot ai, w_{bi} \cdot bi)) = hasActor(c_{ai} \cdot ki, c_{bi} \cdot li)$ *nếu* r_i *là quan hệ phụ thuộc chủ từ logic, nghĩa là đã xử lý trường hợp câu chủ động và bị động.*
- $F(r_i(w_{ai} \cdot ai, w_{bi} \cdot bi)) = hasDObj(c_{ai} \cdot ai, c_{bi} \cdot bi)$ *nếu* r_i *là quan hệ phụ thuộc tân từ trực tiếp sau khi đã xử lý trường hợp chủ động và bị động.*
- $F(r_i(w_{ai} \cdot ai, w_{bi} \cdot bi)) = hasIDObj(c_{ai} \cdot ki, c_{bi} \cdot li)$ *nếu* r_i *là quan hệ phụ thuộc tân từ gián tiếp sau khi đã xử lý trường hợp chủ động và bị động.*

$$t = F(Dep_s)$$
$$= F\left(r_1(w_{a1} \cdot k1, w_{b1} \cdot l1)\right)$$
$$F\left(r_2(w_{a2} \cdot k2, w_{b2} \cdot l2)\right) \ldots$$
$$F(r_n(w_{an} \cdot kn, w_{bn} \cdot ln))$$

Định lý 2.2 So sánh hai phát biểu trong ngôn ngữ nhiên qua $\boldsymbol{L_S}$

Trong một ngữ cảnh xác định, cho s_1 *và* s_2 *là hai phát biểu trong ngôn ngữ tự nhiên có kết quả phân tích phụ thuộc lần lượt là* Dep_1 *và* Dep_2. *Giả sử* $t_1 = F(Dep_1)$ *và* $t_2 = F(Dep_2)$. *Khi đó, nếu* $t_1 = t_2$ *thì* s_1 *và* s_2 *có cùng nghĩa trong ngữ cảnh đó.*

2.1.2 Biểu diễn ngữ nghĩa của cụm từ

Khái niệm 2.1 Cấu trúc biểu diễn ngữ nghĩa của cụm từ

Cho một từ điển có từ vựng và ký hiệu nghĩa tương ứng trong ngôn ngữ tự nhiên $\boldsymbol{L_N}$. *Ngữ nghĩa của cụm từ s trong* $\boldsymbol{L_N}$ *là một phát biểu* **t** *trong ngôn ngữ biểu diễn ngữ nghĩa* $\boldsymbol{L_S}$ *theo Error! Reference source not found. sao cho* $\boldsymbol{t=F(Dep_s)}$ *với F là phép biến đổi tập quan hệ phụ thuộc thành phát biểu trong* $\boldsymbol{L_S}$ *theo Error! Reference source not found. và* $\boldsymbol{Dep_s}$ *là tập các quan hệ phụ thuộc của* **s**.

*Để thuận tiện trong quá trình tính toán, ngữ nghĩa của cụm từ được biểu diễn bằng một bộ <**c**,**C**,**R**> trong đó:*

- ***c** là nghĩa từ vựng có vai trò trung tâm của cụm từ.*
- ***C** là một dãy các nghĩa từ vựng có được bằng cách chọn không lặp các từ*

*vựng c_{ai} · ki trong các quan hệ của **t** và lấy chỉ c_{ai}. **C** là một dãy không phân biệt thứ tự.*

- ***R** là một dãy các quan hệ phụ thuộc có được bằng cách lấy toàn bộ các quan hệ trong **t**. **R** là một dãy không có thứ tự.*

2.1.3 Biểu diễn ngữ nghĩa của văn bản

Khái niệm 2.2 Cấu trúc biểu diễn ngữ nghĩa của văn bản

*Cho một từ điển có các từ vựng và ký hiệu nghĩa tương ứng trong ngôn ngữ tự nhiên L_N, một văn bản **Doc**=$\{s_1,s_2,..,s_n\}$ trong L_N với s_i là cụm từ thứ i trong văn bản Doc có ngữ nghĩa tương ứng là bộ $<c_i,C_i,R_i>$. Khi đó, ngữ nghĩa của **Doc** là một bộ <C,R>, trong đó:*

- *C là kết quả nối các dãy C_i*
- *$R = \{R_i|i=1..n\}$*

2.2 ĐỀ XUẤT MÔ HÌNH CHUNG

Mô hình chung được đề xuất để áp dụng cho cách tiếp phân tích tài liệu và truy vấn theo hai mặt thành phần và cấu trúc tương ứng với dãy ***C*** và tập ***R*** theo **Khái niệm 2.2**

2.2.1 Yếu tố ảnh hưởng đến độ chính xác và độ phủ

2.2.1.1 Sự trùng khớp term

Có hai trường hợp:

- Trường hợp thứ nhất, câu truy vấn sử dụng từ ngữ khác với từ ngữ được sử dụng trong văn bản làm giảm độ phủ của kết quả truy hồi.
- Trường hợp thứ hai, câu truy vấn và văn bản chứa những term thường xuất hiện trong đa số các văn bản nhưng những term này không giúp ích trong việc phân biệt nội dung của văn bản làm giảm độ chính xác của kết quả truy hồi.

2.2.1.2 Công thức xếp hạng tài liệu

Theo kết quả nghiên cứu [14] cho thấy có ba yếu tố quan trọng ảnh hưởng đến kết quả xếp hạng.

1. Khả năng phân biệt nội dung của term. Term chỉ xuất hiện trong một số tài liệu có liên quan đến nhau. Yếu tố này được thể hiện qua chỉ số IDF.
2. Tần số xuất hiện TF của term trong tài liệu.
3. Độ dài của tài liệu. Tài liệu càng dài thì nội dung càng có nhiều chủ đề.

2.2.2 Biểu diễn văn bản và truy vấn

Văn bản và truy vấn đối với mô hình đề xuất có dạng là một bộ $< C, R >$ và truy vấn được xử lý như một văn bản chỉ có một cụm từ. Trong bộ $< C, R >$, C là một chuỗi các đặc trưng thành phần (từ, nghĩa từ vựng, khái niệm, …) của văn bản và R là tập hợp các chuỗi chứa các đặc trưng cấu trúc (bi-gram, quan hệ phụ thuộc, ..) của văn bản. Khi đó, theo mô hình vector, văn bản và truy vấn sẽ được biểu diễn bằng hai ma trận tương ứng với thành phần C và thành phần R. Trong đó:

Thành phần C được biểu diễn như một văn bản thông thường bằng một ma trận Term-Document như **Hình 2.1** và thành phần R được biểu diễn theo mức câu bằng một ma trận Term-Sentence như **Hình 2.2**

		Document			
		d_1	d_2	..	d_n
Term	c_1	TF_{11}	TF_{12}		TF_{1n}
	c_2	TF_{21}	TF_{22}		TF_{2n}
	c_m	TF_{m1}	TF_{m2}		TF_{mn}

Hình 2.1 Ma trận Term-Document được lập cho thành phần C trong văn bản trong đó c_i là các nghĩa từ vựng có trong chuỗi C của tất cả văn bản, d_j là văn bản thứ j trong tập tài liệu, TF_{ij} là giá trị tần số của nghĩa c_i có trong văn bản d_j

		Sentence							
		d_1			..	d_n			
		s_1	..	s_{k1}	..	s_1	..	s_{kn}	
Term	$r_1<x_{u1},y_{v1}>$	$TF_{1,1,1}$		$TF_{1,1,k1}$	..	$TF_{1,n,1}$		$TF_{1,n,kn}$	
	$r_2<x_{u2},y_{v2}>$	$TF_{2,1,1}$		$TF_{2,1,k1}$	..	$TF_{2,n,1}$		$TF_{2,n,kn}$	
	$r_m<x_{um},y_{vm}>$	$TF_{m,1,1}$		$TF_{m,1,k1}$	..	$TF_{m,n,1}$		$TF_{m,n,kn}$	

Hình 2.2 Ma trận Term-Sentence được lập cho thành phần R trong văn bản trong đó r_i $<x_{ui},y_{vi}>$là các quan hệ trên các nghĩa từ vựng có trong chuỗi R trong tất cả văn bản, d_j là văn bản thứ j trong tập tài liệu, s_{jk} là chuỗi quan hệ nghĩa thứ k trong văn bản d_j, $TF_{j,i,k}$ là giá trị tần số của quan hệ phụ thuộc nghĩa $r_i<x_{ui},y_{vi}>$ có trong chuỗi quan hệ phụ thuộc thứ k tương ứng với cụm từ s_k trong văn bản d_j

2.2.3 Tính toán độ liên quan giữa văn bản và truy vấn

Được tính theo **Khái niệm 2.11** là:

$$d(T,q) = \alpha \times d_c(C_T, C_q) + (1-\alpha) \times d_r(R_T, R_q)$$

- Việc tính toán $d_c(C_T, C_q)$ và $d_r(R_T, R_q)$ sẽ được tính toán dựa trên khoảng cách ngữ nghĩa được nêu trong **Khái niệm 2.6** và **Khái niệm 2.10**.

2.3 ĐỘ ĐO KHOẢNG CÁCH NGỮ NGHĨA

Khoảng cách ngữ nghĩa trong luận án được phát triển từ khoảng cách ngữ nghĩa của cụm từ [18] với cơ sở là khoảng cách Jaccard-Tanimoto[37].

2.3.1 Khoảng cách Jaccard-Tanimoto

Khoảng cách Jaccard-Tanimoto [37] giữa hai tập hợp ***A*** và ***B***, ký hiệu ***J_d(A,B)*** được tính dựa trên chỉ số Jaccard ***J(A,B)*** như sau:

$$J_d(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Trong trường hợp $X=\{x_i\}$ và $Y = \{y_i\}$ là hai vector ***n*** chiều trong đó giá trị mỗi chiều là một số thực không âm, khoảng cách Jaccard giữa X và Y được tính theo công thức:

$$J_d(X,Y) = 1 - \frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)}$$

2.3.2 Độ đo khoảng cách

Độ đo khoảng cách ngữ nghĩa được xây dựng trên cơ sở độ đo khoảng cách Jaccard-Tanimoto [37] qua các khái niệm được trình bày chi tiết trong luận án. Các khái niệm này gồm

- **Khái niệm 2.3** – Độ đo khoảng cách giữa hai nghĩa từ vựng
- **Khái niệm 2.4** – Độ đo khoảng cách ngữ nghĩa giữa một nghĩa từ vựng đến một tập các nghĩa từ vựng
- **Khái niệm 2.5** – Độ đo khoảng cách ngữ nghĩa giữa hai tập nghĩa từ vựng
- **Khái niệm 2.6** – Độ đo khoảng cách ngữ nghĩa có trọng số giữa hai tập nghĩa từ vựng
- **Khái niệm 2.7** – Độ đo khoảng cách ngữ nghĩa giữa hai quan hệ phụ thuộc trong L_S
- **Khái niệm 2.8** – Độ đo khoảng cách ngữ nghĩa giữa một quan hệ phụ thuộc và một tập các quan hệ phụ thuộc trong L_S
- **Khái niệm 2.9** – Độ đo khoảng cách ngữ nghĩa giữa hai tập quan hệ phụ thuộc
- **Khái niệm 2.10** – Độ đo khoảng cách ngữ nghĩa có trọng số của hai tập quan hệ phụ thuộc
- **Khái niệm 2.11** – Độ đo khoảng cách ngữ nghĩa giữa một văn bản và một truy vấn

2.3.3 Các trọng số

2.3.3.1 Tần số của term

Trong mô hình đề xuất của luận án, tần số của một term được tính bằng đúng số lần xuất hiện của nó.

2.3.3.2 Độ quan trọng của term

Độ quan trọng của term, thể hiện bằng chỉ số IDF, tính theo công thức đã được công bố [38] như sau:

$$IDF(t) = log\left(\frac{ndocs}{ndoc_t + 1} + 1\right) \quad (2.1)$$

Luận án đề xuất sử dụng thêm trọng số phản ánh vị trí của term trong đồ thị ngữ nghĩa theo **Khái niệm 2.10**

2.3.3.3 Độ dài của văn bản

Đối với thành phần C, độ dài văn bản theo C được tính theo công thức:

$$sum_C = \frac{nterms}{\sqrt{\sum_i \sum_j \theta_{ij}}} \quad (2.2)$$

Trong đó:

- $nterms$ là số lượng nghĩa từ vựng khác nhau có trong văn bản hoặc của truy vấn.
- θ_j là trọng số vị trí của nghĩa từ vựng thứ j trong cụm từ thứ i của văn bản hoặc của truy vấn.

Đối với thành phần R, độ dài được tính theo mức câu vì việc so khớp diễn ra trên mức câu theo

Khái niệm 2.11 và được tính theo công thức:

$$sum_{\theta_i} = \frac{nterms_i}{\sqrt{\sum_j \rho_j}} \quad (2.3)$$

Trong đó:

- $nterms_i$ là số lượng quan hệ phụ thuộc khác nhau có trong ngữ nghĩa của cụm từ thứ i của văn bản hoặc của truy vấn.
- ρ_j là trọng số vị trí của quan hệ phụ thuộc thứ j có trong ngữ nghĩa của cụm từ thứ i của văn bản hoặc của truy vấn.

Công thức này được đề xuất vì:

- Chiều dài của văn bản không tăng tuyến tính theo tỉ lệ tăng của số lượng term.
- Trong trường hợp mỗi term xuất hiện một lần thì chiều dài văn bản bằng căn bậc hai của số lượng term. Trong trường hợp tần số của mỗi term cao thì, về ý nghĩa, độ dài văn bản không lớn.

2.3.3.4 Tỉ lệ giữa khoảng cách và số term trùng khớp

Tỉ lệ này không sử dụng trong công thức tính khoảng cách mà được sử dụng như một độ ưu tiên khi xếp hạng. Tỉ lệ này được sử dụng như một heuristic nhằm tăng cường kết quả

xếp hạng danh sách tài liệu liên quan. Tỉ lệ giữa khoảng cách ngữ nghĩa của văn bản T và truy vấn q và số term trùng khớp, ký hiệu là $\delta_{T,q}$, được đề xuất trong luận án như sau:

$$\delta_{T,q} = \frac{d_c(C, C_q)}{\sum_{C \cap C_q} \theta_i} \quad (2.4)$$

Tỉ lệ này cho thấy khoảng cách trong trường hợp số lượng term trùng khớp ít sẽ cao hơn khoảng cách trong trường hợp số lượng term trùng khớp nhiều.

2.4 CHỈ MỤC NGỮ NGHĨA

Chỉ mục ngữ nghĩa được xây dựng theo cấu trúc chỉ mục nghịch đảo [38]

2.4.1 Chỉ mục lớp nghĩa

Chỉ mục lớp nghĩa SCI (Semantic Class Index), dùng để tính toán khoảng cách ngữ nghĩa theo đặc trưng nghĩa từ vựng, theo **Khái niệm 2.11**

Lớp nghĩa	**Số Tài liệu**	**w**	**Postings list**				
c_1	N_1	idf_{c_1}	L_1	→	d_{11} $w_{c_1}^{d_{11}}$	…	d_{1k} $w_{c_1}^{d_{1k}}$
c_2	N_2	idf_{c_2}	L_2	→	d_{21} $w_{c_2}^{d_{21}}$	…	d_{2k} $w_{c_2}^{d_{2k}}$
…	…	…	…	→	…	…	…
c_m	N_m	idf_{c_m}	L_m	→	d_{m1} $w_{c_m}^{d_{m1}}$	…	d_{mk} $w_{c_m}^{d_{mk}}$

a)

d_1	d_2	…	d_{n_D}
sum_1	sum_2	…	sum_{n_D}

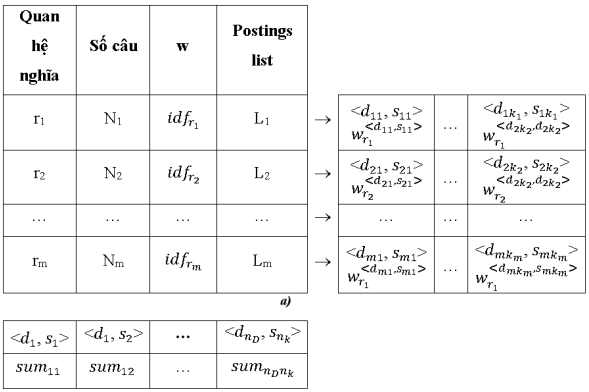
b)

Hình 2.3 Tổ chức chỉ mục lớp nghĩa gồm:

a) Từ điển và postings list b) Tổng trọng số của các lớp nghĩa trong từng tài liệu

2.4.2 Chỉ mục quan hệ nghĩa

Chỉ mục quan hệ nghĩa SRI (Semantic Relation Index), dùng để tính toán khoảng cách ngữ nghĩa theo quan hệ phụ thuộc trong ngữ nghĩa theo **Khái niệm 2.11**



Hình 2.4 Tổ chức chỉ mục quan hệ nghĩa gồm a) Từ điển và postings list
b) Tổng trọng số các quan hệ nghĩa trong một câu

2.5 TRUY HỒI CHỈ MỤC NGỮ NGHĨA

2.5.1 Truy hồi chỉ mục lớp nghĩa

Được thực hiện theo phương pháp truy hồi chỉ mục nghịch đảo trên chỉ mục lớp nghĩa SCI.

2.5.2 Truy hồi chỉ mục quan hệ nghĩa

Được thực hiện theo phương pháp truy hồi chỉ mục nghịch đảo trên chỉ mục quan hệ nghĩa **SRI**

2.5.3 Tính toán khoảng cách ngữ nghĩa

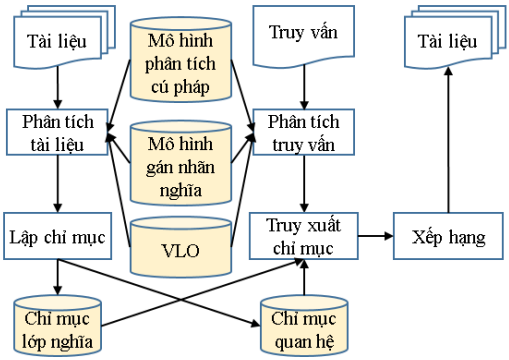
Khoảng cách ngữ nghĩa giữa biểu diễn ngữ nghĩa của văn bản và truy vấn được tính toán theo **Khái niệm 2.11**.

2.5.4 Tính độ liên quan để xếp hạng

Khoảng cách ngữ nghĩa giữa truy vấn và các tài liệu chỉ là một trong những chỉ tiêu ảnh hưởng đến kết quả xếp hạng các tài liệu được truy hồi. Để nâng cao kết quả xếp hạng cần phải sử dụng thêm một đại lượng là tỉ lệ khoảng cách ngữ nghĩa theo số term trùng khớp theo **Công thức (2.4)**.

2.6 MÔ HÌNH HỆ THỐNG

Mô hình hệ thống truy hồi văn bản tiếng Việt dựa trên ngữ nghĩa phát triển từ mô hình hệ thống truy hồi văn bản [19] trong đó áp dụng các phương pháp biểu diễn và phân tích tài liệu, lập chỉ mục, truy hồi chỉ mục và tính toán độ liên quan để xếp hạng đã trình bày trong các **Mục 2.1**, **Mục 2.3**, **Mục 2.4** và **Mục 2.5**. Mô hình hệ thống này gồm các thành phần được biểu diễn trong **Hình 2.5**

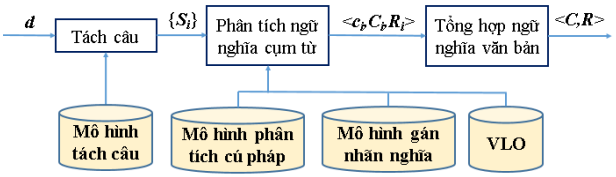


Hình 2.5 Mô hình hệ thống tìm kiếm văn bản tiếng Việt dựa trên ngữ nghĩa

2.6.1 Thành phần phân tích tài liệu

Đầu vào: Từng tài liệu văn bản d có trong tập văn bản D

Đầu ra: Ngữ nghĩa $< C, R >$ tương ứng với mỗi tài liệu văn bản d



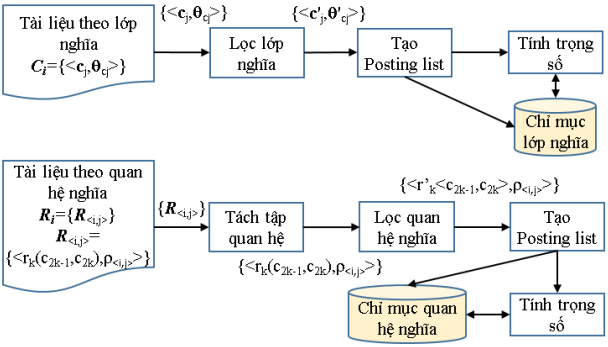
Hình 2.6 Sơ đồ thành phần phân tích tài liệu

2.6.2 Thành phần lập chỉ mục

Đầu vào: Từng cấu trúc ngữ nghĩa của tài liệu văn bản $<C_i,R_i>$ tương ứng với tài liệu d_i

Đầu ra:

- Chỉ mục lớp nghĩa (SCI) được lập cho từng tập các lớp nghĩa C_i tương ứng với tài liệu d_i
- Chỉ mục quan hệ nghĩa (SRI) được lập cho từng tập quan hệ phụ thuộc $R_{<i,j>}$ tương ứng với cụm từ thứ j trong tài liệu d_i



Hình 2.7 Sơ đồ thành phần lập chỉ mục

2.6.2.1 Lập chỉ mục lớp nghĩa

Quá trình lập chỉ mục lớp nghĩa được thực hiện qua ba bước xử lý chính:

- **Lọc lớp nghĩa**. Các lớp nghĩa không thuộc nhóm danh từ, danh từ riêng, động từ, tính từ và nhóm không xác định sẽ được loại bỏ.

- **Tạo Posting List**.

2.6.2.2 Lập chỉ mục quan hệ nghĩa

Quá trình lập chỉ mục quan hệ nghĩa được thực hiện qua bốn bước xử lý chính:

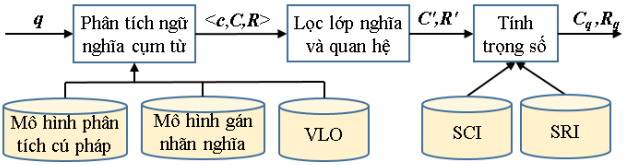
- **Tách tập quan hệ**. Tách tập $\boldsymbol{R_i}$ của mỗi tài liệu $\boldsymbol{d_i}$ thành các tập quan hệ phụ thuộc $\boldsymbol{R}_{<i,j>}$ tương ứng với câu thứ $\boldsymbol{j}$ của tài liệu $\boldsymbol{d_i}$
- **Lọc quan hệ nghĩa**. Loại bỏ các quan hệ phụ thuộc trong $\boldsymbol{R}_{<i,j>}$ mà một trong hai thành phần phụ thuộc là lớp nghĩa không thuộc nhóm danh từ, danh từ riêng, động từ, tính từ hoặc nhóm không xác định.
- **Tạo Posting List.**

2.6.3 Thành phần phân tích truy vấn

Đầu vào: Một câu truy vấn $\boldsymbol{q}$

Đầu ra: Tập $\boldsymbol{C}=\{<c_j,\theta_{cj}>\}$ và $\boldsymbol{R}=\{<r_j<c_x,c_y>,\rho_{rj}>\}$ là tập các lớp nghĩa và tập các quan hệ phụ thuộc nghĩa đã được tính trọng số của truy vấn $\boldsymbol{q}$

Thành phần phân tích truy vấn có sơ đồ được trình bày như **Hình 2.8**



Hình 2.8 Sơ đồ thành phần phân tích truy vấn

2.6.4 Thành phần Truy hồi chỉ mục

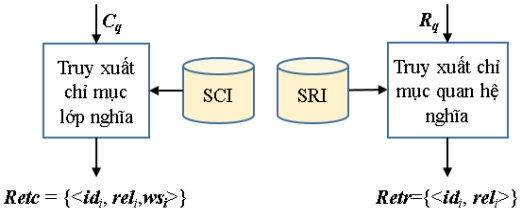
Đầu vào:

- Tập các lớp nghĩa $\boldsymbol{C_q}$
- Tập các quan hệ nghĩa $\boldsymbol{R_q}$
- Chỉ mục ***SCI***
- Chỉ mục ***SRI***

Đầu ra:

- Tập ***Retc*** = {<id_i,rel_i,ws_i>} là danh sách mà mỗi phần tử của nó là chỉ số id_i, độ liên quan rel_i của tài liệu id_i và C_q và tổng trọng số của các lớp nghĩa trùng nhau giữa tài liệu id_i và C_q
- Tập ***Retr*** = {<id_i,rel_i >} là danh sách mà mỗi phần tử của nó là chỉ số id_i và độ liên quan rel_i của tài liệu id_i và R_q

Sơ đồ của thành phần Truy hồi chỉ mục được trình bày trong **Hình 2.9**. Thành phần này có hai luồng xử lý song song. Luồng thứ nhất truy hồi chỉ mục lớp nghĩa ***SCI*** được xử lý như **Mục 2.5.1.** Luồng thứ hai truy hồi chỉ mục quan hệ nghĩa ***SRI*** được xử lý như **Mục 2.5.2**



Hình 2.9 Sơ đồ thành phần truy hồi chỉ mục

2.6.5 Thành phần Xếp hạng

Đầu vào:

- Tập các tài liệu liên quan từ kết quả truy hồi lớp nghĩa ***Retc***={<id_i,$relc_i$,ws_i>} chứa chỉ số tài liệu liên quan id_i, khoảng cách tới truy vấn $relc_i$ và tổng trọng số các lớp nghĩa trùng giữa tài liệu và truy vấn.
- Tập các tài liệu liên quan từ kết quả truy hồi quan hệ nghĩa ***Retr***={<id_i,$relr_i$>} chứa các cặp gồm chỉ số tài liệu id_i và khoảng cách $relr_i$

Đầu ra: Tập ***Ret***={<id_i,rel_i >} là danh sách tài liệu được xếp thứ tự giảm dần theo giá trị độ liên quan.

Thành phần này tính toán độ liên quan của từng tài liệu và xếp hạng theo **Mục 2.5.3** và **Mục 2.5.4**

2.7 CÁC THAM SỐ CỦA MÔ HÌNH

Mô hình có các tham số có thể thay đổi như sau:

2.7.1 VLO

VLO có thể được thay đổi theo từng lĩnh vực hoặc từng ngôn ngữ khác nhau tùy theo yêu cầu áp dụng mô hình truy hồi của luận án.

2.7.2 Mô hình phân tích cú pháp phụ thuộc

Mô hình phân tích cú pháp phụ thuộc được huấn luyện trên ngữ liệu phân tích cú pháp phụ thuộc (dependency treebank).

2.7.3 Mô hình gán nhãn nghĩa

Mô hình gán nhãn nghĩa được huấn luyện trên ngữ liệu gán nhãn nghĩa. Các nhãn nghĩa này phải tương thích với VLO.

2.7.4 Hệ số kết hợp kết quả so khớp

Hệ số α $(0 \leq \alpha \leq 1)$ là hệ số kết hợp cho biết mức độ ảnh hưởng của khoảng cách theo lớp nghĩa và khoảng cách theo quan hệ phụ thuộc trong khoảng cách chung giữa văn bản và truy vấn.

2.7.5 Hệ số điều chỉnh trọng số vị trí

Hệ số $\omega \in R$ là hệ số điều chỉnh giá trị của trọng số nghĩa từ vựng và trọng số quan hệ phụ thuộc trong ngữ nghĩa của cụm từ.

CHƯƠNG 3. CƠ SỞ TRI THỨC NGỮ NGHĨA TỪ VỰNG TIẾNG VIỆT

3.1 ONTOLOGY LÀ GÌ?

Ontology là một hệ thống các khái niệm được phân chia theo bản thể của nó, nghĩa là các khái niệm được phân chia đựa vào sự ra đời của nó chứ không phải được phân chia theo từng quan điểm khác nhau.

3.2 NÉT NGHĨA LÀ GÌ?

Khái niệm nét nghĩa (semantic feature), theo tác giả Hoàng Phê[4] là:
"*những thành tố ngữ nghĩa chung cho nghĩa của các từ thuộc cùng một nhóm từ, hoặc riêng cho nghĩa của một từ, đối lập với nghĩa của những từ khác trong cùng một nhóm. Nét nghĩa được diễn đạt bằng từ (hoặc tổ hợp từ)*".
Khi xét nghĩa từ vựng là một tổ hợp của các nét nghĩa thì có thể nhận thấy các ba đặc điểm sau:

1) Nếu hai nghĩa từ vựng $\boldsymbol{sense_a}$ và $\boldsymbol{sense_b}$ có càng nhiều nét nghĩa trùng nhau sẽ $\boldsymbol{sense_a}$ và $\boldsymbol{sense_b}$ càng giống nhau.
2) Nếu hai nghĩa từ vựng $\boldsymbol{sense_a}$ và $\boldsymbol{sense_b}$ sau khi bỏ qua các nét về phương ngữ và nguồn gốc mà chúng hoàn toàn giống nhau thì có thể xem được xem chúng đồng nghĩa.
3) Nghĩa từ vựng $\boldsymbol{sense_a}$ là thượng vị của nghĩa từ vựng $\boldsymbol{sense_b}$ nếu tập các nét nghĩa của $\boldsymbol{sense_b}$ chứa tập các nét nghĩa của $\boldsymbol{sense_a}$

Ba đặc điểm này được sử dụng để xác định các lớp nghĩa và mối liên hệ thượng-hạ vị giữa chúng trong VLO.

3.3 CƠ SỞ TRI THỨC NGỮ NGHĨA TỪ VỰNG TIẾNG VIỆT LÀ GÌ?

Cơ sở tri thức ngữ nghĩa từ vựng tiếng Việt, Vietnamese Lexicon Ontology – VLO, là nơi lưu trữ nghĩa của từ vựng dưới dạng các ký hiệu và các quan hệ phụ thuộc về ngữ pháp và ngữ nghĩa giữa chúng.

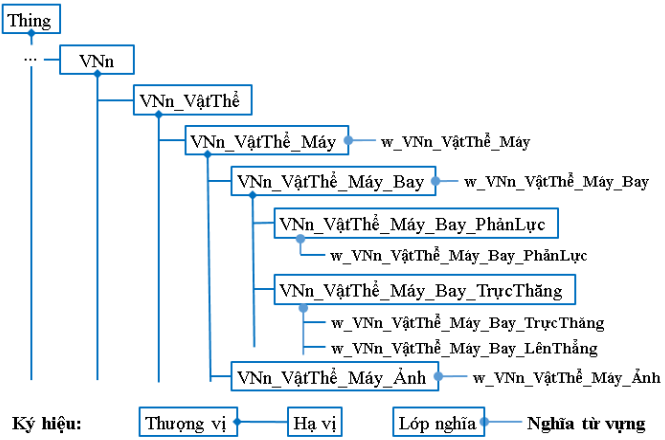
3.4 LÝ DO XÂY DỰNG VLO

- Có hệ thống nhãn nghĩa để chú giải nghĩa cho từ vựng.
- Xác lập các quan hệ phụ thuộc giữa các nhãn nghĩa theo kết quả phân tích các quan hệ phụ thuộc của các ngữ đoạn hoặc câu được sử dụng thực tế để sử dụng như là những ràng buộc ngữ nghĩa trong phân tích cú pháp và ngữ nghĩa.

3.4.1 Thể hiện chi tiết nghĩa của từ vựng

Mỗi nghĩa từ vựng (sense) khác nhau được thể hiện bằng một ký hiệu khác nhau và mỗi nghĩa từ vựng (sense) được phân lớp dựa trên từ loại, nguồn gốc và các nét nghĩa [3] của nó.

Hình 3.1 minh họa cách phân lớp nghĩa từ vựng trong VLO qua các từ "*máy ảnh*", "*máy bay phản lực*", "*máy bay trực thăng*" và "*máy bay lên thẳng*".



Hình 3.1 Minh họa cách phân lớp nghĩa từ vựng trong VLO

3.4.2 Thể hiện chi tiết các ràng buộc giữa các nghĩa từ vựng

Những ràng buộc chi tiết về ngữ nghĩa là điều kiện để nâng cao độ chính xác trong kết quả phân tích cú pháp[20].

3.4.3 Có khả năng suy diễn các quan hệ phụ thuộc

Có thể cài đặt cơ chế suy diễn các quan hệ phụ thuộc trong khung vị từ và các quan hệ phụ thuộc bổ nghĩa.

3.5 CẤU TRÚC CỦA CƠ SỞ TRI THỨC NGỮ NGHĨA TỪ VỰNG TIẾNG VIỆT

3.5.1 Các thành phần trong VLO

Cơ sở tri thức ngữ nghĩa từ vựng tiếng Việt – VLO, về hình thức là một bộ <***C***,***Sense***,***R***,***Dep***,***L***>, trong đó:

- ***C***={c_i} là tập các phân lớp c_i, gọi là lớp nghĩa (semantic class). Lớp nghĩa được sử dụng như một nhãn nghĩa trong đó tên lớp là tên nhãn nghĩa. Tập ***C*** có đặc điểm là điều kiện đảm bảo VLO là một ontology sau:

  Cho c_1, $c_2 \in C$, khi đó xảy ra một trong ba trường hợp sau:

  - $c_1 \cap c_2 = \varnothing$
  - $c_1 \cap c_2 = c_1$
  - $c_1 \cap c_2 = c_2$

- ***Sense***={$sense_i$} là tập các nghĩa từ vựng $sense_i$ sao cho:

  $\forall\ sense_i \in Sense, \exists\ c_j \in C: sense_i \in c_j$

  Nếu c_j là lớp nghĩa nhỏ nhất chứa $sense_i$ thì c_j được gọi là "*lớp nghĩa của* $sense_i$", các trường hợp khác thì c_j được gọi là "*lớp nghĩa* chứa $sense_i$"

- ***R***={***hasMod***, ***hasRMod***, ***hasPComp***, ***hasRPComp***, ***hasActor***, ***hasDObj***, ***hasIDObj***, ***hasComp***, ***hasRComp***, ***hasConj***, ***hasSyn***} là tập các quan hệ phụ thuộc.

- ***Dep***={r_i<s_x,s_y>} là tập các quan hệ phụ thuộc r_i<s_x,s_y>, trong đó $r_i \in R$, $s_x, s_y \in Sense$

- ***L*** là tập từ vựng tương ứng với tập nghĩa ***Sense***. mỗi phần tử trong ***L*** là biểu diễn từ vựng của một hoặc nhiều phần tử trong ***Sense***.

3.5.2 Các đặc điểm của VLO

1) VLO là một hệ thống phân lớp các nghĩa, có dạng cây đa phân.
2) Các lớp nghĩa con có ngữ nghĩa là sự kết hợp nghĩa của lớp nghĩa tổ tiên và nét nghĩa riêng của nó.
3) Các nghĩa từ vựng (tương ứng với sense trong WordNet) là thực thể của lớp nghĩa.
4) Nếu ***c*** Lớp nghĩa của nghĩa từ vựng ***sense*** thì:
   - Trường hợp ***c*** có duy nhất ***sense*** thì nghĩa của ***c*** cũng chính là nghĩa của ***sense***.
   - Trường hợp ***c*** có ***sense'*** và ***sense'*** đồng nghĩa với ***sense*** thì ***c*** là nghĩa chung của ***sense'*** và ***sense***. Nếu bỏ qua nét phương ngữ hay nguồn gốc của ***sense*** và ***sense'*** thì ***c*** là nghĩa của ***sense*** và ***sense'***
5) Cho hai nghĩa từ vựng $sense_a$ và $sense_b$ có lớp nghĩa của chúng lần lượt là c_a và c_b. Nếu $sense_a$ là thượng vị của nghĩa từ vựng $sense_b$ thì $c_b \subset c_c$
6) Nếu hai nghĩa từ vựng $sense_a$ và $sense_b$ cùng thuộc một lớp nghĩa từ vựng thì các quan hệ phụ thuộc nào xác lập trên $sense_a$ thì cũng xác lập trên $sense_b$
7) Nếu nghĩa từ vựng $sense_a$ là thượng vị của nghĩa từ vựng $sense_b$ thì các quan hệ phụ thuộc nào xác lập trên $sense_a$ thì cũng xác lập trên $sense_b$ (theo cơ chế suy diễn).
8) Cho hai nghĩa từ vựng $sense_a$ và $sense_b$ có lớp nghĩa của chúng lần lượt là c_a và c_b. Nếu có quan hệ phụ thuộc $r<sense_a, sense_b>$ thì quan hệ phụ thuộc $r<c_a, c_b>$ có cùng ngữ nghĩa với $r<sense_a, sense_b>$ theo đặc điểm thứ 4).

3.5.3 Xây dựng VLO

VLO được xây dựng từ dữ liệu gồm 500 ngữ đoạn và câu từ một số tin tức về khoa học của Báo trực tuyến VNExpress được thu thập vào thời điểm năm 2012 và 343 câu từ các văn bản về động lực học chất lưu được dịch từ bộ dữ liệu tiếng Anh Cranfield. Thống kê chi tiết về từ vựng và nghĩa từ vựng được trình bày trong ***Mục A.2.3.3*** của ***Phụ lục A***.

3.6 MỘT SỐ VẤN ĐỀ KHI XÂY DỰNG VLO

3.6.1 Tính khách quan

Việc xây dựng VLO phục vụ cho mục đích thử nghiệm trong luận án được thực hiện bởi một người. Để tăng cường tính khách quan trong quá trình xây dựng VLO, luận án đã thực hiện hai điều sau:

- Tham khảo nghĩa của từ vựng trong các từ điển tiếng Việt của dự án VLSP và Hán-Việt từ điển trích dẫn.
- Áp dụng kết quả phân loại từ trong từ điển tiếng Việt của dự án VLSP, các danh mục phân loại từ và danh sách các lớp nghĩa cơ bản của tác giả Đinh Điền [1].

3.6.2 Chi phí xây dựng

Chi phí xây dựng VLO cao, tập trung tại bước phân tích, đối chiếu và tổng hợp nghĩa của từ vựng, từ khoảng 10 phút đến 45 phút để xử lý một câu tiếng Việt tùy theo số lượng từ vựng và mức độ thông dụng của từ vựng.

3.6.3 Đánh giá VLO

VLO góp phần làm tăng kết quả xác định các quan hệ phụ thuộc đúng ngữ nghĩa (F_1=0.5570) so với kết quả phân tích cú pháp phụ thuộc tiếng Việt (F_1=0.3837) lên hơn 1,45 lần (145%)

3.7 KẾT CHƯƠNG

Hệ thống nhãn nghĩa và các ràng buộc ngữ nghĩa trong VLO đã làm tăng kết quả xác định các quan hệ phụ thuộc đúng ngữ nghĩa theo Universal Dependency so với kết quả phân tích cú pháp phụ thuộc tiếng Việt lên hơn 1.45 lần (145%).

Ngữ liệu được tạo ra trong quá trình xây dựng VLO là một tài nguyên quan trọng trong nghiên cứu gán nhãn nghĩa tự động và phân tích ngữ nghĩa tự động cho câu tiếng Việt.

CHƯƠNG 4. PHƯƠNG PHÁP PHÂN TÍCH NGỮ NGHĨA CỤM TỪ TIẾNG VIỆT

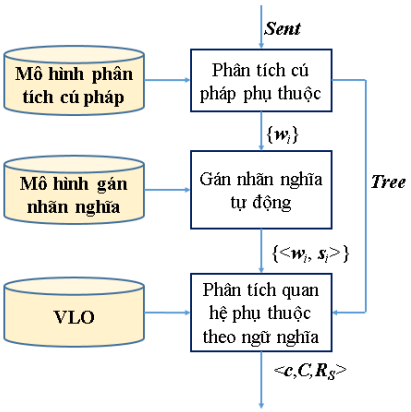
4.1 PHÂN TÍCH NGỮ NGHĨA CỦA CÂU

4.1.1 Bài toán

Mục tiêu phân tích ngữ nghĩa của câu là xác định các quan hệ phụ thuộc theo Stanford Dependency [15]. Hiện tại, đã có kết quả nghiên cứu về phân tích cú pháp phụ thuộc cho tiếng Việt của Nguyen Quoc Dat và các đồng tác giả [49] có độ chính xác P=0.7353 nhưng chưa có kết quả nghiên cứu về phân tích ngữ nghĩa câu tiếng Việt theo quan hệ phụ thuộc Stanford Dependency.

4.1.2 Hướng giải quyết vấn đề

Sử dụng các tập luật được soạn thủ công để điều chỉnh kết quả phân tích cú pháp phụ thuộc tự động từ mô hình phân tích cú pháp phụ thuộc đã được huấn luyện[49].



Hình 4.1 Sơ đồ quá trình phân tích ngữ nghĩa của câu tiếng Việt

Quá trình phân tích ngữ nghĩa của câu từ tiếng Việt được thực hiện theo sơ đồ trong
Hình 4.1

4.2 GÁN NHÃN NGHĨA CHO TỪ VỰNG

Luận án sử dụng một phương pháp dựa trên thông kê điều chỉnh từ phương pháp Transformation Based Learning nhưng bỏ qua việc chuyển đổi nhãn. Phương pháp này tương tự như phương pháp Maximum Entropy nhưng thay đổi tiêu chuẩn tối ưu từ maximum entropy sang maximum likelihood.

4.3 PHÂN TÍCH QUAN HỆ PHỤ THUỘC THEO NGỮ NGHĨA CÂU

4.3.1 Rút gọn quan hệ phụ thuộc

Áp dụng kết quả nghiên cứu về Collapsed Dependency của Schuster và Manning (2016) [62] và Ruppert (2015) [59].

4.3.2 Áp dụng các ràng buộc nghĩa và mở rộng quan hệ nghĩa

4.3.2.1 Chuyển về dạng đồ thị

Bước này chuyển cấu trúc phụ thuộc từ dạng cây về dạng đồ thị.

4.3.2.2 Điều chỉnh từ ghép

Khi kết quả phân đoạn từ ở bước phân tích hình thái chưa chính xác, dựa vào nghĩa của từng từ đơn và các ràng buộc ngữ nghĩa trong VLO để điều chỉnh từ đơn thành từ ghép.

4.3.2.3 Điều chỉnh nhãn quan hệ phụ thuộc và thành phần trung tâm

Các trường hợp cần điều chỉnh nhãn quan hệ phụ thuộc và thành phần trung tâm như sau:

- **Trường hợp 1 – Điều chỉnh trường hợp động từ làm định ngữ**: điều chỉnh lại thành một trong các loại quan hệ hasActor, hasDObj, hasIDObj dựa trên các ràng buộc trong VLO.
- **Trường hợp 2 – Điều chỉnh vai trò chủ từ, tân từ trực tiếp và tân từ gián tiếp trong câu bị động**: dựa trên các ràng buộc trong VLO.
- **Trường hợp 3 – Điều chỉnh quan hệ phụ thuộc không thỏa ràng buộc**: dựa vào các ràng buộc trong VLO để tìm từ lân cận thỏa quan hệ phụ thuộc.

4.3.2.4 Điều chỉnh vị trí của quan hệ phụ thuộc

Trong quá trình điều chỉnh các quan hệ phụ thuộc có thể dẫn đến trường hợp vi phạm ràng buộc Projectivity[13]. Vì vậy, cần phải xác từ vựng ở trí thích hợp và thỏa ràng buộc ngữ nghĩa trong VLO để điều chỉnh.

4.3.2.5 Mở rộng quan hệ phụ thuộc

Mở rộng quan hệ phụ thuộc cho các từ trong một ngữ đoạn có chứa các từ có quan hệ phụ thuộc liên từ với nhau.

4.3.2.6 Chuyển đổi về dạng biểu diễn ngữ nghĩa

Áp dụng **Định lý 2.1**

4.3.3 Biểu diễn theo cấu trúc ngữ nghĩa

Biến đổi kết quả phân tích ngữ nghĩa của câu về dạng $< c, C, R_S >$

4.4 ĐÁNH GIÁ KẾT QUẢ PHÂN TÍCH NGỮ NGHĨA

4.4.1 Đánh giá kết quả gán nhãn nghĩa

Kết quả gán nhãn nghĩa tự động theo nghiên cứu của luận án có độ chính xác đạt 0.7949 cao hơn so với kết quả gán nhãn theo phương pháp Maximum Entropy có độ chính xác đạt 0.7008

4.4.2 Đánh giá kết quả phân tích ngữ nghĩa

Kết quả phân tích ngữ nghĩa đạt F_1=0.5498 trong trường hợp phân biệt các loại quan hệ bổ nghĩa và F_1=0.557 trong trường hợp phân biệt các loại quan hệ bổ nghĩa. Nếu dùng kết quả phân tích cú pháp phụ thuộc tiếng Việt [49] thì kết quả của hai trường hợp tương ứng là F_1=0.33, F_1=0.3837

4.4.3 Đánh giá tác dụng của việc phân tích ngữ nghĩa

Thử nghiệm thực hiện theo hướng truy hồi câu (Sentence Retrieval) với dữ liệu gồm 720 câu hoặc ngữ đoạn và 30 ngữ đoạn truy vấn tiếng Việt. [18] cho thấy việc phân tích ngữ nghĩa có tác dụng nâng cao kết quả truy hồi câu với độ đạt F_1 đạt 86.5% cao hơn khi dùng từ đơn tiếng Việt (F_1 đạt 70.3%) và từ ghép tiếng Việt (F_1 đạt 71.9%).

4.5 KẾT CHƯƠNG

Phương pháp phân tích ngữ nghĩa trong luận án là cần thiết với kết quả phân tích quan hệ phụ thuộc hơn hẳn kết quả phân tích cú pháp phụ thuộc hiện tại cho tiếng Việt.

Kết quả phân tích ngữ nghĩa cải tiến độ F_1 trong truy hồi câu ("sentence retrieval").

CHƯƠNG 5. THỬ NGHIỆM VÀ ĐÁNH GIÁ

5.1 CÁC CHỈ SỐ ĐÁNH GIÁ

5.1.1 Độ chính xác, độ phủ và độ F

Độ chính xác (Precision) của phương pháp khi xử lý truy vấn q_i được ký hiệu là P_i và được tính theo công thức 5.1 sau:

$$P_i = \frac{|G_i \cap Rel_i|}{|Rel_i|} \quad (5.1)$$

Độ phủ của phương pháp khi xử lý truy vấn q_i được ký hiệu là R_i và được tính theo công thức 5.2:

$$R_i = \frac{|G_i \cap Rel_i|}{|G_i|} \quad (5.2)$$

Để thuận tiện cho việc so sánh kết quả, độ đo F kết hợp độ chính xác P và độ phủ R theo công thức 5.3 sau:

$$F_\beta = \frac{(1+\square^2)\times P\times R}{\square^2\times P + R} \quad (5.3)$$

Trong đó β là tham số để điều chỉnh tỉ lệ đóng góp của độ chính xác vào độ F. với $\beta=1$, độ đo F_1 (độ đo điều hòa) được tính theo công thức 5.4:

$$F_1 = \frac{2\times P\times R}{P + R} \quad (5.4)$$

5.1.2 Độ chính xác bộ phận

Công thức 5.5 tính độ chính xác $P_i@k$ tại k phần tử đầu tiên khi xử lý truy vấn q_i:

$$P_i@k = \frac{|G_i \cap Rel_i@k|}{|Rel_i@k|} \quad (5.5)$$

Trong đó, $R_i@k$ là tập hợp gồm k phần tử đầu tiên của tập hợp đã sắp thứ tự Rel_i

5.1.3 Độ chính xác trung bình

Độ chính xác trung bình *AP* của một truy vấn q_i được tính bằng trung bình của tất cả độ chính xác $P_i r$ tại n_r điểm có độ phủ r thay đổi. Công thức 5.6 tính AP như sau:

$$AP_i = \frac{1}{n_r} \sum_r P_i r \tag{5.6}$$

Độ chính xác trung bình MAP của tất cả truy vấn được tính theo công thức 5.7 sau:

$$MAP = \frac{1}{|Q|} \sum_i AP_i \tag{5.7}$$

TREC đề xuất việc tính AP_i trên 11 điểm có độ phủ là 0, 0.1, 0.2, .., 1.0 như được trình bày trong [38] để thống nhất kết quả khi so sánh các phương pháp. Công thức 5.8 tính độ chính xác nội suy $P_{inter}r$ tại độ phủ r như sau:

$$P_{inter}r = \max_{r<r'} Pr' \tag{5.8}$$

5.2 BỘ DỮ LIỆU THỬ NGHIỆM

Bộ dữ liệu thử nghiệm được sử dụng trong luận án VN-CRANFIELD-1 được dịch từ bộ dữ liệu thử nghiệm Cranfield gồm 100 truy vấn và 508 văn bản viết về động lực học. Kết quả thử nghiệm hai mô hình vector với công thức xếp hạng TF.IDF và mô hình xác suất với công thức xếp hạng BM25 trên VN-CRANFIELD-1cho kết quả khá tương đồng so với kết quả truy vấn của hai mô hình này trên bộ dữ liệu Cranfield gốc.

5.3 CÀI ĐẶT THỬ NGHIỆM

Các chương trình truy hồi văn bản được cài đặt thử nghiệm được đặt tên gồm TF.IDF, BM25, SEMDORE và QRYEXP.

5.3.1 Chương trình TF.IDF

- Sử dụng Apache Lucene để lập chỉ mục và truy hồi theo mô hình vector.
- Sử dụng công thức tính toán độ liên quan được điều chỉnh từ công thức tính khoảng cách cosine.

- Các văn bản và các truy vấn đã được phân tích sẵn.

5.3.2 Chương trình BM25

- Sử dụng Apache Lucene để lập chỉ mục và truy hồi theo mô hình xác suất.
- Sử dụng công thức tính toán độ liên quan theo công thức BM25.

5.3.3 Chương trình SEMDORE

- Được cài đặt theo đúng mô hình được đề xuất.
- Công thức tính độ liên quan dựa trên các công thức tính khoảng cách được sử dụng trong **Mục 2.6.4**

5.3.4 Chương trình QRYEXP

- Được cài đặt theo hướng Automatic Query Expansion.
- Kế thừa từ chương trình BM25.

5.3.5 Chương trình WE

- Được cài đặt theo mô hình vector với công thức xếp hạng là độ đo Cosine giữa hai vector ngữ nghĩa của văn bản và của truy vấn.
- Vector ngữ nghĩa của văn bản (hoặc câu truy vấn) là vector trọng tâm có trọng số của các vector từ có trong văn bản (hoặc câu truy vấn). Trọng số được tính là giá trị $TF \times IDF$ của từ tương ứng.
- Vector ngữ nghĩa của mỗi từ được ước lượng với mạng nơron theo phương pháp word2vec của tác giả Mikolov[64]

5.3.6 Chương trình LDA

- Phân tích tài liệu theo mô hình chủ đề LDA do Blei[13] đề xuất với thư viện Mallet[1].
- Công thức xếp hạng là độ đo có kết quả truy hồi cao nhất được chọn theo thử nghiệm thức tế từ 3 độ đo Cosine, Log-Likelihood và

[1] http://mallet.cs.umass.edu/

Kullback-Leibler divergence tính trên phân phối xác suất của tài liệu và truy vấn theo các chủ đề.

5.4 CÁC THỬ NGHIỆM

5.4.1 Thử nghiệm về ảnh hưởng của mô hình

Mô hình đề xuất có kết quả truy hồi văn bản MAP trong trường hợp có stemming và không có stemming (0.4516 và 0.4263) cao hơn cả hai mô hình vector (0.4421 và 0.4063) và mô hình xác suất với công thức xếp hạng BM25 (0.4456 và 0.4112) với dữ liệu Cranfield gốc bằng tiếng Anh.

5.4.2 Thử nghiệm về ảnh hưởng của term

Mô hình truy hồi được nghiên cứu trong luận án có thể đạt được kết quả MAP là 0.3822 cao hơn mô hình vector có MAP là 0.3688 nhưng thấp hơn mô hình BM25 có MAP là 0.3825 do độ chính xác gán nhãn nghĩa còn thấp. Nếu chỉ thay nghĩa của từ bằng biểu diễn văn bản của chính từ đó thì mô hình của luận án đạt kết quả MAP là 0.4045, cũng là kết quả cao nhất.

Thử nghiệm so sánh kết quả truy hồi khi dùng cùng hệ thống được nghiên cứu với kết quả phân tích cú pháp so sánh với kết quả phân tích ngữ nghĩa. Kết quả truy hồi khi chỉ dùng từ ghép đạt MAP là 0.3849 cao hơn khi chỉ dùng lớp nghĩa đạt Map là 0.3636. Kết quả truy hồi khi chỉ dùng quan hệ phụ thuộc theo cú pháp đạt MAP là 0.2241 thấp hơn khi chỉ dùng quan hệ phụ thuộc theo ngữ nghĩa đạt MAP là 0.2785.

Thử nghiệm gia tăng kích thước VLO bằng cách phân tích thêm 169 câu và cập nhật các ràng buộc ngữ nghĩa và các nghĩa từ vựng từ các câu này vào VLO. Sau đó thử nghiệm mô hình với VLO đã được tăng cường cho thấy mô hình truy hồi của luận án có kết quả MAP là 0.3845 cao hơn kết quả của chương trình BM25 với MAP là 0.3825.

Kết quả cho thấy, nếu dùng kết quả phân tích ngữ nghĩa theo nghiên cứu của luận án đã có sự cải tiến trong kết quả truy hồi văn bản và việc gia tăng kích thước của VLO có thể gia tăng kết quả truy hồi văn bản.

5.4.3 So sánh với một phương pháp Automatic Query Expansion

So sánh kết quả truy hồi của mô hình được đề xuất (MAP=0.3822) với kết quả của mô hình theo cách tiếp cận Automatic Query Expansion (chương trình **QRYEXP**) (MAP=0.3708) cho thấy mô hình của luận án cho kết quả tốt hơn.

5.4.4 So sánh với một phương pháp sử dụng vector ngữ nghĩa

So sánh kết quả truy hồi của mô hình được đề xuất (MAP=0.3822) với kết quả của mô hình vector sử dụng vector ngữ nghĩa của từ vựng (chương trình **WE**) (MAP=0.2020) cho thấy mô hình của luận án cho kết quả tốt hơn.

5.4.5 So sánh với một phương pháp sử dụng LDA

So sánh kết quả truy hồi của mô hình được đề xuất (MAP=0.3822) với kết quả của mô hình chủ đề với LDA (chương trình **LDA**) (MAP=0.1664) cho thấy mô hình của luận án cho kết quả tốt hơn.

5.5 KẾT CHƯƠNG

Mô hình truy hồi văn bản của luận án có độ chính xác trung bình cao hơn mô hình vector và mô hình xác suất với công thức xếp hạng BM25.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết luận

Mục đích của luận án khi nghiên cứu mô hình truy hồi văn bản tiếng Việt dựa trên ngữ nghĩa là giải quyết bài toán truy hồi văn bản trên biểu diễn ngữ nghĩa của văn bản đã đạt được với các kết quả sau:

- Thứ nhất, đề xuất cấu trúc tổ chức và phương pháp xây dựng VLO, một ontology về ngữ nghĩa từ vựng tiếng Việt, có vai trò của hệ thống nhãn nghĩa và ràng buộc ngữ nghĩa cho mục đích phân tích ngữ nghĩa cho câu tiếng Việt.
- Thứ hai, đề xuất phương pháp và kỹ thuật phân tích ngữ nghĩa của câu tiếng Việt theo NN-BD-NN dựa trên kết quả phân tích cú pháp phụ thuộc và VLO.
- Thứ ba, đề xuất mô hình truy hồi văn bản tiếng Việt theo hướng tiếp cận phân tích ngữ nghĩa của ngôn ngữ học tính toán và các phương pháp giải quyết cho từng bài toán của mỗi thành phần trong mô hình.

Kiến nghị

Từ những kết quả đạt được trong luận án, các vấn đề cần được tiếp tục nghiên cứu như sau:

- Thứ nhất, nghiên cứu xây dựng ngữ liệu được chú giải nghĩa từ vựng và quan hệ phụ thuộc theo Stanford Dependency cho tiếng Việt khi phát triển VLO để tạo điều kiện thực hiện các nghiên cứu về phân tích ngữ nghĩa tiếng Việt.
- Thứ hai, tiếp tục nghiên cứu phương pháp gán nhãn nghĩa với số lượng nhãn nghĩa rất lớn để tăng độ chính xác của kết quả gán nhãn nghĩa tự động.

- Thứ ba, nghiên cứu phương pháp phân tích quan hệ phụ thuộc theo Stanford Dependency cho tiếng Việt theo hướng tiếp cận học máy dựa trên ngữ liệu đã được chú giải nghĩa từ vựng và quan hệ phụ thuộc.
- Thứ tư, nghiên cứu sử dụng NN-BD-NN kết hợp với VLO để suy diễn và áp dụng vào hướng nghiên cứu Question Answering.

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ CÓ LIÊN QUAN ĐẾN LUẬN ÁN

Danh mục Bài báo hội nghị

1 Tuyen Thi-Thanh Do. "A concept identification method for Vietnamese concept-based information retrieval system". Proceedings of iiWAS 2012, Dec 3-5 2012, Bali, Indonesia. ACM Conference Proceeding series, pages: 149-152, ISBN: 978-1-4503-1306-3. 2012.

2 Tuyen Thi-Thanh Do. "Building a Vietnamese Lexicon Ontology for Syntactic parsing and Document Annotation". In Proceedings of iiWAS 2013, 2-4 Dec 2013, Vienna, Austria. ACM Conference Proceeding series, pages: 619-623, ISBN 978-1-4503-2113-6. 2013.

3 Tuyen Thi-Thanh Do. "Ontology-based Annotation and Indexing for Vietnamese Text Document". In Proceedings of iiWAS 2013, 2-4 Dec 2013, Vienna, Austria, ACM Conference Proceeding series, pages: 363-367, ISBN 978-1-4503-2113-6. 2013.

4 Tuyen Thi-Thanh Do. "A Preliminary Study on Semi-automatic Construction of Sense Tagged Corpus with WordNet Senses Using Semantic Vector". ICIST 2017, pages: 490-496, DOI: 10.1109/ICIST.2017.7926810. 2017.

Danh mục Bài báo tạp chí

5 Tuyen Thi-Thanh Do, Dang Tuan Nguyen. *Phrasal Semantic Distance for Vietnamese Textual Document Retrieval*. Tạp chí Tin học điều khiển [Journal of Computer Science and Cybernetics], Vol. 32, No. 3, pp: 185-202 . 2015.

6 Tuyen Thi-Thanh Do, Dang Tuan Nguyen. *A Framework for Vietnamese Text Document Retrieval System Based on Phrasal Semantic Analysis*. International Journal of Simulation Systems, Science & Technology - IJSSST, Vol. 15, No. 4 . 2014.

Danh mục Đề tài nghiên cứu khoa học

7 Đỗ Thị Thanh Tuyền, Nguyễn Tuấn Đăng . "Mô hình tìm kiếm văn bản tiếng Việt dựa trên ngữ nghĩa cụm từ truy vấn". Đề tài nghiên cứu khoa học cấp ĐHQG loại C năm 2013.