ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Phạm Thị Thu Uyên

TRÍCH RÚT MỐI QUAN HỆ NGỮ NGHĨA VÀ ÁP DỤNG CHO HỆ THỐNG HỎI ĐÁP TỰ ĐỘNG TIẾNG VIỆT

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

Hà Nội - 2009

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Phạm Thị Thu Uyên

TRÍCH RÚT MỐI QUAN HỆ NGỮ NGHĨA VÀ ÁP DỤNG CHO HỆ THỐNG HỎI ĐÁP TỰ ĐỘNG TIẾNG VIỆT

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

Cán bộ hướng dẫn: **PGS.TS Hà Quang Thuỵ**

Cán bộ đồng hướng dẫn: **Cử nhân Trần Mai Vũ**

Hà Nội - 2009

Lời cảm ơn

Trước tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Phó Giáo sư Tiến sĩ Hà Quang Thụy và Cử Nhân Trần Mai Vũ, người đã tận tình chỉ bảo và hướng dẫn tôi trong suốt quá trình thực hiện khoá luận tốt nghiệp.

Tôi chân thành cảm ơn các thầy, cô đã tạo những điều kiện thuận lợi cho tôi học tập và nghiên cứu tại trường Đại Học Công Nghệ.

Tôi cũng xin gửi lời cảm ơn tới các anh chị và các bạn sinh viên trong nhóm "Khai phá dữ liệu" đã giúp tôi rất nhiều trong việc hỗ trợ kiến thức chuyên môn để hoàn thành tốt khoá luận.

Cuối cùng, tôi muốn gửi lời cảm vô hạn tới gia đình và bạn bè, những người thân yêu luôn bên cạnh và động viên tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi xin chân thành cảm ơn !

Sinh viên

Phạm Thị Thu Uyên

Tóm tắt

Với sự ra đời và phát triển ngày càng mạnh mẽ trên World Wide Web đặt ra thách thức đòi hỏi việc khai thác thông tin một cách hiệu quả. Mặc dù chất lượng của các máy tìm kiếm đã được cải thiện nhưng kết quả trả về chỉ là những tài liệu có liên quan. Vì thế, hệ thống hỏi đáp ra đời là một nhu cầu cấp thiết, cung cấp cho người dùng câu trả lời ngắn gọn và chính xác nhất. Đây là một bài toán khó đối với hầu hết các ngôn ngữ nói chung trên thế giới nói chung cũng như hệ thống tiếng Việt nói riêng.

Khoá luận tập trung vào nghiên cứu các phương pháp xây dựng hệ thống hỏi đáp và đề xuất đề xuất mô hình cho hệ thông hỏi đáp tự động cho tiếng Việt dựa vào phương pháp trích rút quan hệ ngữ nghĩa bằng cách kết hợp hai phương pháp Snowball của Agichtein, Gravano [1] và phương pháp trích rút mối quan hệ sử dụng sử máy tìm kiếm của Ravichandran, Hovy [25] cho tập văn bản tiếng Việt. Thực nghiệm ban đầu của mô hình cho thấy hệ thống có thể trả lời chính xác được 89,1% câu hỏi người dùng đưa vào và khả năng đưa ra câu trả lời là 91,4%. Dựa vào kết quả trên, chúng tôi nhận thấy phương pháp trích rút mối quan hệ ngữ nghĩa được triển khai cho ngôn ngữ tiếng Việt là khả quan, phục vụ tốt cho việc xây dựng hệ thống hỏi đáp.

Mục lục

Danh sách các bảng

Danh sách hình vẽ

Danh sách các chữ viết tắt

Q&A	Question Answering
SEAL	Set Expands for Any Language
PMI	Pointwise Mutual Information
NP	Noun Phrase
UMLS	Unified Medical Language System
FSS	Fixed Seed Size
ISS	Increase Seed Size

Mở đầu

Các bài toán cơ bản cho trong xử lý ngôn ngữ tự nhiên vẫn luôn nhận được sự quan tâm đặc biệt từ các nhà nghiên cứu. Đây là nền tảng cho việc xây dựng và phát triển các bài toán ứng dụng khác. Trích rút mối quan hệ ngữ nghĩa cho một tập văn bản cũng là một trong số đó, nó đóng vài trò ngày càng quan trọng trong xử lý ngôn ngữ tự nhiên. Bài toán này tiến hành trích rút mối quan hệ giữa các khái niệm về mặt ngữ nghĩa hoặc dựa vào mối quan hệ xác định trước tìm kiếm những thông tin phục vụ cho quá trình xử lý khác. Trích rút mối quan hệ được ứng dụng nhiều cho các bài toán như: Hệ thống hỏi đáp [11,16,20,25], phát hiện ảnh qua đoạn văn bản [7], tìm mối liên hệ giữa bệnh-genes [27],…. Vì thế, vấn đề trích rút mối quan hệ ngữ nghĩa nhận được sự quan tâm rất lớn từ các nhà nghiên cứu, các hội nghị lớn trên thế giới trong những năm gần đây như: Colling, ACL, Senseval,… Đồng thời, trích rút mối quan hệ ngữ nghĩa cũng là một phần trong các dự án quan trọng mang tầm cỡ quốc tế trong lĩnh vực khai phá tri thức như: ACE (Automatic Content Extraction)[1], DARPA EELD (Evidence Extraction and Link Discovery)[2], ARDA-AQUAINT (Question Answering for Intelligence), ARDA NIMD (Novel Intelligence from Massive Data). Global WordNet[3].

Trong những năm gần đây, mặc dù đã có nhiều phương pháp mới được đưa ra nhưng bài toán trích rút mối quan hệ ngữ nghĩa vẫn được nhận sự quan tâm từ các nhà nghiên cứu cho các ngôn ngữ nói chung và tiếng Việt nói riêng. Tương tự đối với tiếng Anh, trích rút mối quan hệ ngữ nghĩa cũng đang là một vấn đề được đề cập trong các bài toán về xử lý văn bản tiếng Việt. Việc tìm ra một phương pháp tối ưu cho ngôn ngữ tiếng Việt còn đang là một vấn đề còn gặp nhiều khó khăn do hiện tại các kĩ thuật về xử lý ngôn ngữ, tài nguyên ngôn ngữ học cũng như các kĩ thuật học máy phục vụ cho quá trình xử lý còn đang được hoàn thiện. Vì thế, nhiều bài toán xử lý cho ngôn ngữ tiếng Việt còn gặp nhiều hạn chế.

Mục tiêu của khoá luận này là khảo sát, nghiên cứu để đưa ra một phương pháp trích rút mối quan hệ ngữ nghĩa tối ưu nhất cho ngôn ngữ tiếng Việt. Để tiếp cận mục tiêu

[1] *http://www.itl.nist.gov/iad/894.01/tests/ace/*.

[2] *http://w2.eff.org/Privacy/TIA/eeld.php*

[3] *http://www.globalwordnet.org*

này, khoá luận nghiên cứu và giới thiệu các phương pháp trích rút mối quan hệ ngữ nghĩa đang được quan tâm nhất hiện nay. Từ đó, đưa ra một phương pháp trích rút mối quan hệ ngữ nghĩa cho ngôn ngữ tiếng Việt bằng cách kết hợp giữa phương pháp trích rút mối quan hệ ngữ nghĩa sử dụng máy tìm kiếm [25] và phương pháp Snowball [1]. Bên cạnh đó, khoá luận cũng áp dụng phương pháp trích rút mối quan hệ ngữ nghĩa để giải quyết cho bài toán mà cũng đang nhận được sự quan tâm không kém – đó là xây dựng hệ thống hỏi đáp. Thông qua việc xây dựng hệ thống hỏi đáp tự động (question answering), hệ thống cũng đánh giá được hiệu quả của phương pháp xử lý cho bài toán trích rút mối quan hệ ngữ nghĩa mà khoá luận đưa ra.

Nội dung của khoá luận được chia thành các chương như sau:

Chương 1: Trình bày khái quát về bài toán trích rút mối quan hệ ngữ nghĩa. Chương này đề cập tới khái niệm quan hệ ngữ nghĩa, các loại quan hệ ngữ nghĩa, bài toán trích rút mối quan hệ ngữ nghĩ. Chương 1 cũng giới thiệu khái quát về hệ thống hỏi đáp tự động và một số hệ thống hỏi đáp sử dụng trích rút mẫu quan hệ ngữ nghĩa

Chương 2: Các phương pháp trích rút mẫu quan hệ ngữ nghĩa. Đây là chương trình bày tất các phương pháp trích rút mẫu quan hệ ngữ nghĩa sử dụng kĩ thuật bootstrapping theo hướng tiếp cận học bán giám sát. Đồng thời đưa ra phương pháp trích rút mẫu quan hệ ngữ nghĩa phù hợp nhất đối với tài liệu tiếng Việt.

Chương 3: Mô hình hệ thống hỏi đáp tiếng Việt sử dụng trích rút mối quan hệ ngữ nghĩa. Trình bày mô hình trích rút mẫu quan hệ ngữ nghĩa, phương pháp sinh tự động tập thực thể từ dữ liệu web. Từ đó đưa ra mô hình cho hệ thống hỏi đáp tiếng Việt áp dụng trích rút mối quan hệ ngữ nghĩa.

Chương 4: Thực nghiệm, kết quả và đánh giá. Tiến hành thực nghiệm việc sinh thực thể tự động, thực nghiệm trích rút mối quan hệ ngữ nghĩa và thực nghiệm hệ thống hỏi đáp tự động tiếng Việt.

Phần kết luận và hướng phát triển khoá luận: Tóm lược những điểm chính của khoá luận. Chỉ ra những điểm cần khắc phục, đồng thời đưa ra những hướng nghiên cứu trong thời gian sắp tới.

Chương 1. Khái quát bài toán trích rút mối quan hệ ngữ nghĩa

Để hiểu và giải quyết được bài toán trích rút mối quan hệ ngữ nghĩa, đòi hỏi chúng ta cần phải nắm vững được định nghĩa quan hệ ngữ nghĩa là gì, các đặc trưng của quan hệ ngữ nghĩa, các loại quan hệ ngữ nghĩa,… Vì thế, khoá luận trong chương này giới thiệu các vấn đề liên quan tới bài toán trích rút mối quan hệ ngữ nghĩa, làm tiền để cho việc giải quyết bài toán.

1.1 Quan hệ ngữ nghĩa

Quan hệ ngữ nghĩa (semantic relation) là một khái niệm trong ngôn ngữ học. Việc xác định quan hệ ngữ nghĩa nhận được sự rất nhiều quan tâm từ các nhà nghiên cứu về ngôn ngữ học cũng như xử lý ngôn ngữ tự nhiên.

Có rất nhiều khái niệm hay định nghĩa về quan hệ ngữ nghĩa đã được đưa ra. Theo nghĩa hẹp, Birger Hjorland đã định nghĩa **quan hệ ngữ nghĩa** [29]: *Là mối quan hệ về mặt ngữ nghĩa giữa hai hay nhiều khái niệm. Trong đó, khái niệm được biểu diễn dưới dạng từ hay cụm.*

Ví dụ: Ta có một câu "Hội Lim được tổ chức ở Bắc Ninh"

=> (Hội Lim, Bắc Ninh) có mối quan hệ là "tổ chức"

Xác định các mối quan hệ ngữ nghĩa giữa các khái niệm là một vấn đề quan trọng trong tìm kiếm thông tin. Việc làm rõ mối quan hệ giữa các khái niệm sẽ làm tăng tính ngữ nghĩa cho câu hay tập tài liệu. Đồng thời, khi tìm kiếm thông tin một vấn đề nào đó, ta có thể có được những thông tin về các vấn đề khác liên quan tới nó. Vì vậy, để tìm kiếm được những thông tin chính xác, chúng ta cần biết các loại mối quan hệ giữa các khái niệm và đồng thời tìm hiểu các phương pháp để xác định được mối quan hệ đó.

1.2 Các loại quan hệ ngữ nghĩa

Quan hệ ngữ nghĩa thể hiện mối quan hệ giữa các khái niệm, khái niệm ở đây có thể là một từ hoặc một cụm danh từ. Chúng được biểu diễn dưới dạng cấu trúc phân cấp thông qua các mối quan hệ. Dựa vào những đặc trưng và đặc tính ngữ nghĩa, ta có thể phân thành nhiều loại mối quan hệ khác nhau.

Theo Girju, một số mối quan hệ ngữ nghĩa quan trọng là thường dùng để thể hiện mối quan hệ giữa các khái niệm như: hyponymy/ hypernymy (is - a), meronymy/holonymy (part - whole), synonymy và antonymy [12].

- ***Hyponymy**:* Là một quan hệ thượng hạ vị (quan hệ giữa hai từ, trong đó một từ luôn bao gồm ngữ nghĩa của từ kia, nhưng không ngược lại). Đây là mối quan hệ ngữ nghĩa cơ bản, được sử dụng với mục đích phân loại những thực thể khác nhau để tạo ra các ontology có phân cấp.

  Ví dụ: "*Động vật*" bao gồm cả "*con chó*".

- ***Meronymy**:* Là một quan hệ ngữ nghĩa thể hiện mối quan hệ bộ phận – toàn phần (part-whole) giữa hai khái niệm. Mối quan hệ ngược lại được gọi là *holonymy*

  Ví dụ: "*tay*" là một phần của "*cơ thể con người*" ("*hand*" is a part of the "*human body*"). "*Cơ thể con người*" có một phần là "*tay*" ("*human body*" is a holonymy of "*hand*")

- ***Synonymy***: Hai từ được xem là synonymy nếu chúng cùng đề cập tới một khái niệm ngữ nghĩa, hay chúng đồng nghĩa với nhau.

  Ví dụ: "*Hoa hồng*" và "*Phần trăm*" đều chỉ về tiền trả cho người làm trung gian, mô giới trong việc giao dịch, mua bán.

- ***Antonyms***: Chúng biểu diễn mối quan hệ của hai khái niệm trái ngược nhau.

  Ví dụ: Lạnh – Ấm, Mua – bán, thành công – thất bại,…

Synonymy và antonymy đóng vai trò quan trọng trong ngôn ngữ tự nhiên. Nó giúp cho việc diễn tả tránh sự lặp lại giữa các câu khi nói về cùng một sự việc bằng cách sử dụng từ đồng nghĩa (synonymy) hoặc từ trái nghĩa (antonyms) để thể hiện sự phủ định

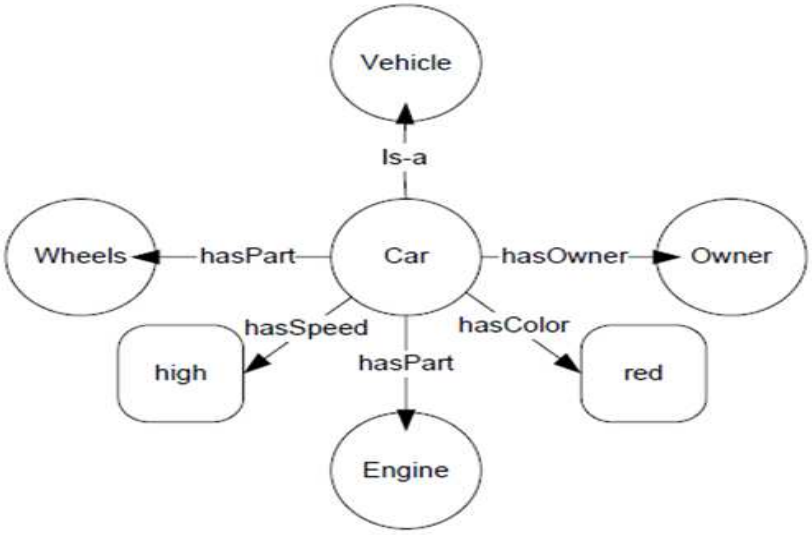
Ví dụ: Bạn A học không tốt. Học lực của bạn A là kém.

Hai câu trên về mặt ý nghĩa là hoàn toàn giống nhau, tuy nhiên việc sử dụng các từ đồng nghĩa để tránh sự lặp lại khi diễn tả sự việc "Ban A học kém".

Ngoài ra, các khái niệm và mối quan hệ giữa chúng cũng được thể hiện thông qua các tập corpus, từ điển ngôn ngữ học,….Vì thế, cũng có rất nhiều mối quan hệ khác nhau để

biểu diễn chúng. Ví dụ: WordNet[4] là một từ điển trực tuyến trong Tiếng Anh, được phát triển bởi các nhà từ điển học trường đại học Princeton. WordNet bao gồm 100.000 khái niệm bao gồm danh từ, động từ, tính từ, phó từ liên kết với nhau thông qua 17 mối quan hệ (được mô tả trong bảng 1) [12]. Thông thường, người ta thường hay sử dụng WordNet cho việc tìm kiếm các mối quan hệ ngữ nghĩa. Đồng thời, dựa vào các mối quan hệ này, một từ trong WordNet có thể tìm được các mối liên hệ với các khái niệm khác.

Ví dụ: Từ "car" trong WordNet có thể tìm được mối liên hệ với các từ như: Vehicle, Owner, Wheels, high,… thông qua các mối quan hệ như: is-a, has part, hasOwner, hasSpeed,… (như hình 1)



Hình 1. Mối liên hệ giữa từ "car" với các từ khác thông qua các mối quan hệ

Các từ được tổ chức dưới dạng synset, tức là một tập hợp gồm các từ đồng nghĩa (synonyms), hay một nhóm các khái niệm có liên quan với nhau.

Ví dụ, "exploration" và "geographic expedition" là các từ đồng nghĩa (synonym), vì thế chúng được nhóm với nhau trong một synset {exploration, geographic expedition}

Wordnet bao gồm những từ và các quan hệ phổ biến trong tiếng Anh. Ngoài các mối quan hệ giữa các danh từ là *hypernymy/hyponymy(is-a), meronymy/holonymy (a-part),*

[4] http://wordnet.princeton.edu/

synonymy, antonymy. Mối quan hệ ngữ nghĩa còn có giữa các động từ, thể hiện qua các mối quan hệ là *cause-to, entail*. Ngoài ra, *attribute* thể hiện mối quan hệ ngữ nghĩa giữa tính từ và danh từ.

Bảng 1. Mối quan hệ ngữ nghĩa trong WordNet

Mối quan hệ	Các khái niệm được liên kết với nhau bởi mối quan hệ	Ví dụ
Hypernymy (is - a)	Danh từ - Danh từ Động từ - Động từ	Cat is-a feline Manufacture is-a make
Hyponymy (reverse is-a)	Danh từ - Danh từ Động từ - Động từ	Feline reverse is-a cat Manufacture reverse is-a make
Is-part-of	Danh từ - Danh từ	Leg is-part-of table
Has-part	Danh từ - Danh từ	Table has-part leg
Is-member-of	Danh từ - Danh từ	UK is-member-of NATO
Has-member	Danh từ - Danh từ	NATO has-member UK
Is-suff-of	Danh từ - Danh từ	Carbon is-stuff-of coal
Has-stuff	Danh từ - Danh từ	Coal has-stuff carbon
Cause-to	Động từ - Động từ	To develop cause-to to grow
Entail	Động từ - Động từ	To snore entail to sleep
Atribute	Tính từ - Danh từ	Hot attribute temperature
Synonymy (synset)	Danh từ - Danh từ Động từ - Động từ	Car synonym automobile To notice synonym to observe

	Tính từ - Tính từ Phó từ - Phó từ	Happy synonym content Mainly synonym primarily
Antonymy	Danh từ - Danh từ Động từ - Động từ Tính từ - Tính từ Phó từ - Phó từ	Happines antonymy unhappiness To inhale antonymy to exhale Sincere antonymy insincere Always antonymy never
Similarity	Tính từ - Tính từ	Abridge similarity shorten
See-also	Động từ - Động từ Tính từ - Tính từ	Touch see-also touch down Inadequate see-also insatisfactory

1.3 Bài toán trích rút mối quan hệ ngữ nghĩa

- ***Định nghĩa bài toán trích rút mối quan hệ ngữ nghĩa.***

Như đã giới thiệu, các khái niệm có chứa trong một tập câu hay tập tài liệu luôn có mối liên hệ với nhau thông qua các mối quan hệ ngữ nghĩa. Các mối quan hệ này thường được ẩn giấu trong các câu, việc tìm ra các mối quan hệ ngữ nghĩa là rất cần thiết, nhằm phục vụ cho các bài toán xử lý ngôn ngữ. Vì thế, bài toán trích rút mối quan hệ ngữ nghĩa được đặt ra và yêu cầu cần phải được giải quyết.

Roxana Girju đã phát biểu bài toán trích rút mối quan hệ ngữ nghĩa [14] như sau: *Nhận đầu vào là các khái niệm hay thực thể, thông qua tập tài liệu không có cấu trúc như các trang web, các tài liệu, tin tức,... ta cần phải xác định được các mối quan hệ ngữ nghĩa giữa chúng*.

Các ví dụ về trích rút mối quan hệ ngữ nghĩa [14]:

[**Saturday's snowfall**]$_{\text{TEMP}}$ topped [**a record in Hartford, Connecticut**]$_{\text{LOC}}$ with [**the total of 12/5 inches**]$_{\text{MEASURE}}$, [**the weather service**]$_{\text{TOPIC}}$ said. The storm claimed its fatality Thursday when [**a car driven by a** [**college student**]$_{\text{PART-WHOLE}}$]$_{\text{THEME}}$ skidded on

[**an interstate overpass**]$_{LOC}$ in [**the mountains of Virginia**]$_{LOC/PART-WHOLE}$ and hit [**a concrete barrier**]$_{PART-WHOLE}$, police said.

Các mối quan hệ được trích rút như sau:

TEMP (Saturday, snowfall)

LOC (Hartford Connecticut, record)

MEASURE(total, 12.5 inches)

TOPIC (weather, service)

PART-WHOLE (student, college)

THEME (car, driven by a college student)

LOC (interstate, overpass)

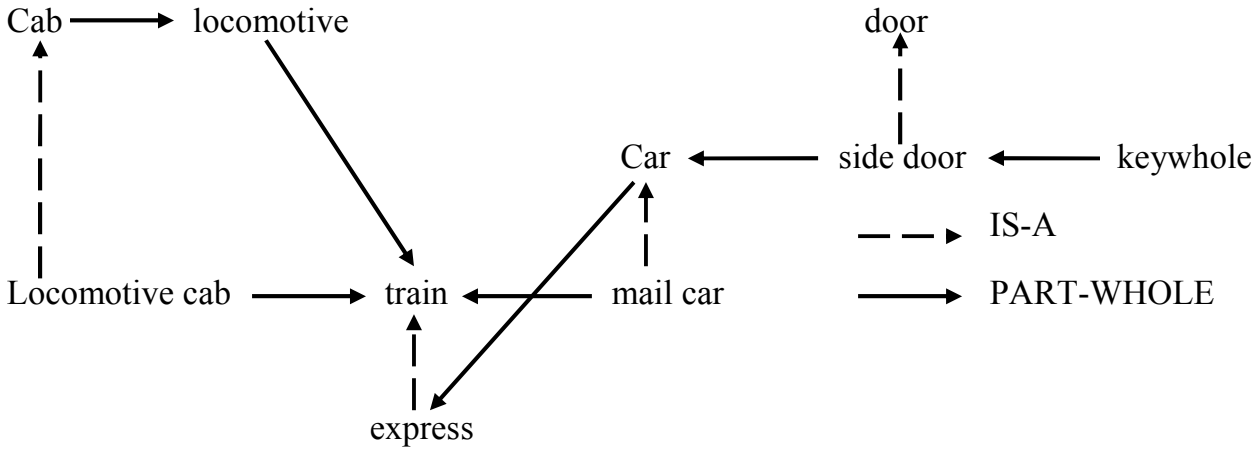
LOC (mountains, Virginia)

PART-WHOLE/LOC (mountains, Virginia)

PART-WHOLE (concrete, barrier)

Ví dụ 2:

The car's mail messenger is busy at work in [**the mail car**] $_{PART-WHOLE}$ as the train moves along. Through the open [**side door**]$_{PART-WHOLE}$ **of the car**]$_{PART-WHOLE}$, moving scenery can be seen. The worker is alarmed when he hears an unusual soud. He peeks through [**the door's keyhole**]$_{PART-WHOLE}$ leading to the tender and [**locomotive cab**] $_{PART-WHOLE}$ and sees the two bandits trying to break through [**the**[**express car**] $_{PART-WHOLE}$ **door**] $_{PART-WHOLE}$



- ***Ứng dụng của bài toán.***

Ngoài việc làm giàu thêm lượng thông tin, trích rút mối quan hệ được xem là một phương pháp hiệu quả để đưa ra phương pháp xử lý cho các hệ thống [15], như: Hệ thống

hỏi đáp (Question Answering) [11,16,20,25], KB construction [24], phát hiện ảnh qua đoạn văn bản (text-to-image generation) [7], tìm mối quan hệ bệnh tật - Genes (gene-disease relationships) [27], ảnh hưởng qua lại giữa protein-protein (Protein-Protein interaction)[17]…

Mặc dù được áp dụng cho nhiều bài toán, nhưng hiện nay trích rút mối quan hệ được tập trung nhiều nhất trong bài toán xây dựng hệ thống hỏi đáp. Việc nghiên cứu và xây dựng hệ thống hỏi đáp cũng đang là một trong các bài toán nhận được sự quan tâm lớn từ các nhà nghiên cứu hiện nay.

Trong phần tiếp theo, khoá luận nêu khái quát về bài toán xây dựng hệ thống hỏi đáp (question answering) bằng việc áp dụng phương pháp trích rút mối quan hệ ngữ nghĩa.

1.4 Hệ thống hỏi đáp dựa trên trích rút quan hệ ngữ nghĩa

1.4.1 Khái niệm hệ thống hỏi đáp

Từ những năm 1960, các nhà nghiên cứu đã nghiên cứu và tiến hành xây dựng hệ thống hỏi đáp. Đồng thời, world wide web ra đời và phát triển đã trở thành một kho dữ liệu khổng lồ. Hệ thống hỏi đáp ra đời, đã trở thành một công cụ khai thác các tài nguyên web nhằm tìm kiếm câu trả lời. Từ những quan tâm và yêu cầu thực tế, việc xây dựng hệ thống hỏi đáp ngày càng trở nên cấp thiết.

Hệ thống hỏi đáp tự động [35]: *Là hệ thống được xây dựng để thực hiện việc tìm kiếm tự động câu trả lời từ một tập lớn các tài liệu cho câu hỏi đầu vào một cách chính xác và ngắn gọn.*

Đã có rất nhiều hệ thống được ra đời áp dụng nhiều phương pháp khác nhau. Từ năm 2000, phương pháp trích rút mối quan hệ ngữ nghĩa đã được sử dụng và đã có nhiều hệ thống hỏi đáp được ra đời, như: Webclopedia[16], OntotripleQA[25],…

Mặc dù áp dụng phương pháp trích rút mối quan hệ ngữ nghĩa nhưng vẫn tuân theo quy trình xử lý cũng như các kĩ thuật xử lý ngôn ngữ vẫn phải được sử dụng để tiến hành xây dựng hệ thống hỏi đáp. Một số vấn đề quan tâm cũng như các bước xử lý cơ bản sẽ được trình bày ở phần tiếp theo.

1.4.2 Một số vấn đề quan tâm khi thiết kế hệ thống hỏi đáp

Vào năm 2002, một nhóm các nhà nghiên cứu đã đưa ra một số vấn đề cần quan tâm khi xây dựng một hệ thống hỏi đáp như sau [5]:

- *Loại câu hỏi*: Câu hỏi trong ngôn ngữ tự nhiên rất đa dạng, ẩn ý, nhập nhằng và phụ thuộc vào ngữ cảnh. Một số loại câu hỏi đang được quan tâm trong hệ hống hỏi đáp như câu hỏi về sự vật, sự kiện, định nghĩa, danh sách, quá trình, cách thức, lý do… Mỗi loại câu hỏi có những đặc trưng và khó khăn riêng, đòi hỏi phải có các chiến lược để trả lời chúng.
- *Xử lý câu hỏi*: Một câu hỏi có thể được diễn đạt qua nhiều cách khác nhau. Vì thế, xử lý câu hỏi là xác định được các câu hỏi tương tự, các quan hệ ngữ pháp, loại câu hỏi, đồng thời có thể chuyển một câu hỏi phức tạp thành chuỗi các câu hỏi đơn giản hơn.
- *Ngữ cảnh:* Câu hỏi thường được gắn với ngữ cảnh và câu trả lời cũng được đưa ra trong một ngữ cảnh xác định. Việc sử dụng các thông tin về ngữ cảnh giúp hệ thống hỏi đáp hiểu câu hỏi một cách rõ ràng, loại bỏ được các nhặp nhằng và tăng tính chính xác khi trả lời câu hỏi.
- *Nguồn dữ liệu*: Nguồn dữ liệu cho hệ thống hỏi đáp rất phong phú, có thể là sách, báo chí hay các trang web. Tuy nhiên cần đảm bảo nguồn dữ liệu có độ tin cậy và thông tin chính xác cao.
- *Trích xuất câu trả lời:* Việc trích xuất câu trả lời phụ thuộc vào nhiều yếu tố: độ phức tạp của câu hỏi, loại câu hỏi có được từ quá trình xử lý câu hỏi, dữ liệu chứa câu trả lời, phương pháp tìm kiếm và ngữ cảnh,... Câu trả lời cho người dùng cần phải đảm bảo chính xác.

1.4.3 Một số hệ thống hỏi đáp tiêu biểu

Cùng với sự phát triển bủng nổ của world wid web và sự quan tâm của các nhà nghiên cứu, đã có rất nhiều hệ thống hỏi đáp được ra đời. Một số hệ thống hỏi đáp tiêu biểu được biết đến như sau: Answer.com[5], START[6], Ask Jeeves[7], Webclopedia [16] and

[5] www.answers.com

[6] www.ai.mit.edu/projects/infolab

MURAX [21],... Trong đó, một số hệ thống hỏi đáp đã sử dụng phương pháp trích rút mối quan hệ như: Webclopedia[16], OntotripleQA [25], …

- Năm 2000, Hovy, Gerber và Hermjakob đã giới thiệu hệ thống hỏi đáp tự động Webclopedia [16]. Với mỗi câu hỏi đầu vào, hệ thống sẽ xác định câu hỏi thuộc loại nào, từ đó đưa ra một tập các mẫu cho loại câu hỏi đó và một tập các mẫu cho câu trả lời tương ứng. Sau đó, sử dụng tập mẫu câu trả lời để tìm ra những đoạn văn, những câu có chứa các thông tin liên quan và trích xuất ra câu trả lời cuối cùng đáp ứng yêu cầu người dùng.
- Năm 2002, Ravichandran và Hovy cũng đưa ra một phương pháp trích rút mối quan hệ tự động cho hệ thống hỏi đáp tự động [25]. Nhận đầu vào là những ví dụ của một loại câu hỏi (bao gồm những khái niệm là câu hỏi và câu trả lời), từ đó cho tiến hành học để trích rút mẫu và những ví dụ mới cho loại câu hỏi đó. Sau đó sẽ tiến hành trả lời dựa trên tập mẫu đã được xây dựng.
- Năm 2004, Kim, Lewis, Martinez và Goodall cũng đưa ra một hệ thống hỏi đáp OntotrileQA [20] sử dụng kĩ thuật trích rút mối quan hệ ngữ nghĩa cho các thực thể trên ontoloty đã được gán nhãn bằng tay.
- Năm 2009, một hệ thống hỏi đáp đã được xây dựng dựa vào việc trích xuất tự động các từ, khái niệm và mối quan hệ [11]. Ở đây, Fahmi đã tăng độ bao phủ các mối quan hệ bằng việc cho việc học bán giám sát để sinh tự động các mẫu quan hệ từ một tập dữ liệu lớn. Mục đích của ông là làm tăng độ chính xác bằng việc sử dụng những thông tin từ Unified Medical Language System (UMLS) và sử dụng việc lựa chọn những mối quan hệ liên quan tới các từ trong lĩnh vực y tế.

Như vậy, phương pháp trích rút mối quan hệ ngữ nghĩa cũng được sử dụng nhiều có việc xây dựng hệ thống hỏi đáp. Đồng thời, qua quá trình khảo sát và nghiên cứu, chúng tôi nhận thấy phương pháp này hầu như đều tiến hành bằng việc trích rút các mẫu quan hệ cho những mối quan hệ ngữ nghĩa đã được xác định trước.

[7] www.ask.com

1.5 Tóm tắt chương một

Trong chương này, khoá luận giới thiệu khái quát về bài toán trích rút mối quan hệ ngữ nghĩa, một số loại quan hệ ngữ nghĩa và ứng dụng của trích rút mối quan hệ ngữ nghĩa cho bài toán xây dựng hệ thống hỏi đáp. Trong chương tiếp theo, khoá luận nêu rõ các phương pháp trích rút mẫu quan hệ ngữ nghĩa và đưa ra phương pháp trích rút mẫu quan hệ ngữ nghĩa phù hợp với ngôn ngữ tiếng Việt.

Chương 2. Các phương pháp trích rút mẫu quan hệ ngữ nghĩa

Thông thường, việc xác định các mối quan hệ ngữ nghĩa thường do các chuyên gia tiến hành. Ví dụ, trong việc xây dựng WordNet, có rất nhiều nhà nghiên cứu đã tham gia xây dựng và phát triển trong nhiều năm, như: Geoge A. Miller[8], Christiane Fellbaum[9], Randee Tengi[10],… Đây là một công việc rất tốn thời gian cũng như chi phí cho việc xây dựng tài nguyên. Chính vì yêu cầu đó, đòi hỏi cần phải có một phương pháp để phát hiện tự động các mối quan hệ.

Hiện nay, các giải pháp nhằm giải quyết vấn đề này tập trung vào việc sử dụng các phương pháp học máy để trích rút mẫu tự động như: học không giám sát, học giám sát (Phương pháp trích xuất dựa vào các đặc trưng (feature based) [19], phương pháp trích xuất dựa vào tập nhân (kernel based)[6],…), học bán giám sát (DIRPE [4], Snowball [1], KnowItAll [9, 10], TextRunner [3],...). Trong các phương pháp đó, học bán giám sát được xem như là một phương pháp tối ưu để giảm thiểu chi phí cũng như tài nguyên xây dựng. Hướng tiếp cận chính được sử dụng cho việc học hiện nay thường sử dụng kĩ thuật bootstrapping. Kĩ thuật này nhận đầu vào là một tập nhỏ các hạt giống (seed) của một mối quan hệ cụ thể đã được xác định trước, từ đó tiến hành cho học để trích xuất ra một tập các mẫu quan hệ ngữ nghĩa và tiến hành sinh thêm tập seed mới. Kết quả thu được là một tập dữ liệu lớn biểu diễn mối quan hệ được quan tâm.

2.1 Phương pháp DIRPE

Vào năm 1998, Brin đã giới thiệu một phương pháp học bán giám sát cho việc trích rút mẫu quan hệ ngữ nghĩa[4]. Phương pháp được tiến hành với mối quan hệ "**author – book**" với tập dữ liệu ban đầu khoảng 5 ví dụ cho mối quan hệ này. Hệ thống DIRPE mở rộng tập ban đầu thành một danh sách khoảng 15.000 cuốn sách.

[8] http://wordnet.princeton.edu/~geo/

[9] http://wordnet.princeton.edu/~fellbaum/

[10] http://wordnet.princeton.edu/~rit/

Mô tả phương pháp DIRPE như sau:

- Xây dựng tập seed ban đầu để gán nhãn cho một số dữ liệu. Kí hiệu tập seed ban đầu là <A, B>.
- Tìm được một tập các câu có chứa đủ các thành phần của tập seed ban đầu.
- Dựa vào tập câu đã tìm được, tiến hành tìm các mẫu quan hệ giữa các thành phần của seed ban đầu. Brin định nghĩa mẫu ban đầu rất đơn giản, bằng việc giữ lại khoảng 10 kí tự trước thành phần seed đầu tiên và giữ lại phía sau thành phần thứ hai 10 kí tự. Mẫu quan hệ được biểu diễn dưới dạng sau:

[order, author, book, prefix, suffix, middle]

- Từ những mẫu mà chưa được gán nhãn ta thu được một tập các seed (author, book) mới và thêm những seed mới vào tập seed cho mối quan hệ đó.
- Quay lại bước 2 để tìm ra những seed và mẫu mới.

<u>Ví dụ:</u>

Tập seed ban đầu **(Arthur Conan Doyle, The Adventures of Sherlock Holmes)**. Và một tập các tài liệu bao gồm các cặp seed ban đầu

- ***Xác định mẫu quan hệ.***

Mẫu quan hệ có dạng như sau: **[order, author, book, prefix, suffix, middle]**

Dựa vào tập tài liệu, ta thu tập các câu có chứa tập seed ban đầu. Từ tập câu này, tiến hành trích xuất các mẫu quan hệ. (như hình 3).

Câu	Mẫu được trích xuất					
	Order	**Author**	**Book**	**Prefix**	**Suffix**	**Middle**
Read The Adventures of Sherlock Holmes by Arthur Conan Doyle online or in you email	0	Arthur Conan Doyle	The Adventures of Sherlock Holmes	Read	online or,	By
Know that Sir Arthur Conan Doyle wrote The Adventures of Sherlock Holmes, in 1892	1	Arthur Conan Doyle	The Adventures of Sherlock Holmes	now that Sir	In 1892	Wrote
When Sir Arthur Conan Doyle wrote The Adventures of Sherlock Holmes in 1892 he was high	1	Arthur Conan Doyle	The Adventures of Sherlock Holmes	When Sir	In 1892 he	Wrote
...	…	…	…	…	..	…

Hình 2. Các câu và mẫu được trích xuất

Từ đó trích xuất ra được một tập các mẫu:

[0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, when Sir, in 1892 he, wrote]

…

Sau khi được tập mẫu trên, chúng ta tiến hành so khớp (matching) các thành phần giữa, trước và sau của mỗi mẫu để gom nhóm chúng lại thành từng nhóm và loại bỏ những mẫu trùng nhau. Từ đó, ta thu được những mẫu đại diện cho một nhóm các mẫu có dạng như sau:

[từ phổ biến nhất của prefix, author, middle, book, từ phổ biến nhất của suffix]

Mẫu trích rút: **[sir, Arthur Conan Doyle, wrote, The Adventures of Sherlock Holmes, in 1892]**

- ***Việc sinh seed mới.***

Từ những mẫu hoàn chỉnh, ta xét tới những mẫu còn khuyết một vài thành phần, ví dụ như sau: **[Sir, ???, wrote, ??? in 1892]**

Sử dụng những tập mẫu như trên để tìm kiếm những tài liệu khác

"Sir *Arthur Conan Doyle* **worte** *Speckled Band* **in 1892**, that is aroud 662 years apart which would make the stories"

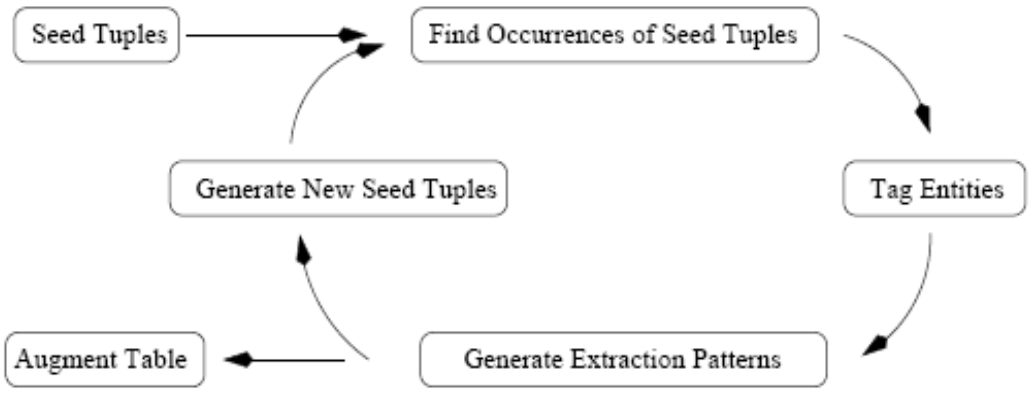
…

Từ tập câu tìm kiếm được, ta có thể trích xuất ra được những tập seed mới: (Arthur Conan Doyle, Speckled Band)

Phương pháp đạt hiệu quả cao trên dữ liệu html cho việc xác định tập mẫu và sinh seed mới. Vì thế, dựa trên ý tưởng của phương pháp DIPRE, vào năm 2000 Agichtein và Gravano đưa một phương pháp Snowball [1] tiến hành thực hiện trên dữ liệu không cấu trúc, xây dựng độ đo để đánh giá độ tin cậy cho việc sinh tập mẫu quan hệ và tập seed mới được sinh ra và bổ sung thêm việc nhận dạng thực thể. Phương pháp này được trình bày chi tiết hơn ở mục tiếp theo.

2.2 Phương pháp Snowball

Snowball là hệ thống trích rút mối quan hệ mà tập mẫu và tập seed mới được sinh ra được đánh giá chất lượng trong quá trình xử lý[1]. Họ thực nghiệm trên mối quan hệ **"tổ chức – địa điểm" ("organization – location").** Với tập seed ban đầu như: Microsoft – Redmond, IBM – Armonk, Boeing – Seatile, Intel – Santa Clara. Kiến trúc của Snowball được minh hoạ như hình dưới đây:



Hình 3. Kiến trúc của hệ thống Snowball

Phương pháp Snowball bao gồm các bước sau:

<u>Bước 1:</u> Học bán tự động để rút mẫu (extraction pattern)

Snowball bắt đầu thực hiện với tập seed ban đầu và một tập văn bản (tập huấn luyện). Các seed này mô tả đúng đắn về một mối quan hệ nào đó.

<u>Ví dụ:</u> Quan hệ: <ORGANIZATION, LOCATION>. Mỗi seed sẽ bao gồm hai thực thể A, B có mối quan hệ với nhau theo dạng: <A, B> hay <thực thể 1, thực thể 2>

Với mỗi seed <A, B>, tiến hành tìm dữ liệu là các câu có chứa cả A và B. Hệ thống sẽ tiến hành phân tích, chọn lọc và rút trích các mẫu. Sau đó, Snowball sẽ tiến hành phân cụm tập các mẫu bằng cách sử dụng hàm Match để ước tính độ tương đồng giữa các mẫu và xác định một vài ngưỡng tương đồng t_{sim} cho việc gom nhóm các cụm. Việc tính độ tương đồng sử dụng hàm Match(mẫu1, mẫu2) như sau:

Match(mẫu1, mẫu2) = (prefix1.prefix2) + (suffix1.suffix2) + (middle1.middle2)

Các mẫu sau khi tìm thấy, sẽ được đối chiếu lại với kho dữ liệu ban đầu để kiểm tra xem chúng có tìm ra được các bộ dữ liệu seed mới <A', B'> nào không. Seed mới <A', B'> sẽ nằm một trong các trường hợp sau:

- **Positive:** Nếu <A',B'> đã nằm trong danh sách seed
- **Negative:** Nếu <A', B'> chỉ có đúng một trong hai (A' hoặc B') xuất hiện trong danh sách seed.
- **Unknown:** Nếu <A', B'>, cả A', B' đều không xuất hiện trong danh sách seed. Tập Unknown được xem là tập các seed mới cho vòng lặp sau.

Snowball sẽ tính độ chính xác của từng mẫu dựa trên số Positive và Negative của nó và chọn ra top N mẫu có điểm số cao nhất. Độ tin tưởng của mẫu được tính theo công thức:

$$belief\ \ (P) = \frac{P.postive}{(P.postive\ \ + P.negative\ \ )}$$

<u>Bước 2</u>: Tìm các seed mới cho vòng lặp học tiếp theo

Với mỗi mẫu trong danh sách top N được chọn sẽ là các cặp trong tập seed mới, tiếp tục được đưa vào vòng lặp mới.

Tương tự như với mẫu thì các cặp này cũng được ước tính như sau:

$$conf(T) = 1 - \prod_{i=0}^{|p|}(1 - belief(P))$$

Hệ thống sẽ chọn ra được M cặp được đánh giá tốt nhất và M cặp này được dùng làm seed cho quá trình rút mẫu kế tiếp. Hệ thống sẽ tiếp tục được quay lại bước 1.Quá trình trên tiếp tục lặp cho đến khi hệ thống không tìm được cặp mới hoặc lặp theo số lần mà ta xác định trước.

2.3 Phương pháp trích xuất mẫu tự động sử dụng máy tìm kiếm

Năm 2002, Ravichandran và Hovy đã áp dụng kĩ thuật bootstrapping để tìm mẫu quan hệ và những seeds mới cho những câu hỏi liên quan tới ngày sinh. Tận dụng nguồn tri thức lớn từ các máy tìm kiếm như Google, Yahoo,…, phương pháp này sử dụng máy tìm kiếm phục vụ cho việc sinh mẫu quan hệ một cách tự động dựa vào các tài liệu web[25].

Thuật toán được mô tả qua các bước sau:

- Chọn các ví dụ của từng loại câu hỏi đã xác định trước.
  <u>Ví dụ:</u> Câu hỏi về ngày tháng năm sinh, và "Mozart 1756"
- Chọn những khái niệm có ở câu hỏi và câu trả lời là query để đưa vào máy tìm kiếm. Tiến hành download 1000 trang web tài liệu có liên quan, chọn tập các câu có chứa cả những khái niệm trong câu hỏi và câu trả lời.
- Tìm những xâu con hoặc các cụm có chứa các khái niệm trong câu hỏi và câu trả lời

<u>Ví dụ</u>:

- The great composer ***Mozart (1756-1791)*** achieved fame at a young age

- ***Mozart (1756 – 1791)*** was a genius
- The whole world would always be indebted to the great music of ***Mozart (1756-1791)***

Ta có thể nhận thấy xâu Mozart (1756-1791) đều xuất hiện trong cả 3 câu và nó mang đầy đủ thông tin cho câu trả lời

- Tiến hành thay thế những từ trong câu hỏi và câu trả lời bằng những tag.

Ví dụ: <NAME> (<ANSWER> - 1791)

Để đánh giá được độ chính xác của mỗi mẫu, đối với phương pháp trên thì người ta sử dụng thuật toán sau [25]:

- Sử dụng các keyword của câu hỏi như các câu truy vấn cho máy tìm kiếm. Tiến hành download 1000 trang web đầu tiên.
- Tách câu cho tập tài liệu trên, ta thu thập được một tập các câu chỉ chứa các keyword có chứa trong câu hỏi
- Với mỗi mẫu đã tìm ra ở thuật toán 1, tiến hành kiểm tra độ chính xác của từng mẫu bằng cách:
  - Kiểm tra mẫu với thẻ <ANSWER> đã được match bởi một số từ nào đó
  - Kiểm tra mẫu với thẻ <ANSWER> được match với câu trả lời đúng
- Tính độ chính xác của mỗi mẫu bằng công thức sau: $P = C_a/C_o$ với

  C_a: tổng số mẫu với câu trả lòi là đúng

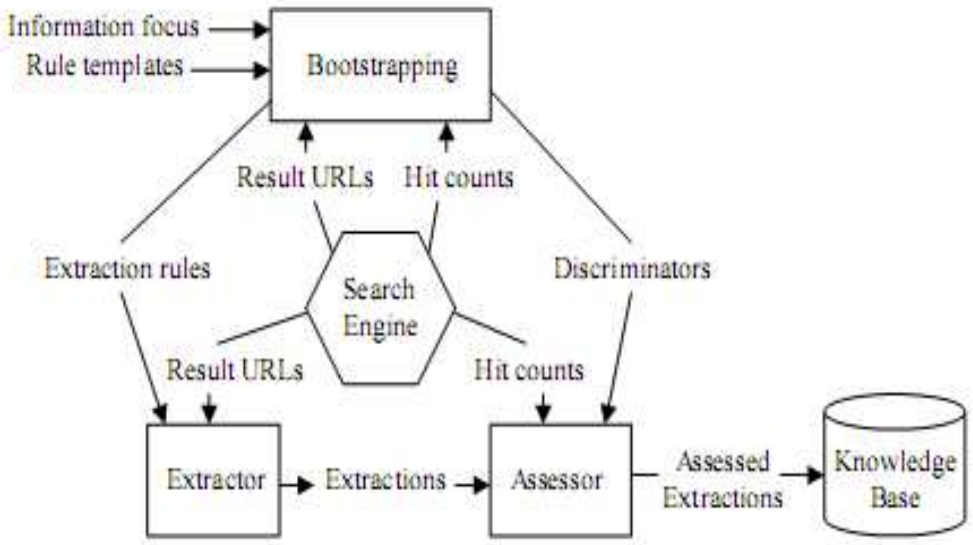
  C_o: Tổng số mẫu mà câu trả lời đúng được thay thế bởi một số từ nào đó
- Giữ lại những mẫu thích hợp mà sau khi tiến hành matching

2.4 Phương pháp KnowItAll

Phương pháp KnowItAll tiến hành trích rút ra những sự vật, khái niệm và các mẫu quan hệ từ các trang web. KnowItAll được mở rộng từ một ontology và dựa vào một tập các luật để từ đó trích rút ra các luật cho mỗi lớp và các quan hệ trong ontology [9]. Hệ thống này dựa vào miền dữ liệu và ngôn ngữ để làm đầy ontology với những sự vật và các mối quan hệ.

Đầu vào của KnowItAll là một tập các lớp thực thể được trích xuất, ví dụ như thành phố (city), nhà khoa học (scientist), bộ phim (movies),…. Và kết quả là một danh sách các thực thể được trích xuất từ các trang web. Các mẫu sử dụng đã được gán nhãn

bằng tay, những mẫu này được xây dựng dựa vào việc tách cụm danh từ (Noun Phrase chunker). Lược đồ hệ thống KnowItAll được thể hiện như hình sau [10]:



Hình 4. Lược đồ các thành phần chính của KnowItAll

Những module chính của KnowItAll như sau:

- **Trích rút (Extractor):** KnowItAll tạo ra một tập các luật trích xuất cho mỗi lớp và các mẫu chung cho nhiều mối quan hệ khác nhau.

  Ví dụ: Những mẫu chung được trích xuất như sau:

  - NP1 {“,” } “**such as**” Nplist2
    - … **including** cities **such as** Birmingham, Montgomery, Mobile, Huntsville,…
    - … publisher of books **such as** Gilamesh, Big Tree, the Last Little Cat …
  - NP1 {“,” }“**and other**” NP2
  - NP1 {“,” } “**including**” NPlist2
  - NP1 “**is a**” NP2
  - NP1 “**is the**” NP2 “**of**” NP3
  - “**the**” NP1 “**of**” NP2 “**is**” NP3

Đối với các mẫu trên thì đầu của mỗi cụm danh từ (noun phrase - NP) trong NPList2 là một ví dụ của lớp trong NP1. Mẫu này có thể được tạo ra để tìm tên các thành phố, sách, ….

Ví dụ1: Một lớp Class1 là "City" thì luật được tìm thấy là những từ như "cities such as" và trích xuất ra những từ đầu của các danh từ là những từ có khả năng.

Predicate: Class1

Pattern: NP1 "such as" NPlist2

Constraints: head(NP1) = plural (label(Class1)) &
properNoun(head(each(NPlist2)))

Bindings: Class1(head(each(NPlist2)))

Cho một câu sau: "*We provide tours to cities such as: Paris, Nice and Monte Carlo*", KnowItAll trích xuất ra được 3 ví dụ trong lớp City từ câu trên là: ***Paris, Nice và Monte Carlo***

Ví dụ 2*:* Trích xuất ra một luật cho mối quan hệ hai ngôi

NP1 "plays for" NP2

& properNoun(head(NP1))

& head(NP2) = "Seattle Mariners"

=>

instanceOf(Athlete, head(NP1))

& instanceOf(SportsTeam, heah(NP2))

& playsFor(head(NP1), head(NP2))

Keywords: "plays for", "Seattle Mariners"

- **Giao diện máy tìm kiếm (Search Engine Interface)**: KnowItAll tự động lấy những câu truy vấn dựa vào việc trích xuất luật. Mỗi luật có các câu truy vấn được tạo ra từ các từ khoá (keyword) có trong các luật.

Ví dụ: Với một luật sẽ đưa ra câu truy vấn "cities such as" vào máy tìm kiếm. Sau đó tiến hành down các trang web có chứa từ khoá, áp dụng module trích xuất (extractor) để chọn ra những câu thích hợp từ các trang web.

Ở đây, KnowItAll đã sử dụng 12 máy tìm kiếm là: Google, AltaVista, Fast,….

- **Đánh giá (Assessor):** KnowItAll sử dụng thống kê các truy vấn của máy tìm kiếm để ước tính khả năng trích rút các mẫu trong module trích rút (Extractor). Đặc biệt, Module Assessor sử dụng một dạng thông tin (pointwise mutual information - PMI) giữa các từ và các cụm từ được ước lượng từ các trang web được trả về từ máy tìm kiếm.

Ví dụ: Giả sử rằng module Extractor đã đề xuất "Liege" là tên của một thành phố. Nếu PMI giữa "Liege" và một cụm từ như "city of Liege" là cao, điều này sẽ đưa ra một tính hiển nhiên rằng "Liege" là một ví dụ chắc chắn thuộc lớp City. Module Assessor ước tính PMI giữa các ví dụ được trích xuất và những cụm từ kết hợp với các thành phố. Việc thống kê điều này thông qua cách phân lớp Naïve Bayes.

2.5 Phương pháp TextRunner

Đối với các phương pháp như DIPRE, Snowball, KnowITAll thì các loại quan hệ thường được định nghĩa trước. TextRunner thì ngược lại, phương pháp này không cần dữ liệu ban đầu mà tự động phát hiện ra các mối quan hệ [3].

Ví dụ:

Trích xuất bộ dữ liệu ba thành phần được thể hiện bởi mối quan hệ nhị phân (Arg1, relation, Arg2) từ câu "*EBay was originally founded by Pierre Omidyar*".

EBay was originally ***founded by*** **Piere Omidyar**

(Ebay, founded by, Pierre Omidyar)

TextRunner bao gồm các module chính sau đây:

- **Self-Supervised Learner**: Đầu tiên, tự động gán nhãn cho tập dữ liệu nhỏ để huấn luyện. Tiếp theo, sử dụng nhãn này để gán nhãn cho dữ liệu để huấn luyện dựa vào Naïve Bayes

Việc trích xuất được biểu diễn dưới dạng sau $t = (e_i, r_{i,j}, e_j)$ với e_i, e_j là các xâu biểu diễn cho các thực thể, $r_{i,j}$ là một xâu biểu diễn mối quan hệ giữa chúng. Với mỗi

câu được phân tích cú pháp, hệ thống sẽ tìm ra tất cả những cụm danh từ (noun pharse). Với mỗi cặp cụm danh từ (e_i, e_j), i < j, hệ thống tìm ra vị trí của chúng và tìm một cụm từ biểu diễn mối quan hệ $r_{i,j}$ trong bộ dữ liệu t.

- **Single-Pass Extractor**: Trích xuất ra những bộ dữ liệu cho tất cả những mối quan hệ có thể xảy ra. Module này không sử dụng bộ phân tích cú pháp. Extractor sẽ tìm ra các bộ dữ liệu ứng viên từ các câu, tiến hành phân loại các ứng cử viên và giữ lại những ứng viên có kết quả nhãn tốt.
- **Redundacy-Based Assessor**: Assessor tiến hành thống kê mỗi bộ dữ liệu được giữ lại dựa vào mô hình xác suất được giới thiệu trong [8]

2.6 Nhận xét

Năm 2007, cũng như các nhà nghiên cứu quan tâm đến phương pháp trích rút mẫu quan hệ ngữ nghĩa, Nguyen Bach [2] đã tổng hợp và đưa ra nhận xét sau khi tiến hành so sánh các phương pháp DIPRE, Snowball, KnowItAll và TextRunner với nhau (theo bảng 2).

Dựa vào bảng trên, ta có thể nhận thấy: Đối với phương pháp TextRunner và KnowItAll sử dụng các kĩ thuật xử lý ngôn ngữ (phân tích cú pháp, tách cụm danh từ).Vì thế, hai phương pháp khó có thể áp dụng cho tài liệu tiếng Việt vì đối với ngôn ngữ tiếng Việt, các kĩ thuật xử lý ngôn ngữ, tài nguyên ngôn ngữ học cũng như các kĩ thuật học máy đã xây dựng nhưng chưa đưa ra được kết quả tốt nhất. Đây là một vấn đề khó khăn ảnh hưởng không nhỏ đến các nghiên cứu về xử lý ngôn ngữ đối với tiếng Việt.

Đồng thời, Snowball là phương pháp cải tiến, mở rộng của phương pháp DIPRE. Phương pháp này biễn diễn các mẫu dưới dạng các vector từ có trọng số nên mẫu sinh ra có khả năng khái quát cao. Ngoài ra, snowball cũng đưa ra phương pháp tìm kiếm, trích chọn và đánh giá độ tin cậy của seed mới và mẫu mới được sinh ra. Vì thế, tập dữ liệu mới (mẫu quan hệ và tập seed mới) được sinh ra có độ tin cậy cao, chính xác từ những dữ liệu nhỏ ban đầu.

Bảng 2. So sánh các phương pháp trích rút mẫu quan hệ ngữ nghĩa

	DIPRE	Snowball	KnowItAll	TextRunner
Dữ liệu ban đầu	Có	Có	Có	Không
Mối quan hệ định nghĩa trước	Có	Có	Có	Không
Công cụ NLP được sử dụng	Không	Có: NER (Nhận dạng thực thể)	Có: NP chunker (tách cụm danh từ)	Có: dependency parser, NP chunker (Phân tích cú pháp, tách cụm danh từ)
Loại mối quan hệ	Hai ngôi	Hai ngôi	Một ngôi / Hai ngôi	Hai ngôi
Ngôn ngữ phụ thuộc	Không	Có	Có	Có
Việc phân loại (classifier)	Matching với mẫu trích xuất	Matching sử dụng hàm có độ tương đồng	Phân loại Naïve Bayses	Phân loại nhị phân tự giám sát
Tham số đầu vào	2	9	>=4	N/A

Ngoài ra như đã trình bày, phương pháp rút trích mẫu sử dụng máy tìm kiếm tận dụng được miền tri thức nền lớn từ nguồn dữ liệu các máy tìm kiếm như: Google, Altavista, Yahoo,… Vì vậy, số lượng mẫu cũng như seed mới có thể tìm kiếm được sẽ đầy đủ hơn trong tập dữ liệu web khổng lồ mà chi phí ít, hiệu quả đạt được lại cao. Tuy

nhiên, đối với phương pháp này thì chưa đưa ra kĩ thuật để sinh thêm những bộ dữ liệu mới.

Dựa vào những ưu điểm, nhược điểm trên của các phương pháp, đồng thời dựa vào điều kiện thực tế về ngôn ngữ tiếng Việt (phương pháp xử lý, tài nguyên ngôn ngữ học, kĩ thuật học máy), đối với khoá luận này, tôi quyết định sử dụng phương pháp cho việc trích rút mối quan hệ bằng cách kết hợp giữa hai phương pháp Snowball và phương pháp sử dụng máy tìm kiếm để trích xuất ra mối quan hệ ngữ nghĩa hai ngôi trong tập văn bản tiếng Việt.

Tuy nhiên, đối với phương pháp kết hợp này, đòi hỏi phải tiến hành bước nhận dạng các thực thể, đây là một bước bắt buộc để đảm bảo quá trình sinh tập seed mới cũng như việc trích rút ra được các mẫu có độ chính xác cao. Hiện nay, việc nhận dạng cũng như sinh tự động các thực thể từ tập dữ liệu Web lớn cũng là một vấn đề được quan tâm và cần phải được giải quyết cho ngôn ngữ tiếng Việt.

2.7 Tóm tắt chương hai

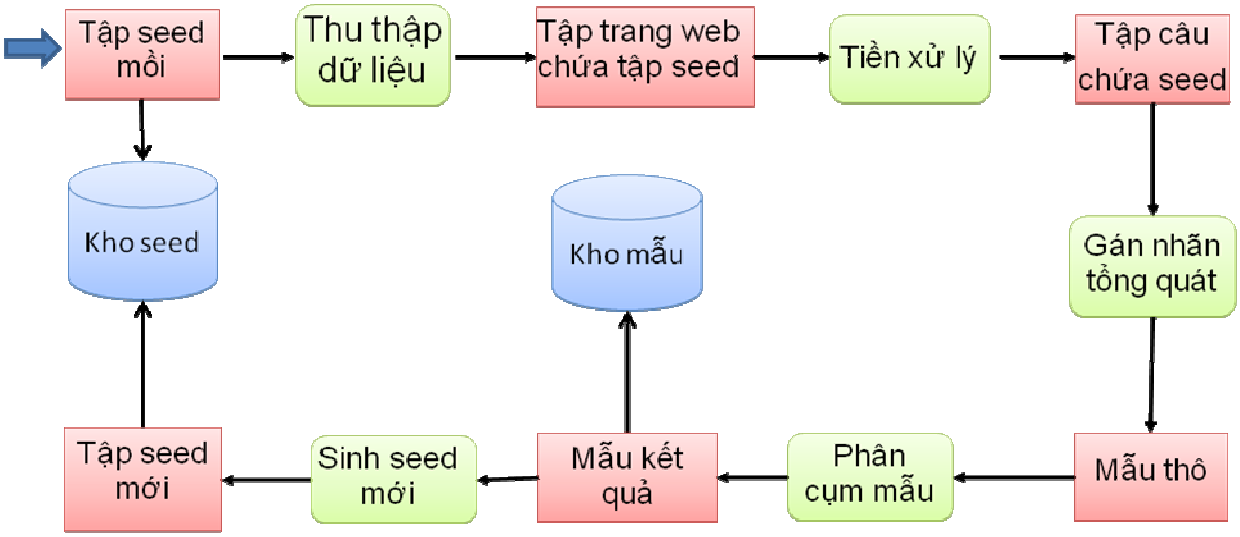
Trong chương hai, khoá luận đã giới thiệu chi tiết các phương pháp để tiến hành trích rút mối quan hệ ngữ nghĩa và đưa ra được phương pháp trích rút mối quan hệ ngữ nghĩa phù hợp với kho văn bản tiếng Việt là kết hợp phương pháp Snowball và phương pháp trích rút sử dụng máy tìm kiếm. Đồng thời, cũng giới thiệu một hệ thống sinh tự động tập thực thể cho nhiều ngôn ngữ trên thế giới và bước đầu có những kết quả cho ngôn ngữ tiếng Việt. Trong chương tiếp theo, khoá luận sẽ giới thiệu mô hình trích rút mối quan hệ và các phương pháp liên quan. Sau đó, áp dụng việc trích rút mối quan hệ ngữ nghĩa vào bài toán xây dựng hệ thống hỏi đáp tự động cho kho văn bản tiếng Việt.

Chương 3. Mô hình hệ thống hỏi đáp tiếng Việt sử dụng trích rút quan hệ ngữ nghĩa.

3.1 Mô hình trích rút mẫu quan hệ ngữ nghĩa

Qua quá trình khảo sát các phương pháp trích rút mẫu quan hệ ngữ nghĩa và dựa trên điều kiện thực tế về kĩ thuật xử lý ngôn ngữ, tài nguyên ngôn ngữ học cũng như các kĩ thuật học máy phục vụ cho quá trình xử lý ngôn ngữ tiếng Việt, khoá luận đề xuất phương pháp là kết hợp giữa phương pháp Snowball [1] và phương pháp sử dụng máy tìm kiếm [25]. Dưới dây là mô hình cho việc trích rút mẫu quan hệ ngữ nghĩa.

- **Đầu vào**: Tập dữ liệu seed mồi ban đầu, các seed gồm hai thành phần <thực thể 1, thực thể 2>
- **Đầu ra**: Tập seed mới và mẫu mới được sinh ra và được lưu vào Cơ sở dữ liệu
- **Phương pháp giải quyết và mô hình**:



Hình 5. Mô hình trích rút mẫu quan hệ ngữ nghĩa

- **<u>Bước 1</u>: Thu thập dữ liệu**
  - Nhằm tận dụng miền tri thức nền lớn từ các máy tìm kiếm như: Google, Yahoo, Altavisa,…. Ở bước này, ta sử dụng phương pháp rút trích mẫu quan hệ từ máy tìm kiếm [Mục 2.3]. Với đầu vào là một tập seed ban đầu được xây dựng bằng

tay, thông qua máy tìm kiếm ta tìm được một tập các trang web có chứa đầy đủ hai thành phần của tập seed này.

- **<u>Bước 2:</u> Tiền xử lý**
    - Loại bỏ thẻ HTML, lấy nội dung chính của từng trang web.
    - Tách câu trên tập dữ liệu thu được và giữ lại những câu chứa cả hai thành phần của seed.
    - Tách từ trong tiếng Việt. Loại bỏ từ dừng cho tập câu này
    - Áp dụng phương pháp sinh tự động tập thực thể để mở rộng tập thực thể từ những thực thể ban đầu cho từng mối quan hệ đã được xác định trước các nhãn thực thể. Phương pháp này được trình bày ở phần tiếp theo.
- **<u>Bước 3:</u> Gán nhãn tổng quát**
    - Dựa vào tập thực thể mở rộng, tiến hành tìm và xác định nhãn cho các thực thể có chứa trong tập câu thu được ở bước trên.
    - Sau khi các thực thể được gán nhãn, xác định các thành phần trái, thành phần phải, thành phần giữa cho các thực thể có chứa trong tập seed dựa vào tập câu thu được.
    - Biểu diễn các thành phần trái, thành phần phải và thành phần giữa dưới dạng các vector, ta thu được một tập các mẫu thô.
- **<u>Bước 4:</u> Phân cụm mẫu.**
    - Tiến hành so khớp các thành phần trái, thành phần phải và thành phần giữa cho các mẫu thô để loại bỏ các mẫu thô trùng.
    - Dựa theo phương pháp Snowball, xác định các mẫu quan hệ được thực hiện bằng việc phân cụm mẫu thô. Mỗi cụm đại diện bởi một mẫu và quá trình phân cụm mẫu được thực hiện như sau: Với những mẫu thô mới được sinh ra, tiến hành tính độ tương đồng với các mẫu đại diện theo công thức sau:

Match(mẫu1, mẫu2) = (prefix1.prefix2) + (suffix1.suffix2) + (middle1.middle2)

Nếu độ tương đồng vượt qua một ngưỡng xác định, thì mẫu thô đó sẽ thuộc vào nhóm có độ tương đồng với nó cao nhất. Ngược lại, mẫu đó sẽ là đại diện cho một nhóm mới được sinh ra.

- **<u>Bước 5</u>: Sinh seed mới**
    - Những mẫu tổng quát đã thu được sẽ làm đầu vào cho vào máy tìm kiếm để tìm ra tập các câu có chứa các mẫu đó.
    - Nhận dạng các thực thể có chứa trong tập câu dựa vào tập các thực thể mở rộng.
    - Kiểm tra độ tin cậy của các seed mới được sinh ra. Những seed vượt qua được giá trị ngưỡng thì giữ chúng lại.
- Sau đó quay lại bước 1, sử dụng tập seed mới thu được cùng với tập seed ban đầu đưa vào máy tìm kiếm để tiến hành sinh tập seed mới và tìm thêm tập mẫu quan hệ mới cho mối quan hệ đó. Vòng lặp sẽ được dừng khi số lượng seed mới hoặc mẫu mới không còn được tiếp tục sinh ra.

Với tập seed và mẫu mới được sinh ra sau mỗi vòng lặp, việc đánh giá độ chính xác của chúng được sử dụng theo phương pháp Snowbal [Mục 2.2].

Công thức đánh giá mẫu mới được sinh ra như sau:

$$belief\ (P) = \frac{P.postive}{(P.postive\ + P.negative\ )}$$

Công thức đánh giá các seed mới được sinh ra trong vòng lặp tiếp theo:

$$conf(T) = 1 - \prod_{i=0}^{|p|}(1 - belief(P))$$

3.2 Phương pháp sinh tự động thực thể từ tập dữ liệu Web lớn

Một trong các vấn đề đòi hỏi trong việc trích rút mối quan hệ ngữ nghĩa là việc xác định các thực thể đã được gán nhãn trong tập tài liệu. Hiện nay quá trình nhận dạng thực thể có một số phương pháp được đưa ra như [28]: xác định thực thể dựa trên luật (rule-based named entity detection), dựa vào tập từ điển (exact dictionary-based chunking), và nhận dạng thực thể sử dụng xác suất thống kê (running a statistical Named entity recognizer). Tuy nhiên, việc sử dụng thống kê cho vấn đề này lại có khả năng gây ra sai

số trong khi đó quá trình trích rút thì đòi hỏi các nhãn phải độ chính xác cao. Vì thế, khoá luận này tập trung vào việc nghiên cứu xác định các thực thể bằng cách sử dụng gán nhãn dựa vào luật và từ điển.

Hiện nay, đối với ngôn ngữ tiếng Việt, có một nghiên cứu có liên quan đến bài toán nhận dạng thực thể ở Việt Nam là công cụ VN-KIM IE được xây dựng bởi một nhóm nghiên cứu do phó giáo sư tiến sĩ Cao Hoàng Trụ đứng đầu, thuộc trường Đại học Bách Khoa Thành phố Hồ Chí Minh [30]. Tuy nhiên, phương pháp này hiệu quả chưa cao khi nhận dạng nhiều nhãn thực thể, trong khi yêu cầu của việc nhận dạng thực thể phục vụ cho việc trích rút mối quan hệ thì đòi hỏi độ chính xác lớn. Trên thế giới, đã có rất nhiều hệ thống đã giải quyết được bài toàn này cho nhiều loại ngôn ngữ. Một trong số đó là hệ thống Boowa[11], ra đời vào năm 2008 do Wang và Cohen xây dựng, hệ thống này xây dựng nhằm phục vụ cho việc tìm kiếm tự động các thực thể dựa vào một tập nhỏ các thực thể đã được gán nhãn trước[26]. Hệ thống, đã được tiến hành thực nghiệm và đem lại kết quả tốt cho một số loại ngôn ngữ như: Tiếng Anh, tiếng Nhật và tiếng Hàn Quốc, tiếng Trung Quốc,….

Hệ thống được xây dựng dựa vào hệ thống SEAL (Set Expander for Any Language) tiến hành mở rộng tập thực thể một cách tự động bằng việc phân tích nguồn tài liệu từ web. Wang và Cohen đã nghiên cứu và thực nghiệm việc sinh tự động tập thực thể bằng nhiều phương pháp khác nhau [26]. Trong đó, hai phương pháp được sử dụng là: sử dụng việc mở rộng giám sát và kĩ thuật bootstrapping. Cả hai quá trình được bắt đầu bởi một tập nhỏ seed ban đầu. Có rất nhiều cách để lựa chọn tập seed ban đầu, như: Lựa chọn tập seed ban đầu với số lượng cố định (Fixed Seed Size - FSS) và số lượng seed có thể gia tăng (Increasing Seed Size - ISS). Đồng thời, để đánh giá được tập thực thể sinh ra, hai ông cũng đã tiến hành thực nghiệm trên bốn phương pháp đánh giá sau: Random Walk with Restart, Page Rank, Bayesian Sets và Wapper Length.

Dựa trên ý tưởng này, khoá luận tập trung nghiên cứu và tiến hành việc sinh tự động tập thực thể cho ngôn ngữ tiếng Việt. Qua quá trình thực nghiệm, chúng tôi nhận thấy việc sử dụng kĩ thuật bootstrapping kết hợp với ISS và sử dụng hàm đánh giá kết quả là Random Walk with Restart đem lại kết quả cao nhất.

[11] http://boowa.com

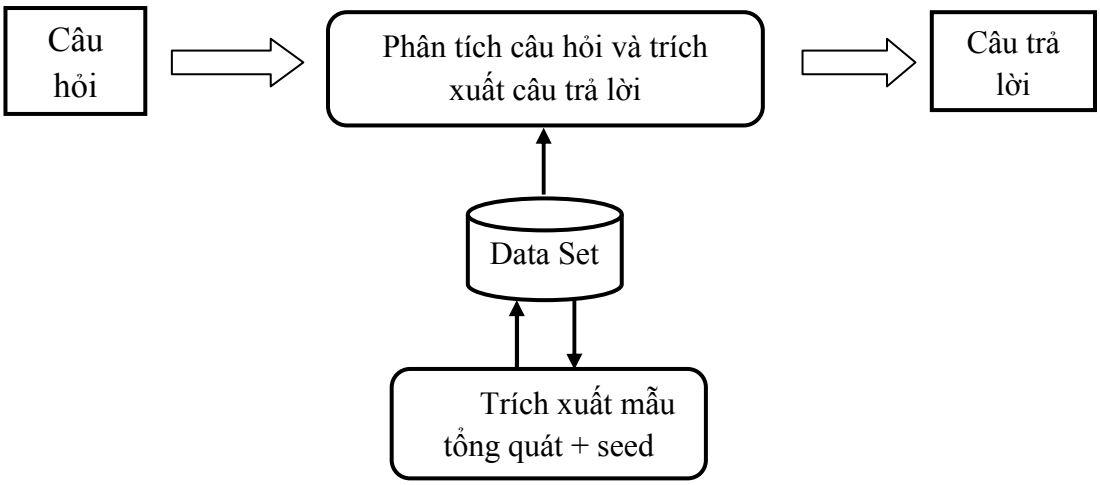
Phương pháp sinh tự động tập thực thể từ các tài liệu web bằng việc sử dụng kĩ thuật bootstrapping kết hợp với số lượng seed có thể gia tăng (ISS) được mô tả như sau:

```
starts ← ϕ, used ← ϕ
for I = 1 to m do
    if I = 1 then
        Seeds ← select_2 (E)
    else
        m = min (3, |used|)
        seeds ← select_m(used) ∪ select1(E)
    end if
    used ← used ∪ seeds
    starts ← expandstarts(seeds)
    ranked_list rankr(starts)
end for
```

Đây là giả mã cho phương pháp sử dụng giám sát mở rộng kết hợp ISS. Đối với phương pháp sử dụng kĩ thuật boostrapping kết hợp ISS thì tương tự. Tuy nhiên có một điểm khác biệt là ngoại trừ vòng lặp đầu tiên, những seed mới ở vòng lặp thứ I thì đề có những thực thể mới có độ rank cao trong vòng lặp thứ i-1

3.3 Mô hình hệ thống hỏi đáp tiếng Việt.

Từ những công trình liên quan được nêu ở các mục trên, khoá luận này đưa ra mô hình áp dụng trích rút mối quan hệ ngữ nghĩa vào hệ thống hỏi đáp tự động tiếng Việt. Phương pháp trích rút mối quan hệ ngữ nghĩa đã trình bày là sự kết hợp giữa hai phương pháp Snowball và phương pháp trích rút mối quan hệ sử dụng máy tìm kiếm. Phương pháp này tận dụng được nguồn tài nguyên dữ liệu trực tuyến khổng lồ nhằm mở rộng cũng như đánh giá được độ chính xác của tập dữ liệu thu được. Dưới dây là mô hình chung của hệ thống



Hình 6. Mô hình của hệ thống hỏi đáp tự động

Dựa vào mô hình, giải quyết bài toán qua 2 pha chính:

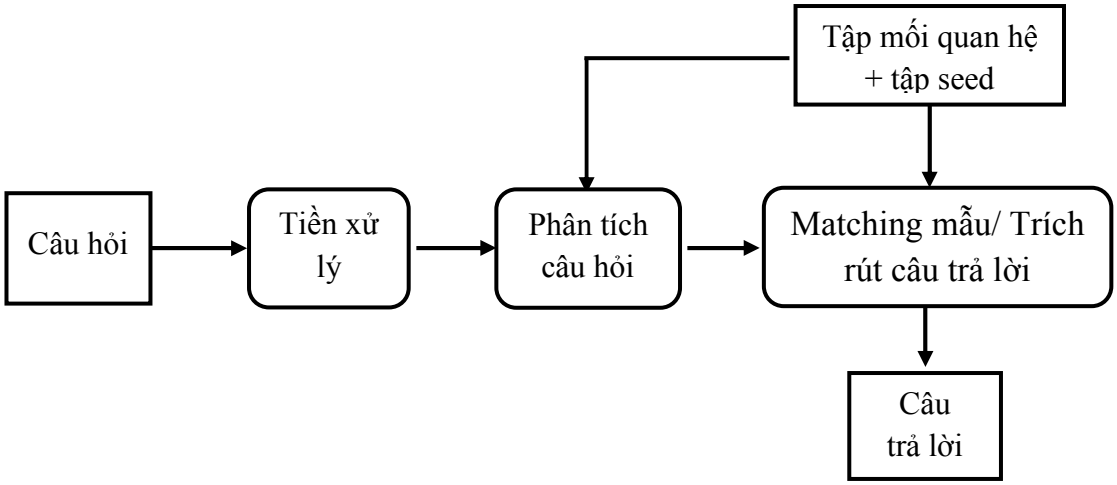
- Pha 1: Trích rút mẫu quan hệ và tập seed
- Pha 2: Phân tích câu hỏi và trích xuất câu trả lời

❖ **Pha 1: Trích rút mẫu quan hệ ngữ nghĩa hai ngôi**

- ✓ ***Input***: Tập các seed ban đầu được xây dựng bằng tay.
- ✓ ***Output:***
  - Tập mẫu tổng quát sử dụng cho việc phân tích câu hỏi và trả lời
  - Tập seed mới
- ✓ ***Phương pháp giải quyết:*** Sử dụng mô hình trích rút mối quan hệ ngữ nghĩa [Mục 3.1]

❖ **Pha 2: Phân tích câu hỏi và tìm câu trả lời**

- ✓ ***Input:*** Câu hỏi tự nhiên do người dùng đưa vào
- ✓ ***Output:*** Câu trả lời ngắn gọn và chính xác
- ✓ **Phương pháp giải quyết:**



Hình 7. Mô hình xử lý cho pha phân tích câu hỏi và trích xuất câu trả lời

<u>Bước 1:</u> Tiền xử lý câu hỏi

- Tiến hành tách từ cho câu hỏi
- Loại bỏ các từ dừng

<u>Bước 2</u>: **Phân tích câu hỏi**

- Dựa vào tập thực thể mở rộng, nhận dạng các thực thể có chứa trong câu hỏi.
- *Xác định nhãn thực thể*: Bằng cách so khớp các thực thể được nhận dạng với các thực thể có chứa trong tập seed.
- *Xác định mẫu quan hệ*: Sau khi xác định được thực thể và các nhãn, xác định các mối quan hệ có liên quan tới thực thể đó và những tập mẫu tương ứng với mối quan hệ này.
- Vector hoá câu hỏi bằng cách biểu diễn các từ trong câu hỏi dưới dạng vector từ

<u>Bước 3:</u> So khớp mẫu

- Tính độ tương đồng giữa vector câu hỏi với tập mẫu tương ứng của các mối quan hệ theo độ đo cosine
- Chọn các mẫu có độ tương đồng cao nhất. Dựa vào đó, ta xác định được mối quan hệ mà câu hỏi đang được đề cập tới.

<u>Bước 4:</u> Trích xuất câu trả lời.

- Sau khi xác định được mối quan hệ được hướng tới và các mẫu quan hệ tương ứng kết hợp với thực thể ban đầu có trong câu hỏi đã được xác định. Tiến hành trích xuất ra câu trả lời là thành phần còn lại của seed.

3.4 Tổng kết chương ba

Trong chương ba, khoá luận đã giới thiệu chi tiết mô hình trích rút mối quan hệ ngữ nghĩa cho tập văn bản tiếng Việt, giới thiệu phương pháp sinh tự động tập thực thể từ những thực thể nhỏ ban đầu đã được gán nhãn trước. Đồng thời, áp dụng trích rút mối quan hệ ngữ nghĩa để xây dựng mô hình cho hệ thống hỏi đáp tự động tiếng Việt. Trong chương tiếp theo, khoá luận sẽ tiến hành thực nghiệm dựa trên mô hình đã xây dựng trên miền dữ liệu là du lịch và sử dụng máy tìm kiếm Google để hỗ trợ cho quá trình thu thập dữ liệu.

Chương 4: Thực nghiệm và đánh giá

4.1 Môi trường và các công cụ sử dụng cho thực nghiệm

- ***Cấu hình phần cứng***

Bảng 3. Cấu hình phần cứng sử dụng trong thực nghiệm

Thành phần	Chỉ số
CPU	1 Pentium IV 3.06 GHz
RAM	1.5 GB
OS	WindowsXP Service Pack 2
Bộ nhớ ngoài	240GB

- ***Môi trường thực nghiệm***
  - Java: Java SE Development Kit (JDK) 6 (gồm jdk1.6.0_04 và jre1.6.0_04)
- ***Công cụ phần mềm sử dụng:***

Bảng 4. Một số phần mềm sử dụng

STT	Tên phần mềm	Tác giả	Nguồn
1	eclipse-SDK-3.4.1-win32		http://www.eclipse.org/downloads
2	Mysql		http://www.mysql.com
3	JvnTextpro	Nguyễn Cẩm Tú	

Ngoài ra các công cụ trên, chúng tôi tiến hành cài đặt các module xử lý dựa trên ngôn ngữ Java, bao gồm các package chính như sau:

- Vqa.datalayer.data: Sử dụng cho việc kết nối Cơ sở dữ liệu. Bao gồm các class: Pattern (sử dụng cho việc sinh tập mẫu), seed (sử dụng cho việc sinh tập seed).
- Vqa.searchEngineIE: Sử dụng cho việc thu thập dữ liệu từ máy tìm kiếm Google.
- Vqa.CharsetDector: Có nhiệm vụ sữa lỗi font chính tả.
- Vqa.util: Bao gồm các hàm tiện ích, như: xử lý chuỗi, loại bỏ dấu câu, từ dừng,…
- Ngoài ra có một số file khác:
    - PatternGenerator.java và SeedGenerator.java: Dùng để sinh ra tập mẫu quan hệ và sinh seed mới
    - QuestionProcessor: Xử lý câu hỏi đầu vào và trích xuất câu trả lời

4.2 Xây dựng tập dữ liệu

Trong khoá luận này, chúng tôi thực nghiệm với tập dữ liệu liên quan tới dữ liệu miền du lịch, sử dụng máy tìm kiếm Google và tiến hành trả lời với tập câu hỏi đơn giản liên quan tới miền du lịch.

- ***Tập các mối quan hệ và dữ liệu seed***

Qua quá trình khảo sát dữ liệu thực tế, để tạo dữ liệu phục vụ cho hệ thống hỏi đáp, chúng tôi liệt kê những mối quan hệ được quan tâm nhiều nhất trong ngành du lịch. Hiện nay chúng tôi có 85 mối quan hệ trong ngành du lịch, ví dụ: lễ hội – địa điểm, bãi biển – địa điểm, đặc sản – địa điểm, núi – chiều cao,… Với 85 mối quan hệ đã thu thập được, chúng tôi tiến hành thực nghiệm trên 10 mối quan hệ.

- **Tập dữ liệu**

Dữ liệu du lịch phục vụ cho hệ thống được crawler về từ các nguồn dữ liệu khác nhau, như là. Nguồn dữ liệu có thể được sử dụng như các website về du lịch, như: wikipedia [35], dulichvietnam.com.vn [31], vietbao.vn [34], travelatvietnam.com [33], e-cadao.com [32], … vì chúng có khả năng trả lời các câu hỏi liên quan tới sự kiện, định nghĩa khái niệm về địa danh, thông tin địa điểm, đặc điểm của khu du lịch,….

Bảng 5. Ví dụ tập các mối quan hệ và các thành phần của seed

Mối quan hệ	Thành phần thứ nhất của seed	Thành phần thứ hai của seed
Lễ hội – Địa điểm	Hội Chùa Keo	Thái Bình
Lễ hội – Địa điểm	Hội Lim	Bắc Ninh
Lễ hội – Địa điểm	Hội Chùa Hương	Hà Tây
Bãi biển – Địa điểm	Quất Lâm	Nam Định
Bãi biển – Địa điểm	Sầm Sơn	Thanh Hóa
Bãi biển – Địa điểm	Đồ Sơn	Hải Phòng
….	….	…..

- **Xây dựng tập thực thể ban đầu cho việc sinh tự động thực thể**

Tương ứng với các mối quan hệ đã được xác định trước, xác định bằng tay nhãn thực thể cho các thành phần trong seed. Với mỗi nhãn, tiến hành tìm các ví dụ cho các thực thể tương ứng.

Bảng 6. Một số thực thể được gán nhãn trước bằng tay

Nhãn thực thể	Một số thực thể được gán nhãn trước
Lễ hội	Lễ hội chùa Hương Hội Lim Hội đền Hùng
Chùa	Chùa Một Cột Chùa Thầy Chùa
Tỉnh, thành phố	Hà Nội Nam Định Hải Phòng
….	….

4.3 Thực nghiệm

4.3.1 Sinh tự động tập thực thể từ dữ liệu web

Trong khoá luận này, chúng tôi tiến hành làm thực nghiệm sinh tập thực thể tự động như sau:

- Tương ứng với mỗi mối quan hệ, gán nhãn cho thực thể trong từng mối quan hệ.
- Với mỗi nhãn thực thể, tiến hành tìm các ví dụ phổ biến nhất theo như bảng 4.
- Nhận đầu vào là các ví dụ của từng mối quan hệ, thông qua module sinh tự động tập thực thể, ta thu được một tập các thực thể có cùng loại nhãn trên.

Khoá luận thực nghiệm trên 10 mối quan hệ. Tương ứng với 10 mối quan hệ đó, tiến hành sinh tự động thực thể cho các nhãn trong mỗi một quan hệ

Bảng 7. Các nhãn thực thể và số lượng thực thể được sinh ra tự động

Nhãn thực thể	**Thực thể ban đầu đã được gán nhãn**	**Số lượng thực thể sinh ra**	**10 thực thể đầu tự động được sinh ra**
Lễ hội	Lễ hội chùa Hương Hội Bà Chúa xứ Hội đền Hùng	194	Hội chùa Thầy Hội đền Thượng Hội chùa Keo Hội đền Chử Đồng Tử Hội mùa xuân hồ Ba Bể Hội Quan Thế Âm Hội đền Cuông Hội Trường yên Hội lăng Ông
Khách sạn	Khách sạn Daewoo Khách sạn Melia	357	Khách sạn Kim Liên Khách sạn khăn quảng đỏ

	Khách sạn Fortuna		Khách sạn Công đoàn Khách sạn Sài Gòn Khách sạn Tây Hồ Khách sạn Dân chủ Khách sạn Hà Nội Khách sạn Hồ Gươm Khách sạn Bông Sen Khách sạn Đông Đô
Công viên	Công viên Thủ Lệ Công viên Thống Nhất Công viên Gia Định	54	Công viên Lênin Công viên Lê Thị Riêng Công viên Hoàng Văn Thụ Công viên nước Hồ Tây Công viên Bách Thảo Công viên Đầm Sen Công viên Tao Đàn Công viên Gò Vấp Công viên Thành Công Công viên Láng Le
Tỉnh - Thành phố	Hà Nội Hải Phòng Hồ Chí Minh	64	Đà Nẵng Nam Định Thái Bình Hải Dương Huế Hải Dương

			Thanh Hoá Bắc Ninh Cần Thơ Vũng Tàu
Chùa	Chùa Dâu Chùa Trấn Quốc Chùa Một Cột	182	Chùa Thiên Mụ Chùa Phật tích Chùa Mía Chùa Tây Phương Chùa Dơi Chùa Quán sứ Chùa Hà Chùa Keo Chùa Tây Phương Chùa Bái Đính
….	….	….	….

Nhận xét:

Đối với những nhãn thực thể phổ biến, số lượng tập thực thể được sinh ra là lớn, độ chính xác cao, đảm bảo cho việc mở rộng và nhận dạng các thực thể, phục vụ tốt cho bài toán trích rút mối quan hệ ngữ nghĩa

4.3.2 Thực nghiệm trích rút mẫu quan hệ ngữ nghĩa trong văn bản tiếng Việt

- ***Thu thập dữ liệu***
  - Tiến hành thu thập dữ liệu với query đầu vào cho máy tìm kiếm Google được biểu diễn dưới dạng như ví dụ sau: "hội chùa hương" + "hà tây" site:vi.wikipedia.org.
  - Tiến hành loại bỏ thẻ html, lấy nội dung chính của trang web
  - Tách câu cho các tài liệu trên.
  - Sử dụng công cụ JvnTextpro [36] để tách từ cho các trang web
  - Loại bỏ từ dừng, thu được một tập các câu có chứa hai thành phần của seed.
- ***Quá trình sinh mẫu quan hệ***
  - Với mỗi câu có chứa hai thành phần của seed. Tìm các chuỗi có ở trong câu trùng với thành phần của seed và thay bằng các nhãn tương ứng của chúng như: <LỄ HỘI>, <DIADIEM>.
  - Xác định các thành phần trái, thành phần phải, thành phần giữa và biểu diễn câu dưới dạng mẫu thô gồm 5 phần: <left, NHÃN 1,middle, NHÃN 2, right>
  - Loại bỏ các mẫu có thành phần giữa là rỗng hoặc dấu : -, (,) …
  - Biểu diễn các thành phần trái, phải và giữa dưới dạng vector từ và trọng số của từng từ trong từng thành phần tương ứng.
  - Tiến hành so khớp các thành phần trái, phải và giữa giữa các mẫu thô với nhau để loại bỏ các mẫu trùng lặp.
  - Phân cụm mẫu: Mỗi cụm được đại diện bởi một mẫu và quá trình phân cụm mẫu được thực hiện theo phương pháp single pass method, tức là: Với những mẫu thô mới được sinh ra, tiến hành tính độ tương đồng với các mẫu đại diện của từng nhóm theo công thức sau:

    Match(mẫu1, mẫu2) = (left1.left2) + (right1.right2) + (middle1.middle2)

    - Nếu độ tương đồng vượt qua ngưỡng cho trước, mẫu mới sinh ra sẽ thuộc nhóm có độ tương đồng nào lớn nhất với mẫu đại diện. Trong quá trình thực nghiệm, tôi lựa chọn ngưỡng cho việc sinh mẫu mới là 0,5.

- Ngược lại, nếu mẫu đó có độ tương đồng nhỏ hơn một ngưỡng xác định thì mẫu đó sẽ là đại diện cho một nhóm mới được sinh ra.

- **Quá trình sinh seed mới:**
  - Sử dụng tập các mẫu đại diện cho từng nhóm được sinh ra trong quá trình sinh mẫu làm đầu vào cho máy tìm kiếm để thu thập các tài liệu có chứa các mẫu đó.
  - Tiến hành loại bỏ thẻ html, lấy nội dung chính của trang web. Tiến hành tách từ, tách câu để lấy ra được một tập các câu có chứa các mẫu đó.
  - Dựa vào tập thực thể mở rộng, nhận dạng các thực thể có ở trong câu
  - Kiểm tra độ chính xác của các seed theo phương pháp Snowball bằng công thức tính độ tin cậy của seed mới như bên dưới.
  - Những seed nào vượt qua một độ tin cây nhất định, lưu các seed đó vào trong cơ sở dữ liệu. Ở đây, qua quá trình thực nghiệm, tôi lựa chọn ngưỡng cho việc sinh seed mới là 0,6

Với tập mẫu và seed mới được sinh ra, được tiến hành đánh giá theo phương pháp Snowball[Mục 2.2]

Công thức tính độ tin cậy của mẫu	$belief(P) = \frac{P.postive}{(P.postive + P.negative)}$
Công thức tính độ tin cậy của seed	$conf(T) = 1 - \prod_{i=0}^{\lvert p \rvert}(1 - belief(P))$

Bảng 8. Các mối quan hệ được chọn làm thực nghiệm

Tên quan hệ	Số lượng tập seed ban đầu	Số lượng mẫu thô	Số lượng mẫu tổng quát	Tập seed mới thu được
Lễ hội-địa điểm	10	509	431	194
Bãi biển – địa điểm	8	3022	1720	203
Chùa chiền – địa điểm	7	1034	756	462
Sông – địa điểm	7	256	145	57
Quán cafe – địa điểm	8	345	314	236
Nhà hàng – địa điểm	8	389	354	563
Khách sạn – địa điểm	8	245	213	346
Siêu thị - địa điểm	8	343	232	132
Công viên – địa điểm	8	234	145	38
Chợ - địa điểm	7	589	430	597

Nhận xét:

Trong quá trình thực nghiệm, tôi chỉ giữ lại các mẫu có độ tin cậy lớn hơn hoặc bằng 0.6 và các seed có độ tin cậy lớn hơn hoặc bằng 0.5. Ta có thể nhận thấy, số lượng mẫu và seed mới được sinh ra khá lớn.

4.3.3 Thực nghiệm phân tích câu hỏi và trích xuất câu trả lời cho hệ thống hỏi đáp tiếng Việt sử dụng phương pháp trích rút mối quan hệ ngữ nghĩa.

- **Tập dữ liệu test:** Chúng tôi xây dựng một bộ câu hỏi gồm 100 câu hỏi đơn giản liên quan đến 10 mối quan hệ được chọn.

- **Độ tương đồng giữa câu hỏi và mẫu:** Trong pha phân tích câu hỏi, chúng tôi sử dụng một hằng số trộn α trong công thức tính toán độ tương đồng giữa câu hỏi và mẫu trả lời.

$$\mathbf{Sim(q,p) = \alpha. Sim1(q,p) + (1- \alpha). Sim2(q,p)}$$

Trong đó:

  - q: Câu hỏi
  - p: Mẫu trả lời
  - Sim1(q,p) là độ tương đồng theo công thức cosin giữa câu hỏi q và mẫu p theo phương pháp tách từ
  - Sim2(q,p) là độ tương đồng theo công thức cosin giữa câu hỏi q và mẫu p theo phương pháp lọc các từ khóa quan trọng theo bộ từ điển danh từ (11745 từ) động từ (8600 từ) và cụm từ (16513 cụm danh từ và cụm động từ).

- **Lựa chọn hằng số trộn**: Nếu α lớn, câu hỏi và mẫu có độ tương đồng cao khi câu hỏi rất giống với mẫu. Nếu α nhỏ, câu hỏi và mẫu chỉ cần có các từ khóa danh từ, động từ giống nhau cũng cho độ tương đồng cao.

- **Lựa chọn ngưỡng tương đồng thấp nhất:** Hệ thống sử dụng một ngưỡng μ về độ tương đồng thấp nhất giữa câu hỏi và mẫu. Khi lựa chọn giá trị của μ cần cân nhắc đến sự cân bằng giữa khả năng trả lời câu hỏi chính xác nhất và khả năng trả lời được nhiều câu hỏi nhất. Nếu μ càng lớn, thì độ tương đồng giữa câu hỏi và mẫu càng cao do đó độ chính xác sẽ tăng, trong khi đó số lượng câu trả lời được sẽ giảm.

μ	**Độ chính xác**	**Khả năng đưa ra câu trả lời**
0.4	85.5%	95,3%
0.5	89,7 %	91,4%
0.6	92,6%	80,3%

Nhận xét

- Một hệ thống hỏi đáp tốt là hệ thống có khả năng đưa ra câu trả lời chính xác nhất và có thể trả lời được nhiều câu hỏi nhất. Theo thực nghiệm chúng tôi nhận thấy, độ chính xác (số lượng câu trả lời đúng trên số câu trả lời hệ thống đưa ra) và khả năng đưa ra câu trả lời (số lượng câu trả lời trên tổng số câu hỏi đưa vào) của hệ thống có quan hệ tỉ lệ nghịch với nhau. Chúng tôi chọn giá trị của $\mu = 0.5$ để đảm bảo độ cân bằng giữa 2 tính chất này của hệ thống.

Ví dụ : Câu hỏi: Nam Định có những bãi biển gì?

Bước1: Nhận dạng thực thể trong câu hỏi dựa trên tập seed. Từ đó xác định được các quan hệ tương ứng và tập mẫu của các quan hệ đó.

- ***Nam Định*** có những bãi biển gì?
- Tìm được một tập các seed có chứa một thành phần là "Nam Định".

Bảng 9. Tập seed tìm được cùng với mối quan hệ tương ứng

Mối quan hệ	Thành phần thứ nhất của seed	Thành phần thứ hai của seed
Bãi biển – Địa điểm	Quất Lâm	Nam Định
Bãi biển – Địa điểm	Hải Thịnh	Nam Định
Lễ hội – Địa điểm	Hội phủ giầy	Nam Định
…	….	….

Bước 2: Biểu diễn câu hỏi dưới dạng vector: <có, bãi_biển>

Bước 3: Tính độ tương đồng giữa câu hỏi với các mẫu trong P.

- Câu hỏi: <có, bãi_biển>
- Mẫu có độ tương đồng cao nhất với câu hỏi: <ĐỊA ĐIỂM> có bãi_biển <BÃI BIỂN> => Quan hệ là: bãi biển – địa điểm

Bảng 10. Tập các mẫu tương ứng với từng mối quan hệ

Mối quan hệ	Mẫu tổng quát
Bãi biển – Địa điểm	<BÃI BIỂN> bãi_biển thuộc <ĐỊA ĐIỂM>
Bãi biển – Địa điểm	<ĐỊA ĐIỂM> có bãi_biển <BÃI BIỂN>
Bãi biển – Địa điểm	…
Lễ hội – Địa điểm	<LỄ HỘI> khai_mạc tại <ĐỊA ĐIỂM>
Lễ hội – Địa điểm	Hằng năm <ĐỊA ĐIỂM> tổ_chức lễ_hội <LỄ HỘI>
Lễ hội – Địa điểm	…
…	….

Bước 4: Tìm câu trả lời

Từ quan hệ bãi biển – địa điểm vừa tìm thấy + tập seed S + thực thể tìm thấy trong câu hỏi, ta đưa ra được câu trả lời

- Quan hệ: Bãi biển – địa điểm
- Tập seed S:

Mối quan hệ	Thành phần thứ nhất của seed	Thành phần thứ hai của seed
Bãi biển – Địa điểm	Quất Lâm	Nam Định
Bãi biển – Địa điểm	Hải Thịnh	Nam Định
Lễ hội – Địa điểm	Hội phủ giầy	Nam Định

- Thực thể trong câu hỏi: ***Nam Định***

⇨ Câu trả lời: ***Quất Lâm, Hải Thịnh***

Nhận xét

Hệ thống hoạt động khá tốt với các câu hỏi đơn giản hỏi về quan hệ ngữ nghĩa hai ngôi xung quanh các quan hệ được quan tâm, đưa ra câu trả lời có độ tin cậy cao. Việc học ra các mẫu tốt, chính xác, thể hiện được đặc trưng của từng quan hệ là rất quan trọng,

ảnh hưởng lớn đến độ chính xác của hệ thống. Dựa vào kết quả thực nghiệm của mô hình hệ thống hỏi đáp, cho thấy việc xây dựng mô hình cho phương pháp trích rút mẫu quan hệ ngữ nghĩa kết hợp giữa phương pháp Snowball và phương pháp trích rút dựa vào máy tìm kiếm là phù hợp với ngôn ngữ tiếng Việt.

Bảng 11. Một số câu hỏi và câu trả lời tương ứng

Câu hỏi	**Câu trả lời**	**Mẫu**	**Độ tương đồng**
Hà Tây có lễ hội gì?	hội chùa hương, hội chùa thầy, hội đánh cá làng me, hội đả ngư, hội làng cổ trai, hội làng đăm, hội rước kẻ giá.	<DIADIEM> có lễ_hội <LEHOI>	0.999999
Lễ hội chùa Hương được tổ chức ở đâu?	Hà Tây	<DIADIEM> tổ_chức lễ <LEHOI>	0.71
Bãi biển Cát bà thuộc thành phố nào	Hải phòng	Bãi_biển <BAIBIEN> thuộc <DIADIEM>	0.81
Ở Nam định có bãi biển gì nổi tiếng?	Quất Lâm, Hải Thịnh	<DIADIEM> có bãi_biển <BAIBIEN>	0.7
Hồ Ba bể ở đâu?	Bắc Kạn	Hồ <HO> nằm ở <DIADIEM>	0.67
Lễ hội chùa Hương tổ chức vào thời gian nào	Hà Tây	<DIADIEM> tổ_chức lễ <LEHOI>	0.63

Kết luận

Nhu cầu xây dựng một hệ thống hỏi đáp tự động cho ngôn ngữ tiếng Việt ngày càng trờ nên cấp thiết nhằm khai thác các dữ liệu web hiệu quả hơn. Các phương pháp được sử dụng cho việc xây dựng hệ thống hỏi đáp rất đa dạng. Vì thế, vấn đề xác định phương pháp xử lý phù hợp với ngôn ngữ ngữ tiếng Việt là một phần quan trọng trong quá trình xây dựng một hệ thống hỏi đáp tự động.

Khoá luận này tiếp cận các vấn đề nói trên, tiến hành nghiên cứu và lựa chọn phương pháp trích rút mẫu quan hệ ngữ nghĩa phục vụ cho việc xây dựng hệ thống hỏi đáp tự động tiếng Việt.

Khoá luận đã đạt được những kết quả sau:

- Tìm hiểu về những vấn đề cần quan tâm khi xây dựng hệ thống hỏi đáp tự động như: việc xác định loại câu hỏi, xử lý câu hỏi, trích xuất câu trả lời, các phương pháp xử lý phù hợp với ngôn ngữ tiếng Việt.
- Nghiên cứu lý thuyết về bài toán trích rút mối quan hệ ngữ nghĩa và các phương pháp trích rút mối quan hệ ngữ nghĩa. Từ đó, đề xuất ra mô hình trích rút mối quan hệ ngữ nghĩa phù hợp với ngôn ngữ tiếng Việt cho những mối quan hệ đã được xác định trước.
- Đồng thời khoá luận đã đưa ra mô hình và xây dựng framework cho hệ thống hỏi đáp tiếng Việt sử dụng phương pháp trích rút mẫu quan hệ ngữ nghĩa trong kho văn bản tiếng Việt để trả lời những câu hỏi trong lĩnh vực liên quan.
- Kết quả của mô hình, độ chính xác là … Từ những kết quả ban đầu đó cho thấy tính đúng đắn của mô hình

Do hạn chế về thời gian và kiến thức có sẵn, khoá luận mới chỉ dừng lại ở mức thử nghiệm mô hình trên một số mối quan hệ phổ biến trong miền dữ liệu du lịch. Trong thời gian tới, tiến hành thực nghiệm trên tất cả các mối quan hệ được quan tâm trên miền dữ liệu du lịch. Đồng thời, mở rộng hệ thống trên miền dữ liệu mở và xây dựng một sản phẩm hỏi đáp tiếng Việt hoàn thiện cung cấp cho người sử dụng.

Tài liệu tham khảo

[1] Eugene Agichtein, Luis Gravano (2000). Snowball: Extracting Relations from Large Plain-Text Collections, *In proceeding of the ACL Conference, 2000,* Department of Computer Science, Columbia University

[2] Nguyen Bach. A survey on relation extraction, 2008. Sameer Badaskar.

[3] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In *Proc. 20th IJCAI*, pp. 2670–2676, Jan. 2007

[4] Brin, S. (1998). Extracting patterns and relations from the world wide web. *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT '98*

[5] Burger, John; Cardie, Claire; Chaudhri, Vinay; Gaizauskas, Robert; Harabagiu, Sanda; Israel, David; Jacquemin, Christian; Lin, Chin-Yew; Maiorano, Steve; Miller, George; Moldovan, Dan; Ogden, Bill; Prager, John; Riloff, Ellen; Singhal, Amit; Shrihari, Rohini; Strzalkowski, Tomek; Voorhees, Ellen; Weischedel, Ralph (2002). "Issues, Tasks and Program Structure to Roadmap Research in Question & Answering(Q&A)" *www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc*

[6] Bunescu, R. C., & Mooney, R. J. (2005a). A shortest path dependency kernel for relation extraction. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 724–731). Vancouver, British Columbia, Canada:* Association

[7] Coyle, B., and Sproat, R. 2001. Wordseye: An automatic text-to-scene conversion system. *Proceedings of the Siggraph Conference, Los Angeles*

[8] D. Downey, O. Etzioni, and S. Soder- land. A Probabilistic Model of Redundancy in Information Extraction. *In Proc. of IJCAI, 2005*

[9] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Web-Scale Information Extraction in KnowItAll. *In WWW, pages 100–110, New York City, New York, 2004.*

[10] Etzioni et al., 2005 O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the Web. *An experimental study. Artificial Intelligence, 165(1), 2005.*

[11] PhD ceremony: I. Fahmi, 14.45 uur, Academiegebouw, Broerstraat 5, Groningen. Thesis: Automatic term and relation extraction for medical question answering system

[12] Corina Roxana Girju (2002). Text mining for semantic relations, *PhD. Thesis*, The University of Texas at Dallas, 2002

[13] Girju R. 2001. Answer Fusion with On-Line Ontology Development. *In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) - Student Research Workshop, (NAACL 2001), Pittsburgh, PA, June 2001.*

[14] Girju R., Badulescu A., and Moldovan D. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. *In the Proceedings of the Human Language Technology Conference, Edmonton, Canada, May-June 2003*

[15] Girju R. Semantic relation extraction and its applications. Course Material. 20th European Summer School in Logic, Language and Information (ESSLLI 2008). Frete und Hansestadt Hamburg, Germany, 4-15 August 2008.

[16] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C-Y Lin (2000). Question Answering in Webclopedia, *Proceedings of the TREC-9 Conference. NIST, Gaithersbur MD*

[17] Minlie Huang and Xiaoyan Zhu and Yu Hao and Donald G. Payan and Kunbin Qu and Ming Li (2004). *Discovering patterns to extract protein-protein interactions from full texts*. **20**. pp. 3604–3612.

[18] Boris Katz (1997). Annotating the World Wide Web using Natural Language. *In Proceedings of the 5th RAIO conference on Computer Assisted information searching on the internet (RIAO'97) 1997*

[19] Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004.*

[20] Kim, S., Lewis, P., Martinez, K. and Goodall, S. (2004) Question Answering Towards Automatic Augmentations of Ontology Instances. In: *The Semantic Web: Research and Applications: First European Semantic Web Symposium, ESWS*, May 2004, Greece

[21] J.Kupiec, MURAX. A robust linguistic approach for question answering using an online encyclopedia. *In R.Korfhage, E.M. Rasmussen, and P.Willett, editors, SIGIR, pages 181-190. ACM, 1993*

[22] C. Kwork, O. Etzioni, and D. S. Weld. Scaling question answering to the web. *In WWW, vol. 10, pages 150-161, Hong Kong, May 2001, IW3C2 and ACM.* www.10.org/cdrom/papers/120/ .

[23] Ryan McDonald, Fernando Periera, Seth Kulick, Scott Winters, Yang Jin and Pete White. Simple Algorithms for Complex Relation Extraction with Applications to Biomedical IE.

[24] D. Moldovan and R. Girju. 2001. An Interactive Tool For The Rapid Development of Knowledge Bases. *In International Journal on Artificial Intelligence Tools (IJAIT)*

[25] Deepak Ravichandran, Eduard Hovy (2002). Learning Surface Text Patterns for a Question Answering System, *In Proceedings of the ACL Conference*, 2002, Information Sciences Institute University of Southern California

[26] Richard C. Wang and William W. Cohen, Iterative Set Expansion of Named Entities using the web. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Pages 1091-1096, 2008*

[27] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, Jun-ichi Tsujii (2006). "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning". *Pacific Symposium on Biocomputing*.

[28] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

[29] http://www.db.dk/bh/Lifeboat_KO/CONCEPTS/semantic_relations.htm

[30] http://www.dit.hcmut.edu.vn/~tru/VN-KIM/products/vnkim-ie.htm

[31] http://dulichvietname.com.vn

[32] http://e-cadao.com

[33] http://travelvietnam.com

[34] http://vietbao.vn

[35] http://wikipedia.org

Công cụ sử dụng

[36] Nguyen Cam Tu (2008). "JVnTextpro: A Java-based Vietnamese Text Processing Toolkit"